UNIVERSIDAD POLITÉCNICA DE MADRID

Escuela Técnica Superior de Ingenieros en Topografía, Geodesia y Cartografía

Metodología para la integración basada en ontologías de información de bases de datos heterogéneas en el dominio hidrográfico

Tesis Doctoral

Autor: Luis Manuel Vilches Blázquez

Directores: Óscar Corcho García
Miguel Ángel Bernabé Poveda

Madrid, 2011

UNIVERSIDAD POLITECNICA DE MADRID

Tribunal nombrado por el Magfco. y Excmo. Sr. Rector de la Universidad Politécnica de Madrid, el día.....de.....de 20.....

Presidente: _____

Vocal: _____

Vocal: _____

Vocal: _____

Secretario: _____

Suplente: _____

Suplente: _____

Realizado el acto de defensa y lectura de la Tesis el día.....de.....de 20.... en la E.T.S.I. /Facultad.....

EL PRESIDENTE LOS VOCALES

A mis padres, madres, hermanas y a Évelin

Agradecimientos

Hay mucha gente a la que quiero agredecer su apoyo durante este tiempo y que de alguna manera, en mayor o menor medida, han contribuido a la realización de este trabajo:

En primer lugar, quiero agradecer a Miguel Ángel Bernabé Poveda su sinceridad en todo momento y la inspiración que tuvo aquella tarde a los píes de la Alhambra paravincularme al director de esta tesis. A Óscar Corcho García quiero agradecerle su interés desde el inicio en el presente trabajo y su empeño para que finalizara esta etapa de mi vida. Gracias a los dos.

Quiero hacer extensivo este agradecimiento a todos los miembros de la extinta Subdirección General de Aplicaciones Geográficas del Instituto Geográfico Nacional (IGN), especialmente a los miembros de los proyectos IDEE y SIGNA. Asimismo, quiero hacer una mención espacial a Antonio F. Rodriguez por "presentarme" la Web Semántica y las ontologías. Siempre recordaré especialmente a Julio Mézcua por su apoyo en momentos difíciles, por disfrutrar de la investigación y por animarme a realizar mi estancia. No quiero olvidarme de todos aquellos que me hicieron sentir un miembro más del IGN y con los que pasé momentos inolvidables (Paquito, Sara, Antonio, Richard, Juan, las chicas del "palomar"…, la buena gente del Atlas, Teledetección y Producción. Y, como no, de Adela Alcázar y Marga Azcárate que vivieron los inicios de *hydrOntology* y que siempre estuvieron dispuestas a colaborar.

Quiera también agradecer a mis compañeros del *Ontology Engineering Group* (OEG) por acogerme como un "informático más" a pesar de ser "uno de los raros", así como por su atención y ayuda en los momentos en los que la he necesitado. En este sentido, quería darle las gracias especialmente a Víctor Saquicela, María Poveda, José Ángel Ramos, a todos aquellos con los que he tenido la oportunidad y suerte de colaborar (Ángel, Boris, Freddy, Elena, Miguel Ángel, César y el equipo de España Virtual) y a los miembros del *workshop* por los buenos momentos que hemos disfrutado y que nos quedan por disfrutar.

En el plano familiar, agradecer a mis padres que me enseñaran la importancia del duro trabajo y a ser constante. Así como por hacerme entender el valor de ciertos aspectos de la vida. A mis otros padres (Agustín y Eli) por todo su cariño, apoyo y buenos consejos. Por último, aunque no menos importante, agradezco los ánimos, paciencia y todo el amor recibido de Évelin, quien ha sufrido, de cerca y en la distancia, los buenos y malos momentos de este trabajo.

Resumen

El objetivo principal de esta tesis se centra en la integración, basada en ontologías, de información de bases de datos geoespaciales heterogéneas con la finalidad de llevar a cabo una visión integrada de la realidad abstraída en distintas fuentes de información.

A pesar de la existencia de múltiples propuestas para la integración de información basada en ontologías, hasta el momento no se han desarrollado metodologías que se puedan aplicar en el área de la información geográfica. Este hecho conlleva, al menos, cuatro importantes limitaciones en los procesos de integración de información:

1. Las propuestas existentes no resultan fácilmente replicables.
2. No es fácil conocer las características de los enfoques propuestos.
3. Los escenarios de integración presentan fuentes de información similares y casos y problemáticas restringidas).
4. No se pueden realizar, de manera óptima, las tareas de planificación necesarias.

Estas carencias demuestran la necesidad de una propuesta metodológica que instruya el proceso de integración basado en ontologías de información geoespacial heterogénea. Por ello, con el objetivo de solventar las carencias detectadas, en este trabajo de tesis doctoral se llevan a cabo las siguientes contribuciones:

En primer lugar, se ha propuesto una metodología para la integración de datos geoespaciales, aplicada sobre un caso práctico en el dominio hidrográfico.

En el proceso de aplicación de esta metodología se han desarrollado recursos ontológicos de dominio y multilingües en el área de la información hidrográfica. Este trabajo ha permitido la generación de siete conjuntos de datos RDF, donde se incluyen características geométricas de la información geoespacial.

Además, se ha publicado información hidrográfica del territorio nacional español conforme a los principios de *Linked Data*, contribuyendo al enriquecimiento de la Web de los Datos con información geoespacial en español.

Finalmente, se han proporcionado diversos enfoques heurísticos para la detección de instancias duplicadas en conjuntos de datos RDF heterogéneos

La validez de estas contribuciones ha sido respaldada por la aplicación a un caso de uso (información hidrográfica) y por una serie de experimentos realizados. Los resultados obtenidos apuntan a la aplicabilidad de la metodología y la viabilidad de modelos (ontologías) y técnicas (enfoques heurísticos).

Abstract

The main objective of this PhD thesis focuses on the ontology-based integration of information related to geospatial heterogeneous databases with the purpose of providing an integrated view of the stored information into different datasets.

Nowadays, there are different proposals for ontology-based integration information. Nevertheless, there is not a methodology that might be applied to the geographical information realm. This issue involves, at least, four important limitations in information integration processing:

1. The existing proposals are not easily replicable.
2. It is not easy to understand the characteristics of the approaches proposed.
3. The integration scenarios are restricted to similar datasets, use cases, and problems.
4. Planification tasks do not perform in a right way.

These limitations prove that there exists a need for a methodological approach to face the ontology-based integration of geospatial heterogeneous information. In this sense, this work proposes the following contributions:

Firstly, we have provided a methodological approach for ontology-based integration of geospatial information. This approach has been applied to a use case in the hydrographical information context.

Besides, we have developed several domain and multilingual ontological resources in the hydrographical realm. These resources have been used for generating seven RDF datasets, which include geometrical characteristics as a particular trait of geospatial data.

Moreover, we have published the Spanish hydrographical information according to the Linked Data principles. This initiative contributes to enrich the Web of Data with geospatial data in the Spanish language.

Finally, we have provided different heuristic approaches for discovering duplicate instances into heterogeneous RDF datasets.

The validity of these contributions has been supported by a use case (hydrographical information) and by a set of experiments performs. The obtained results reveal the applicability of the methodological approach and the feasibility of models (ontologies) and techniques (heuristic approaches).

Índice general

1. Introducción 1

1.1. De "islas de información" a datos integrados y abiertos 1
1.2. Problemas asociados a la integración de datos geográficos 3
1.3. El uso de semántica como solución a los problemas 4
1.4. Propuesta de tesis 4
1.5. Estructura de la tesis 6

2. Análisis de la heterogeneidad en la información geográfica 7

2.1. Conocimientos previos 7
2.1.1. Interoperabilidad 7
2.1.1.1. Interoperabilidad sintáctica 9
2.1.1.2. Interoperabilidad estructural 9
2.1.1.3. Interoperabilidad semántica 10
2.1.2. Ontologías 10
2.2. La heterogeneidad en la información geoespacial 12
2.2.1. Heterogeneidad en información geoespacial 13
2.2.2. Heterogeneidad en otros tipos de información 15
2.3. Una visión general de la problemática de la información geoespacial 21
2.3.1. Estructuración de la información geoespacial 22
2.3.2. Bases de datos geoespaciales 27
2.3.2.1. Datos 28
2.3.2.2. Diversidad de modelos 29
2.3.3. Factor escala 32
2.3.4. Lenguaje natural 33
2.3.4.1. Ambigüedad del lenguaje 34
2.3.4.2. Diferencias semánticas 35
2.3.4.3. Multilingüismo 35
2.4. Conclusiones 39

3. Estado de la cuestión 40

3.1 Ontologías geográficas 40
3.1.1. ¿Qué son las geo-ontologías? 40
3.1.2. Componentes 41
3.1.3. Tipologías de geo-ontolologías 42
3.1.4. Propuestas en el dominio geográfico 44
3.1.4.1. Ontologías geográficas 44
3.1.4.2. Ontologías hidrográficas 45
3.2. Integración de información 46
3.2.1. Marco teórico 46
3.2.2. El problema de la integración 48
3.2.3. Marco de integración de información 49
3.2.4. Modelado 50
Local as View (LAV) 50
Global as View (GAV) 51
Global and Local as View 51
Comparativa de los diferentes enfoques 51
3.2.5. Nuevas arquitecturas en el contexto de la integración de información 52
3.2.6. Propuestas generales de integración de información 52
3.2.7. Tipos de procesos de integración de información 53
3.2.7.1. Integración sintáctica o perspectiva clásica 53

3.2.7.2. Integración estructural 59
3.2.7.3. Integración semántica 68
3.3. Detección de duplicados 74
3.3.1. Un problema con múltiples tratamientos 75
3.3.2. Co-referenica en la Web de los Datos 77
3.3.3. Detección de duplicados en la información geoespacial 78
3.4. Linked Data 79
3.4.1. Linking Open Data 79
3.4.2. Los principios de la Web de Datos Enlazados 80
3.4.3. Linked Data geoespacial 81
3.5. Conclusiones 84

4. Planteamiento 87

4.1. Objetivos 87
4.2. Contribuciones previstas al estado de la cuestión 87
4.3. Premisas, hipótesis y limitaciones de la solución propuesta 89
4.3.1. Premisas 89
4.3.2. Hipótesis 90
4.3.3. Limitaciones 90

5. Metodología para la integración basada en ontologías de información de bases de datos heterogéneas en el dominio geográfico 93

5.1. Definición de metodología 93
5.2. Principios de diseño 95
5.3. Proceso de desarrollo. Descripción de alto nivel 96
5.4. Responsables implicados. Roles 99
5.5. Proceso de desarrollo. Detalles 100
5.5.1. Actividad 1. Estudio de viabilidad 100
5.5.2. Actividad 2. Análisis de fuentes 121
5.5.3. Actividad 3. Integración estructural 129
5.5.4. Actividad 4. Integración semántica 133
5.5.5. Actividad 5. Publicación de datos integrados 141
5.5.6. Actividad 6. Depuración / Post-procesamiento 146
5.5.7. Actividad 7. Modificaciones sobre fuentes de datos 152
5.5.8. Actividad 8. Mantenimiento 156
5.6. Conclusiones 157

6. Desarrollo de recursos ontológicos de dominio en el contexto de la información hidrográfica 160

6.1. Proceso de desarrollo 161
6.2. Contexto metodológico 162
6.3. hydrOntology 164
6.3.1. Especificación de requisitos 164
6.3.2. Preguntas de competencia 164
6.3.3. Fuentes de conocimiento 164
6.3.4. Modelado: Enfoque y criterios de estructuración 169
6.3.4.1. Enfoque 170
6.3.4.2. Criterios de estructuración 171
6.3.5. Características de hydrOntology 172
6.4. Red de ontologías en torno a hydrOntology 174
6.4.1. Especificación de requisitos 174
6.4.2. Preguntas de competencia 174
6.4.3. Fuentes de conocimiento 175
6.4.4. Escenarios 178
6.4.4.1. Reutilización de recursos ontológicos 178
6.4.4.2. Reutilización y reingeniería de recursos ontológicos 179
6.4.5. Características de la red de ontologías 182
6.4.6. Documentación y localización de recursos ontológicos 182
6.5. Conclusiones 185

7. Generación de instancias y publicación de Linked Data en el dominio hidrográfico 186

7.1. Identificación de datos 187
7.2. Generación de RDF 189
7.2.1. Recursos de información 189
7.2.1.1. hydrOntology 189
7.2.1.2. Bases de datos 189
7.2.2. Establecimiento de mappings 189
7.2.3. Generación de resultados 191
7.3. Generación de RDF de la información geométrica 194
7.3.1. Geometría: WKT, WKB y GML 195
7.3.2. Geometría y RDF 195
7.4. Revisión del RDF resultante 197
7.4.1. Análisis. Problemas detectados 197
7.4.2. Problemas solventados 200
7.5. Publicación de datos integrados 204
7.6. Conclusiones 208

8. Detección de instancias duplicadas en conjuntos de datos heterogéneos 210

8.1. Duplicidad de instancias 210
8.2. Problemas en el ámbito de las instancias 212
8.2.1. Nivel léxico 212
8.2.2. Nivel sintáctico 213
8.2.3. Nivel semántico 215
8.2.4. Nivel pragmático 217
8.3. Un acercamiento heurístico para la detección de duplicados 222
8.3.1. Planteamiento 222
8.3.2. Enfoques heurísticos 224
8.3.2.1. Detección de duplicados en un mismo conjunto de datos 225
8.3.2.2. Detección de duplicados en múltiples conjuntos de datos 228
8.4. Conclusiones 238

9. Evaluación 240

9.1. Recursos ontológicos 240
9.2. Metodología de integración de datos geoespaciales 243
9.3. Conjuntos de datos RDF y publicación de Linked Data 246
9.4. Detección de instancias duplicadas 248
9.4.1. Conjunto de pruebas 248
9.4.2. Medidas 250
9.4.3. Experimentación 250
9.4.3.1. Consideraciones previas 251
9.4.3.2. Análisis de resultados 251
9.5. Verificación de hipótesis 259
9.6. Conclusiones 260

10. Conclusiones y líneas futuras de investigación 262

10.1. Conclusiones 262
10.2. Líneas futuras de investigación 266

11. Bibliografía 268

12. Anexo 298

1. Contexto de información 298
2. Análisis de fuentes 301
Descripción de esquemas 301
Comparativa entre esquemas 303
Descripción de datos 304
Comparativa entre datos y revisión de expertos 305

Índice de figuras

Figura 2.1. Interoperabilidad 8
Figura 2.2. Diferentes tipos de heterogeneidad organizados en niveles 12
Figura 2.3. Heterogeneidad y contexto 13
Figura 2.4. Diferente representación geométrica de un mismo fenómeno 13
Figura 2.5. Dimensiones de la heterogeneidad a nivel conceptual 21
Figura 2.6. BTN200 25
Figura 2.7. BTN25…. 25
Figura 2.8. Interfaz de usuario del Getty Thesaurus of Geographic Names 26
Figura 2.9. Clasificación de tipos de ontologías 27
Figura 2.10. Extracto de la tabla de la base de datos del Nomenclátor Geográfico Conciso de España 30
Figura 2.11. Extracto del modelo de la base de datos de la Confederación Hidrográfica del Ebro 31
Figura 3.1. Distinción entre los tipos de conceptos que implican las ontologías geoespacial 42
Figura 3.2. Integración vertical y horizontal 42
Figura 3.3. Integración de datos 50
Figura 3.4. Propuestas generales para integrar información 52
Figura 3.5. Superposición de información en herramientas SIG 54
Figura 3.6. Modelo de referencia de la arquitectura de una IDE 55
Figura 3.7. Arquitectura común del proceso de integración de información 63
Figura 3.8. Componentes de la integración de datos espaciales en la web 66
Figura 3.9. Conjunto de datos pertenecientes a *Linking Open Data* 81
Figura 5.1. Representación gráfica de las relaciones terminológicas en las metodologías 94
Figura 5.2. Visión general de las diversas actividades de la propuesta metodológica y sus respectivas influencias 97
Figura 5.3. Ciclo de vida iterativo 98
Figura 5.4. Visión general de la propuesta metodológica para la integración basada en ontologías de información de bases de datos heterogéneas en el dominio geográfico 99
Figura 5.5. Descomposición de tareas de la Actividad 1 (Estudio de viabilidad) 103
Figura 5.6. Utilización del plugin ODEMapter para la generación de *mappings* 118
Figura 5.7. Descomposición de tareas de la Actividad 2 (Análisis de fuentes) 124
Figura 5.8. Descomposición de tareas de la Actividad 3 (Integración estructural) 130
Figura 5.9. Verificación de *hydrOntology* 133
Figura 5.10. Descomposición de tareas de la Actividad 4 (Integración semántica) 135
Figura 5.11. Resultados de la explotación de *mappings* 138
Figura 5.12. Descomposición de tareas de la Actividad 5 (Publicación datos integrados) 140
Figura 5.13. Desreferencia de un recurso (Conformidad con vocabulario asociado) 143
Figura 5.14. *Desreferencia* de un recurso 144
Figura 5.15. Representación HTML del recurso 145
Figura 5.16. Descomposición de tareas de la Actividad 6 (Depuración / Post-procesamiento) 148
Figura 5.17. Descomposición de tareas de la Actividad 7 (Modificaciones fuentes de datos) 153
Figura 6.1. Visión general del proceso de desarrollo de *hydrOntology* 161
Figura 6.2. Tareas de la actividad de conceptualización conformes a METHONTOLOGY 162
Figura 6.3. Escenarios para la construcción de ontología y redes de ontologías 163
Figura 6.4. Fenómenos en BCN25 167
Figura 6.5. Fenómenos en BCN 200 168
Figura 6.6. Fenómenos del Nomenclátor Geográfico Nacional 169
Figura 6.7. Fenómenos del Nomenclátor Conciso 169
Figura 6.8. Visión general del modelo de *hydrOntology* 173
Figura 6.9. Actividades de reestructuración ontológica 176
Figura 6.10. Niveles de abstracción para la actividad de Reingeniería Ontológica 178
Figura 6.11. Niveles de abstracción para la reingeniería ontológica 179
Figura 6.12. Modelo de alto nivel de la red de ontologías en torno a *hydrOntology* 184

Figura 6.13. Ejemplo de documentación extraído de *hydrOntology* 184
Figura 6.14. Ejemplo de localización extraído de *hydrOntology* 185
Figura 7.1. Estructura del patrón de URI adoptado para la identificación global de instancias 188
Figura 7.2. Estructura del patrón de URI adoptado para la identificación de instancias asociadas a una fuente de datos geográficos 189
Figura 7.3. Visión de alto nivel de los *mappings* entre *hydrOntology* y algunas de las bases de datos estudiadas 190
Figura 7.4. Escenario global de establecimiento de *mappings* entre *hydrOntology* y múltiples bases de datos para upgrade 191
Figura 7.5. Visión general de *mappings* establecidos con el plugin ODEMapster++ 192
Figura 7.6. Transformación de la información geométrica en RDF 195
Figura 7.7. Extracto del RDF geoespacial generado con GEOMETRY2RDF 196
Figura 7.8. Río Ebro y su información geométrica 197
Figura 7.9. Portal web de GeoLinked Data 206
Figura 7.10. Funcionamiento de Pubby 206
Figura 7.11. Visualización de recursos con Pubby 207
Figura 7.12. Visualización de los embalses de España 207
Figura 7.13. Visualización del fenómeno hidrográfico "Río" como *LineStrings* 208
Figura 8.1. Extracto de la tabla "Embalse" de la base de datos de la CHE 215
Figura 8.2. Extracto de la tabla "Rios" de la base de datos de la CHE 216
Figura 8.3. Detección de instancias duplicadas en un mismo conjunto de datos 225
Figura 8.4. Transformación de la codificación de las URI 228
Figura 8.5. Arquitectura desplegada para la detección de duplicados en múltiples conjuntos de datos 231
Figura 8.6. Visión general de la combinación de heurísticas 237
Figura 9.1. GeoLinked Data en la nube de *Linking Open Data* 247
Figura A.1. Visión general del contexto de las bases de datos estudiadas 299

Índice de tablas

Tabla 2.1. Tipos de heterogeneidad organizados conforme a sus caracterizaciones 14
Tabla 2.2. Extracto de tablas correspondientes a diferentes bases de datos con esquema dispar 16
Tabla 2.3. Problemática de la información geoespacial relacionada con tipos de heterogeneidad 24
Tabla 2.4. Fenómenos geográficos procedentes de diversas fuentes topográficas 29
Tabla 3.1. Características de las distintas ontologías hidrográficas 47
Tabla 3.2. Comparativa entre las diferentes propuestas de integración 70
Tabla 3.3. Características de los diferentes *Linked Data* geoespaciales 83
Tabla 5.1. Caso de uso de aplicación de la metodología 101
Tabla 5.2. Productos de tareas, técnicas y participantes de la Actividad 1 102
Tabla 5.3. Cuestiones para definir el dominio de trabajo 104
Tabla 5.4. Respuestas a las cuestiones para la definición del dominio de trabajo 107
Tabla 5.5. Cuestiones para la identificación del propósito 111
Tabla 5.6. Respuestas a las diversas cuestiones para la identificación del propósito 112
Tabla 5.7. Recopilación no exhaustiva de recursos disponibles dentro del dominio hidrográfico 113
Tabla 5.8. Listado de metadatos sobre las fuentes de información 115
Tabla 5.9. Cuestiones sobre el multilingüismo de las fuentes de información 115
Tabla 5.10. Metadatos sobre las fuentes de información 116
Tabla 5.11. Respuestas a las preguntas sobre la componente multilingüe de las fuentes de información 117
Tabla 5.12. Lista de comprobación para la decisión sobre la viabilidad del proyecto 119
Tabla 5.13. Lista de comprobación para la decisión sobre la viabilidad del proceso de integración del conjunto de las bases de datos 120
Tabla 5.14. Productos de tareas, técnicas y participantes de la Actividad 2 124
Tabla 5.15. Metadatos sobre los datos contenidos en las fuentes de información 128
Tabla 5.16. Errores léxicos 129
Tabla 5.17. Sugerencias de cambio 129
Tabla 5.18. Productos de tareas, técnicas y participantes de la Actividad 3 130
Tabla 5.19. Productos de tareas, técnicas y participantes de la Actividad 4 135
Tabla 5.20. Ejemplo de codificación de *mappings* R_2O entre *hydrOntology* y EuroGlobalMap 137
Tabla 5.21. Productos de tareas, técnicas y participantes de la Actividad 5 140
Tabla 5.22. Elementos de metadatos y su descripción asociada 142
Tabla 5.23. Productos de tareas, técnicas y participantes de la Actividad 6 148
Tabla 5.24. Extracto del sitemap de los datos integrados publicados 152
Tabla 5.25. Productos de tareas, técnicas y participantes de la Actividad 7 153
Tabla 5.26. Errores léxicos 154
Tabla 5.27. Identificación de multilingüismo 155
Tabla 5.28. Relación entre fenómenos asociados 155
Tabla 5.29. Presencia de fenómenos múltiples 155
Tabla 5.30. Diversidad de instancias en un único campo 155
Tabla 6.1. Preguntas de competencia de *hydrOntology* 164
Tabla 6.2. Recursos no ontológicos utilizados en el desarrollo de *hydrOntology*. 166
Tabla 6.3. Documento de especificación de requisitos de la ontología 174
Tabla 6.4. Preguntas de competencia de la red de ontologías 175
Tabla 6.5. Recursos ontológicos candidatos a reutilización en la red en torno a *hydrOntology* 177
Tabla 6.6. Recursos ontológicos reutilizados en la red de ontologías en torno a *hydrOntology* 178
Tabla 6.7. Recursos ontológicos sometidos a modificaciones en el desarrollo de la red de ontologías ... 179
Tabla 6.8. Ejemplo de reestructuración de jerarquías 180
Tabla 6.9. Ejemplo de compleción de conocimiento 180
Tabla 6.10. Ejemplo de profundización en el grano de la ontología 181
Tabla 6.11. Ejemplos de restricción del dominio 181
Tabla 6.12. Ejemplo de unificación de nombrado 181
Tabla 6.13. Ejemplo de unificación de idioma 182

Tabla 6.14. Listado de recursos y métricas de la red de ontologías....183
Tabla 7.1. Ejemplo de codificación de los *mappings* R_2O+....193
Tabla 7.2. Ejemplo de consulta ODEMQL....194
Tabla 7.3. Fragmento del GML generado....196
Tabla 7.4. Un ejemplo del WKT generado....196
Tabla 7.5. Ejemplo de problemas de duplicados....198
Tabla 7.6. Ejemplo de registros sin información....198
Tabla 7.7. Ejemplo de repetición del tipo de fenómeno....199
Tabla 7.8. Ejemplo de inclusión del fenómeno en el nombre de instancia....199
Tabla 7.9. Ejemplo de tripletas con correspondencias distintas entre el tipo de fenómeno y el nombre de instancia....199
Tabla 7.10. Ejemplo de tripletas fuera del dominio hidrográfico relacionadas con el fenómeno "Canal"200
Tabla 7.11. Ejemplo de información de otro dominio vinculada al dominio hidrográfico....200
Tabla 7.12. Ejemplo de modificación de la codificación de los *mappings* R_2O+....201
Tabla 7.13. Ejemplo de condición "match-regexp"....201
Tabla 7.14. Ejemplo de codificación para solventar los problemas de conectores....202
Tabla 7.15. Ejemplo de codificación para solventar los problemas de la ausencia del tipo de fenómeno202
Tabla 7.16. Ejemplo de combinación de las condiciones "match-regexp" y "equals"....202
Tabla 7.17. Ejemplo de uso de la condición "match-regexp" para problemas léxico-sintácticos....203
Tabla 7.18. Ejemplo de inclusión del tipo de fenómeno....203
Tabla 7.19. Ejemplo de elemento "attributemap-def" para la inclusión de nombres alternativos....203
Tabla 7.20. Recursos RDF con y sin espacios en blanco....204
Tabla 7.21. Extracto de tripletas RDF procedente del Nomenclátor Geográfico Conciso de España....204
Tabla 8.1. Descripciones similares para diferentes fenómenos....211
Tabla 8.2. Descripciones diferentes para un mismo fenómeno....211
Tabla 8.3. Diferenciación de caracteres (mayúsculas y minúsculas) entre dos bases de datos....213
Tabla 8.4. Lista de errores tipográficos y equivalencias en la base de datos de BCN200....214
Tabla 8.5. Ejemplo de nombres oficiales y/o alternativos en diferentes entradas de una tabla....217
Tabla 8.6. Ejemplo de nombres oficiales y/o alternativos en diferentes tablas de una base de datos....217
Tabla 8.7. Duplicidad de información presente en un extracto de la base de datos de EGM....218
Tabla 8.8. Extracto de la base de datos de ERM con información incompleta....219
Tabla 8.9. Problemas presentes en las instancias en el proceso de integración del conjunto de bases de datos estudiadas organizados en diferentes niveles....223
Tabla 8.10. Conjunto de métricas de similitud utilizadas....229
Tabla 8.11. Resultados métricas igualdad....229
Tabla 8.12. Conjunto de métricas de similitud utilizadas....230
Tabla 8.13. Resultados métricas similitud (umbral 0,96)....230
Tabla 8.14. Recuperación de latitud y longitud....234
Tabla 8.15. Recuperación de la geometría para el formato WKT....235
Tabla 9.1. Comparativa de las ontologías hidrográficas existentes con los ontologías desarrolladas....242
Tabla 9.2. Número de recursos y tripletas generadas en cada fuente....246
Tabla 9.3. Impacto de la iniciativa GeoLinked Data....248
Tabla 9.4. Comparativa de los diferentes *Linked Data* geoespaciales con GeoLinked Data....249
Tabla 9.5. Detección de duplicados en un mismo conjunto de datos....252
Tabla 9.6. Ejemplo de diferenciación de instancias duplicadas....252
Tabla 9.7. Resultados de la detección de duplicados a través de URI....253
Tabla 9.8. Resultados de la detección de duplicados utilizando *labels* y su idioma....253
Tabla 9.9. Resultados de la detección de duplicados mediante las métricas de *Silk*....254
Tabla 9.10. Resultados de la detección de duplicados mediante la comparación de URI y *labels*....255
Tabla 9.11. Resultados de la detección de duplicados mediante hWGS84 (heurísticas) y wgs84 (*Silk*)..255
Tabla 9.12. Resultados de la detección de duplicados mediante la combinación de las heurísticas URI, *Label* e idioma y URI*label*....256
Tabla 9.13. Resultados de la detección de duplicados mediante la combinación de las heurísticas URI, *Label* e idioma, URI*label* y hWGS84....257
Tabla 9.14. Resultados de la detección de duplicados con *Silk* mediante la combinación de métricas de similitud de cadenas y distancia geográfica....257
Tabla 9.15. Relación entre DBpedia y NGCE....258
Tabla 9.16. Relación entre DBpedia y NOMGEO....259
Tabla A.1. Listado de metadatos pertenecientes a la base de datos de BTN25....305

Capítulo 1

Introducción

Existen dos enfoques para la generación de nuevos conocimientos que están ampliamente reconocidos en la investigación científica y que se han considerado válidos durante cientos de años: la investigación experimental y la teórica. Con la aparición de los ordenadores en el siglo XX, aparece un tercer paradigma de investigación científica, que se hace principalmente visible a finales de los años 90. Este paradigma está basado en los aspectos computacionales de la Ciencia; por ejemplo, incorpora simulaciones de fenómenos naturales y de artefactos ingenieriles, todo tipo de experimentos *in-silico*, etc. Este tercer paradigma ha sido denominado como Ciencia computacional o e-Ciencia[1]. Finalmente, en los últimos dos años, y especialmente durante el año 2009, se ha producido un nuevo cambio en la relación entre la computación y la Ciencia, donde el foco ha pasado de centrarse en el impacto de la computación en la Ciencia, y viceversa, al impacto de los datos en la ciencia, en lo que se denomina el **Cuarto Paradigma** (Hey *et al.*, 2009) o la **Ciencia intensiva en datos**. Este paradigma ha emergido rápidamente como el cuarto pilar de la Ciencia, provocando que el ciclo de vida científico habitual, caracterizado por las actividades "crea una hipótesis, diseña y ejecuta el experimento, y analiza resultados", esté siendo reemplazado por el proceso "crea una hipótesis y busca la respuesta en la base de datos" (Lesk, 2004).

Si se considera la creencia no atribuida que afirma que aproximadamente el 80% de toda la información existente tiene un componente geoespacial, lo que implica que una proporción significativa de las bases de datos contienen información directa o indirectamente referenciada a una localización física (Hart *et al.*, 2007), el paradigma de la Ciencia intensiva en datos y la Información Geográfica (IG)[2] están especialmente relacionados.

1.1. De "islas de información" a datos integrados y abiertos

Durante la década de los años 1970 y principios de los 80, la mayoría de aplicaciones de los Sistemas de Información Geográfica (SIG) eran **islas de información**. Estos sistemas eran independientes entre sí y la información que contenían (datos espaciales) era capturada

[1] El término e-Ciencia cubre actualmente tanto el tercer (computacional) como el cuarto paradigma de la Ciencia (intensivo en datos).

[2] El *Open Geospatial Consortium* (OGC) (OpenGIS Consortium, 1999b) define la información geoespacial como la relativa a cualquier entidad que pueda ser aprehendida mirando en los mapas tradicionales, normalmente emitidos en papel, así como los más recientes, interactivos y anotados. El principal propósito de esta información es comunicar el conocimiento que se tiene sobre aquellas entidades que tienen una localización espacial. De esta manera, el conocimiento transmitido por esta información responde a dos tipos de preguntas; "dónde" y "qué".

digitalmente, almacenada, analizada y visualizada. Estos datos raramente se adquirían de fuentes digitales debido a la naturaleza cerrada de los formatos de archivo de las diferentes aplicaciones. Los avances en la tecnología de la información, la creciente demanda por parte de los usuarios para superar las limitaciones existentes y los costes de producción de los datos permitieron a los usuarios compartir datos transfiriéndolos de una *isla* a otra (Bishr, 1998).

El acceso y la utilización de la información geoespacial han cambiado de manera radical en las dos últimas décadas. Este cambio se produce por la constitución de un nuevo paradigma en el campo de la Geomática (Rodríguez *et al.*, 2006). Este nuevo paradigma supone una evolución conceptual y técnica promovida por el paso de los SIG a las Infraestructuras de Datos Espaciales[3] (IDE), consecuencia del impacto generado por la aparición de Internet. Uno de los puntos de partida para este cambio fue la **Orden Ejecutiva 12906** "***Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure***" del presidente de Estados Unidos (11 abril, 1994). Esta orden, impulsada por el Vicepresidente Al Gore y originaría del proyecto NSDI (*National Spatial Data Infrastructure*)[4], en su texto inicial reconoce la importancia de la información geográfica:

> "*La IG es crítica para promover el desarrollo económico, mejorar nuestra gestión de los recursos naturales y proteger el medio ambiente…"*

Asimismo, esta orden originaría una visión de un futuro digital en el que los ciudadanos del mundo podrían interactuar con un globo virtual tridimensional y acceder a grandes cantidades de información científica y cultural para ayudarles a entender sus actividades humanas y a la Tierra.

En Europa esta visión se traduce en el desarrollo de la **Directiva INSPIRE** (*INfrastructure for SPatial InfoRmation in Europe*) (Commission of the European Communities, 2007), cuyo principal objetivo es poner a disposición la información geográfica relevante y de buena calidad para apoyar la formulación, ejecución, seguimiento y evaluación de las políticas comunitarias que tengan dimensión o impacto territorial.

Estas iniciativas (la Orden Ejecutiva 12906 y la Directiva INSPIRE) se alinean con diversos proyectos, iniciativas y disposiciones legales que abogan por promover la libre circulación de la información y que reconocen el derecho de los ciudadanos a acceder a los datos que gestionan las instituciones públicas:

- Convenión de Aarhus (UNECE, 1998). Establece el derecho de los ciudadanos a acceder a la información gubernamental relativa al medio ambiente y a participar de alguna manera en la toma de decisiones que lo afecten.
- *Budapest Open Access Initiative* (Open Society Institute, 2001). Esta iniciativa está firmada por más de 200 universidades, bibliotecas y centros de investigación de todo el mundo. La iniciativa promueve la publicación electrónica de revistas de investigación para ofrecer mayor difusión, mediante mecanismos más rápidos y ágiles que las tradicionales revistas científicas, a los resultados y conclusiones de proyectos de investigación.
- Directiva Europea PSI (*Public Sector Information*) 2003/98/CE (European Parliament and the Council of the European Union, 2003). Esta directiva trata sobre la reutilización, por parte de la sociedad, de la información que gestiona el sector público.

[3] Infraestructuras de Datos Espaciales. Metadatos, conjuntos de datos espaciales y los servicios de datos espaciales; los servicios y tecnologías de red; los acuerdos sobre puesta en común, acceso y utilización (Commission of the European Communities, 2007).
[4] http://www.fgdc.gov/nsdi/nsdi.html

- *Open Archives Initiative*[5]. Esta iniciativa está dirigida a definir especificaciones estándar para hacer interoperables y fácilmente accesibles vía internet los archivos y catálogos de metadatos relativos a documentación científica. De esta manera, esta iniciativa contribuye a la libre circulación de la información.

Hoy día, los gobiernos y las instituciones gubernamentales de todo el mundo están incrementando la publicación de datos en Internet, especialmente en el contexto del denominado ***Open Data Movement***. Compartir estos datos permite una mayor transparencia, ofrece servicios públicos más eficientes y promueve un incremento del uso público y comercial de la información gubernamental. Algunos gobiernos han comenzado a crear catálogos o portales (por ejemplo, los gobiernos de Estados Unidos y Reino Unido) para facilitar al público el acceso y utilización de estos datos (W3C, 2009), disponibles en diferentes formatos (hojas de cálculo, *dumps* de bases de datos relacionales, RDF (*Resource Description Framework*) y con diversidad de contenido (por ejemplo: estadística, transporte, etc.).

1.2. Problemas asociados a la integración de datos geográficos

La necesidad de acceder e integrar recursos de datos (abstracciones de la realidad) creados por productores y usuarios resulta cada vez más acuciante, consecuencia del posicionamiento de la e-Ciencia como un área interdisciplinar (McGuinness *et al*., 2009). Sin embargo, los procesos de captura[6] de la información geográfica, influenciados por ciertas herencias, conllevan la generación, mantenimiento y actualización de información con diferente calidad, estructuración y formato, de forma individualizada e independiente, por cada productor.

Según Gómez-Pérez *et al*. (2008) existen dos puntos de vista principales sobre los problemas de heterogeneidad en la información geoespacial: uno más general y otro más específico. El primero, es decir, el punto de vista más global, supone que la IG es, cada vez más, capturada, gestionada y actualizada con variables niveles de detalle o especificidad (granularidad), calidad y estructura por diferentes productores cartográficos. En la práctica, este hecho provoca la construcción de múltiples conjuntos de bases de datos espaciales con gran heterogeneidad en las clasificaciones de conocimiento y modelos de datos asociados. Esta diversidad implica la coexistencia de una gran variedad de fuentes con diferente información, estructura y semántica sin un contexto general armonizador. Además, esta heterogeneidad combinada con la necesidad de compartir y superponer información, proveniente de diversas fuentes, por parte de diferentes usuarios conlleva varios e importantes problemas cuando se relacionan topónimos (instancias) similares en las tareas de consulta, recuperación, explotación, visualización y actualización de datos geoespaciales (Vilches-Blázquez *et al*., 2007a).

Desde un punto de vista más específico, el concepto más importante dentro del dominio geográfico es el de fenómeno (*feature*), también denominado entidad, elemento u objeto geográfico. Esta afirmación se sustenta en el hecho de que, según OGC (2003), el fenómeno geográfico (*geographic feature*) es el punto de partida para el modelado de la información geoespacial. Por esta razón, la unidad básica de la IG dentro de la mayoría de modelos es el fenómeno[7], es decir, una abstracción de un ente del mundo real asociado con una localización en la superficie terrestre, de la que son recogidos y mantenidos algunos de sus datos (ISO, 2005) para su posterior diseminación. Los fenómenos pueden incluir representaciones de un amplio rango de entes que pueden ser localizados en el tiempo y el espacio, tales como edificios,

[5] http://www.openarchives.org/OAI/OAI-organization.php

[6] Los procesos de captura de la información geoespacial, según (Bishr, 1998), se caracterizan por las siguientes fases: identificación de los fenómenos del mundo real de interés, asignación de nombre, categorización de fenómenos, creación de modelos y, finalmente, introducción de fenómenos e instancias relacionadas en una base de datos.

[7] *Feature. An abstraction of a real world phenomenon associated with a location relative to the Earth, about which data are collected, maintained and disseminated* (ISO, 2005).

ciudades y pueblos, redes geométricas, etc. Esto significa que, tradicionalmente, un fenómeno encierra en un concepto todo lo que un área temática considera sobre un sencillo fenómeno geográfico (Greenwood *et al*., 2003). Desde este punto de vista, se observa que la heterogeneidad asociada al concepto fenómeno aumenta como consecuencia de los intereses y necesidades de los diferentes productores de IG.

Ambos puntos de vista enmarcan múltiples problemas de heterogeneidad presentes en la información geoespacial. Asimismo, la complejidad y riqueza de los datos geográficos junto a la herencia de los procesos productivos presentan otras cuestiones específicas que dificultan la integración de información para conseguir la interoperabilidad semántica de la IG.

1.3. El uso de semántica como solución a los problemas

El papel de la semántica en la interoperabilidad e integración de datos heterogéneos, incluyendo a la información geoespacial, ha sido reconocido desde hace tiempo (Sheth, 1998; Goodchild *et al.*, 1999; Kuhn, 2005). La idea de proporcionar una integración semántica a los datos geoespaciales supone un paso hacia adelante en la posibilidad de superar los problemas producidos por la heterogeneidad presente en la información. Esta integración permite mezclar de algún modo sus propios resultados con los de las bases de datos públicas y/o con otros datos disponibles. Para ello, los resultados (datos) locales deben ser gestionados con suficientes metadatos como para facilitar una búsqueda y recuperación eficiente. Los resultados (datos) públicos deben ser accesibles en formatos interoperables con los datos locales para conseguir una integración eficiente y deben ser documentados utilizando vocabularios controlados que todos entiendan y que reflejen de manera precisa el significado de los datos. De esta manera, un usuario puede no poseer conocimientos de suficiente profundidad como para comprender la colección completa de los datos. Sin embargo, todas las fuentes de información se presentarán al usuario como una única fuente de datos, es decir, como una visión integrada de la realidad. El significado de los datos numéricos de los modelos teóricos, de los programas de análisis, etc., debe estar bien definido, de tal manera que los distintos usuarios puedan interpretar la información integrada desde estas fuentes distribuidas sin ser expertos en todas las áreas sobre las que los datos versan.

En esta línea, el reciente surgimiento de iniciativas como *Linked Data*, presentada como una de las mejores alternativas para la creación de estos espacios de información compartidos (Bizer *et al.,* 2009), resulta una alternativa que posibilita enriquecer la integración clásica de la información geográfica y, además, alcanzar la interoperabilidad semántica entre múltiples conjuntos de datos. La idea de *Linked Data* está vinculada a las mejores prácticas recomendadas para exponer, compartir y conectar conjuntos de datos en la Web Semántica (Bizer *et al.*, 2008). Estas prácticas están siendo adoptadas por un número creciente de productores de información, originando un espacio de datos global que contiene billones de datos, denominada la Web de los Datos (Bizer *et al.,* 2009). En este espacio de datos la información geográfica puede servir de referencia para la integración basada en ontologías (semántica) de información de múltiples dominios.

1.4. Propuesta de tesis

A pesar de la existencia de múltiples propuestas para la integración de información basada en ontologías, descritas en el Capítulo 3, hasta el momento no se han desarrollado metodologías que se puedan aplicar en este área. Este hecho conlleva, al menos, cuatro importantes limitaciones en los procesos de integración de información:

1. Las propuestas existentes no resultan fácilmente replicables.
2. No es fácil conocer las características de los enfoques propuestos.

3. Los escenarios de integración presentan fuentes de información similares y casos y problemáticas restringidas).
4. No se pueden realizar, de manera óptima, las tareas de planificación necesarias.

Estas carencias demuestran la necesidad de una propuesta metodológica que instruya el proceso de integración basado en ontologías de información geoespacial heterogénea. Por ello, con el objetivo de solventar las carencias detectadas, en este trabajo de tesis doctoral se llevarán a cabo las siguientes tareas:

En primer lugar, se propondrá una **metodología para la integración basada en ontologías de información de bases de datos heterogéneas en el dominio geográfico**. En esta metodología se identificarán las distintas las fases, actividades y tareas, así como los participantes implicados, las técnicas a utilizar y las entradas y salidas asociadas a cada tarea de la propuesta metodológica desarrollada. Asimismo, para demostrar la aplicabilidad de esta metodología se llevará a cabo un caso práctico de integración en el dominio hidrográfico.

En el proceso de aplicación de esta metodología se desarrollarán **recursos ontológicos de dominio y multilingües en el área de la información hidrográfica**. Estas ontologías serán utilizadas para la creación de *wrappers*[8] entre dichos recursos y las diversas fuentes de información (bases de datos) estudiadas. Este trabajo permitirá la **generación de diversos conjuntos de datos RDF**, donde se tratarán e incluirán las **características geométricas de la información geoespacial** a través de la incorporación de *Geography Markup Language* (GML) en el RDF generado.

Tras la generación del RDF geoespacial se procederá a la **publicación de la información hidrográfica** del territorio nacional español **conforme a los principios de *Linked Data***. Esta información permitirá el enriquecimiento de la Web de los Datos con información geoespacial en español.

Finalmente, se propondrán **diversos enfoques heurísticos para la detección de instancias duplicadas** en conjuntos de datos RDF heterogéneos y como resultado de dicha detección se formalizarán relaciones (*owl:sameAs*) entre las instancias duplicadas detectadas.

La aplicación de este trabajo al dominio de la información hidrográfica se debe, fundamentalmente, a la importancia vital del agua como recurso. En relación con esto, destacan dos escenarios evidentes. Por un lado, desde una óptica general, es un hecho innegable que la superficie de nuestro planeta (el planeta azul) está compuesta por ¾ partes de agua. Por otro, desde un punto de vista más concreto, los problemas de estiaje en nuestro país requieren una mejor gestión de este recurso.

Desde una óptica más relacionada con la información, la información hidrográfica forma parte de las capas que conforman el sustrato cartográfico de referencia. Este hecho se ve consolidado por la presencia de dicha información entre los datos de referencia de INSPIRE (Infrastructure for Spatial Information in the European Community) (European Parliament *et al*., 2007) (recogida en el Anexo de esta Directiva). Además, desde la perspectiva legal, hay que añadir la importancia de la Directiva Marco del Agua (European Parliament *et al*., 2000), cuyo objetivo principal consiste en el establecimiento de un marco para la protección comunitaria de las aguas superficiales continentales, aguas de transición, aguas costeras y aguas subterráneas.

[8] Los *wrappers* actúan como interfaces de cada fuente de datos, proporcionando una (semi) estructura a aquellas fuentes no estructuradas o bien mapean la estructura de datos original en la búsqueda de un patrón común.

1.5. Estructura de la tesis

Esta tesis se organiza de la siguiente manera:

- El **capítulo 2** se inicia con la introducción de una serie de conocimientos previos relativos al concepto de interoperabilidad y con algunas nociones generales del concepto ontología. A continuación, se presenta un análisis de los tipos de heterogeneidad en la información geográfica y una visión general de los principales problemas de la información geoespacial.

- El **capítulo 3** recoge el estado de la cuestión de los diversos temas de interés en este trabajo. Así, se realiza un análisis de las ontologías desde su vinculación con el dominio geográfico y se revisan un conjunto significativo de ontologías geoespaciales. Se continúa con una descripción de las características fundalmentales del proceso de integración de información y se recogen diversos tipos de proceso de integración. También se describen distintos tratamientos y enfoques para la detección de duplicados y se presenta una visión general de la iniciativa *Linking Open Data* y las propuestas existentes de *Linked Data* geográfico. Finalmente, se exponen unas conclusiones y se recogen diferentes limitaciones y problemas abiertos en los distintos temas tratados, enfatizando aquellos aspectos sobre los que este trabajo aporta solución.

- El **capítulo 4** enuncia los objetivos y las contribuciones de este trabajo, definiendo asimismo las premisas, hipótesis y limitaciones del mismo.

- El **capítulo 5** describe en detalle la metodología propuesta para la integración basada en ontologías de información de bases de datos heterogéneas en el dominio geográfico.

- El **capítulo 6** expone el desarrollo de recursos ontológicos de dominio en el contexto de la información hidrográfica. Este proceso conlleva la construcción de una ontología global de dominio y, con posterioridad, el desarrollo de una red de ontologías. Este capítulo recoge los cambios metodológicos y de enfoque realizados.

- El **capítulo 7** describe el proceso de generación de instancias (RDF) para el dominio hidrográfico, diversos problemas detectados en el RDF resultante y las soluciones a dichos problemas, y el proceso de publicación de los conjuntos de datos generados como *Linked Data*.

- El **capítulo 8** presenta los diferentes enfoques heurísticos desarrollados para la detección de instancias duplicadas en el contexto de conjuntos de datos geoespaciales.

- El **capítulo 9** recoge la evaluación de trabajo realizado. En primer lugar se presenta la evaluación de la metodología desarrollada en este trabajo a través de los principios de diseño considerados y de la aplicabilidad de la misma. Posteriormente, se realiza la evaluación de una muestra de los conjuntos de datos RDF de las distintas fuentes de información. Finalmente, se evalúan los diversos enfoques heurísticos desarrollados y se realiza una comparación de dichos resultados con los obtenidos por el *framework Silk*. La evaluación realizada se basa, principalmente, en las medidas de *precision* (precisión), *recall* (exhaustividad) y Medida E o Medida Armónica.

- Finalmente, el **capítulo 10** recoge las conclusiones de este trabajo, haciendo especial énfasis en sus aportaciones y en las hipótesis presentadas en el Capítulo 4. Además se destacan las futuras líneas de investigación abiertas como continuación de esta tesis.

Capítulo 2

Análisis de la heterogeneidad en la información geográfica

En este capítulo se analizan los tipos de heterogeneidad existentes en la información geográfica y se presenta una visión general de los principales problemas de la información geoespacial, cuestiones que se van a abordar en esta tesis. Previamente, se va a detallar de forma breve un conocimiento previo base para la comprensión y detalle de este trabajo.

El capítulo se estructura de la siguiente manera: La sección 2.1 introduce una serie de **conocimientos previos** relativos al concepto de interoperabilidad y algunas nociones generales sobre un componente fundamental en el proceso de integración semántica: las ontologías. En la sección 2.2 se analizan los diferentes **tipos de heterogeneidad** presentes en la información geográfica y en otros tipos de información (bases de datos y modelos de conocimiento/ontologías). La sección 2.3 presenta una visión general de los principales **problemas de la información geoespacial**, utilizando los tipos de heterogeneidad para contextualizarlos. Finalmente, la sección 2.4 recoge las conclusiones de este capítulo.

2.1. Conocimientos previos

En esta sección se introduce una serie de conocimientos previos relacionados con el concepto de interoperabilidad, describiendo los tres tipos de interoperabilidad más extendidos en la bibliografía científica (sintáctica, estructural y semántica), y algunas nociones generales sobre un componente fundamental en el proceso de integración semántica: las ontologías.

2.1.1. Interoperabilidad

Una de las motivaciones principales para el desarrollo de los diferentes Sistemas de Información Geográfica, también conocidos como Sistemas de Información Territorial, e Infraestructura de Datos Espaciales es hacer más eficiente el trabajo con los geodatos (McKee, 2000; Nerbert, 2001). Para facilitar la comunicación entre estos sistemas se crean los estándares y especificaciones del Comité Técnico 211 de la Organización Internacional para la Estandarización (ISO/TC 211)[9], *World Wide Web Consortium*[10] (W3C) y del *Open Geospatial Consortium*[11] (OGC) con los que se resuelven problemas derivados de la diferencia de formatos de información, consiguiendo interoperabilidad. A continuación, se reflejan las definiciones más extendidas de este concepto;

[9] *ISO/TC 211*. http://www.isotc211.org/
[10] *W3C*. http://www.w3.org/
[11] *OGC*. http://www.opengeospatial.org/

"*La habilidad de dos o más sistemas o componentes de intercambiar información y utilizar la información que ha sido intercambiada*"[12] (Institute of Electrical and Electronics Engineers (IEEE), 1990).

"*La capacidad de comunicarse, para ejecutar programas, o para transferir datos entre varias unidades funcionales de una manera que requiera al usuario tener poco o nada de conocimiento de las características únicas de esas unidades*"[13] (ISO 2382-1, 1993; ISO 19119, 2002). En este sentido, dos componentes X e Y (ver Figura 2.1) pueden interoperar (son interoperables) si X puede enviar peticiones R de servicios a Y, basadas en un mutuo entendimiento de R por X e Y, y si Y puede, de forma similar, devolver mutuamente entendible respuestas S a X (ISO 19119, 2002).

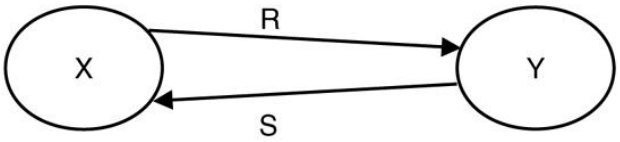


Figura 2.1. Interoperabilidad (ISO 19119, 2002)

La noción de interoperabilidad involucra un proceso de comunicación entre dos agentes (por ejemplo, máquinas y/o seres humanos) que utilizan su propia terminología para abstraer su percepción subjetiva del mundo (Brodeur *et al.*, 2001). Basándose en la analogía de la comunicación humana, se puede observar que las diferentes visiones del mundo y el uso de los diferentes términos pueden conllevar serios problemas cuando los significados intuitivos de los elementos de una terminología no están claros y el entendimiento del proveedor de información difiere del entendimiento del usuario (Nowak *et al.*, 2005).

No obstante, la noción de interoperabilidad necesita ser entendida en un sentido más amplio (Kuhn, 2005). Bajo esta perspectiva se encuentran las diferentes concepciones y/o niveles de interoperabilidad recogidas en (Bishr, 1998; Sheth, 1998; European Commision, 2004; Tolk *et al*., 2007, Christaki *et al.*, 2007; Nilsson *et al*., 2008; Manso *et al.*, 2009). Según Kuhn (2005) la idea de interoperabilidad no debe comprender únicamente una base sintáctica común, sino que además debe contemplar las cuestiones sociales y organizacionales involucradas en la compartición de la información (Harvey *et al*., 1999) para conseguir la interoperación entre personas y sistemas. Asimismo, desde el punto de vista de la información, la capacidad de interoperar entre dos sistemas se debe a que ambos poseen un modelo de información interoperable en un sentido amplio. Para alcanzar la denominada interoperabilidad *sensu stricta* (Vilches *et al*., 2006) el modelo de información de ambos sistemas debe ser interoperable de forma sintáctica, estructural y semánticamente (Sheth, 1998; Jang *et al.*, 2007). Ciñéndose a estos tipos de interoperabilidad y asemejando los mismos a la división de interoperabilidad propuesta por (Hasselbring, 2000), a continuación, este trabajo se centrará en describir los tres tipos mencionados de interoperabilidad (sintáctica, estructural y semántica) que, posteriormente, serán relacionados con homónimos procesos de integración de información.

En esta sección se recogen diferentes perspectivas del concepto interoperabilidad. A continuación, se describe la consideración de la interoperabilidad sintáctica (sección 2.1.1.1), estructural (2.1.1.2) y semántica (2.1.1.3).

[12] "*the ability of two or more systems or components to exchange information and to use the information that has been exchanged*".
[13] "*Interoperability is the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units*" (ISO, 1993; ISO, 2002).

2.1.1.1. Interoperabilidad sintáctica

La rápida evolución tecnológica y la amplia variedad de elecciones para implementar cada sistema pueden llevar a un alto grado de heterogeneidad entre los sistemas de información, incluso aunque estos tengan el mismo dominio de aplicación e idéntico universo de usuarios en mente (Fileto, 2001). Ante esta situación la necesidad de interoperabilidad, concretamente, a nivel sintáctico se pone de manifiesto.

El punto de vista de la implementación (Hasselbring, 2000) o interoperabilidad sintáctica, asegura la existencia de conexión técnica, es decir, que los datos pueden ser transferidos entre diferentes sistemas. Esto posibilita el intercambio de información entre sistemas derivado de la utilización de formatos o estructuras comunes de datos, lenguajes, lógica, registros y ficheros. Se incluyen como aspectos que permiten la interoperabilidad sintáctica los estándares o especificaciones de formatos que estructuran la información de modo que pueda ser procesada e interpretada a ambos extremos de los sistemas. Entre estos estándares y/o especificaciones destaca el lenguaje XML (*eXtensible Markup Language*) junto con los esquemas de datos (XML Schema o DTD (*Document Type Definition*)). Este lenguaje se establece como una vía para proporcionar una sintaxis común en los procesos de intercambio e integración de información heterogénea (Cui *et al.*, 2001; Gaun *et al*., 2003) y, por tanto, puede considerarse un elemento fundamental que permite alcanzar la interoperabilidad sintáctica. Asimismo, en este marco se encuentran especificaciones y estándares que definen los esquemas de intercambio de información y el significado de cada uno de los elementos sin ambigüedades, tales como; *Geographic Markup Language* (GML) (Portele, 2007; ISO 19136, 2007) en el transporte de IG vectorial, *Style Layer Description* (SLD) (Open GIS Consortium (OGC), 2002) en la definición de un estilo de visualización, *Common Query Language* (CQL) o *Filter Encoding*, (FE) (Open GIS Consortium (OGC), 2001) para las consultas y los filtros, etcétera.

En un contexto menos técnico y más próximo al dominio geográfico surgen dos definiciones distintas de interoperabilidad a nivel sintáctico. Por un lado, una definición más general y, por tanto, de aplicación global al mencionado dominio como es la definición de interoperabilidad geográfica (ISO, 2002). Según ISO esta interoperabilidad "*es la habilidad de los sistemas de información de 1) intercambiar libremente todo tipo de información espacial sobre la Tierra y sobre los objetos y fenómenos sobre, encima y debajo de la superficie de la Tierra; y 2) cooperativamente, a través de las redes, correr software capaz de manipular tal información*". Por otro lado, una visión centrada en el ámbito de las IDE, como la propuesta por (The European Parliament and Council of the EU, 2007), definida como "*la posibilidad de combinación de los conjuntos de datos espaciales y de interacción de los servicios, sin intervención manual repetitiva, de forma que el resultado sea coherente y se aumente el valor añadido de los conjuntos y servicios de datos*".

2.1.1.2. Interoperabilidad estructural

En un nivel intermedio entre la interoperabilidad sintáctica y la semántica se encuentra la interoperabilidad estructural. Este tipo de interoperabilidad está relacionada con las diferencias en los esquemas de las diferentes fuentes. Dichas diferencias se ponen de manifiesto con la existencia de disparidad de componentes entre diversos esquemas. Por tanto, este tipo de interoperabilidad es la encargada de reconciliar los problemas relativos a la heterogeneidad estructural.

La interoperabilidad estructural (Shekhar, 2004; Friis-Christensen *et al*., 2005) se va a encargar de proporcionar los medios para especificar esquemas comunes (o metadatos) para compartir. De esta manera, esta interoperabilidad va a proporcionar un nivel básico de transformación de un esquema a otro. El objetivo de este tipo de interoperabilidad, denominada punto de vista conceptual en (Hasselbring, 2000), es mejorar el entendimiento del alcance de las aplicaciones y sus características para descubrir las formas más apropiadas de modelar fenómenos y acciones en una base de datos.

Por otro lado, la situación intermedia, mencionada con anterioridad, de este tipo de interoperabilidad se muestra como un factor clave para la consecución de la interoperabilidad sintáctica y semántica. Así pues, la utilización de modelos estructurales comunes en los sistemas de información conllevan interoperabilidad sintáctica entre los mismos (ISO, 2002). Asimismo, para la consecución de la interoperabilidad semántica de los diferentes tipos de fenómenos existentes en los variados modelos es necesario establecer correspondencias entre las definiciones de los múltiples tipos de fenómenos existentes en las fuentes de información (ISO, 2002).

2.1.1.3. Interoperabilidad semántica

Los expertos de dominio poseen propios acercamientos conceptuales, terminológicos e incluso representaciones de datos elegidas para modelar la realidad y sus problemas (Hasselbring, 2000). A esto se le une el hecho de que XML no es suficiente para determinar la semántica asociada a los datos (Cui *et al.*, 2001). Esta manifiesta heterogeneidad del entendimiento, incluso dentro de un mismo dominio, y las carencias de expresividad semántica de ciertos lenguajes ponen en evidencia la creciente necesidad de consecución de la interoperabilidad semántica.

La interoperabilidad semántica (Sheth, 1998; Ouksel *et al.*, 1999) está encargada de asegurar que el contenido de la información (significados) intercambiada sea entendido de la misma manera por cualquier sistema. Por tanto, dos sistemas son semánticamente interoperables si éstos tienen un entendimiento común de la semántica de la información, permiten combinar información con otros recursos y transmitirlos y procesarlos por los sistemas de una manera significativa. Considerando estas características resulta evidente que la interoperabilidad semántica es la forma más útil de interoperabilidad. Esta afirmación es resultado de la dificultad de imaginar, en el mundo real, a dos agentes interoperando de manera satisfactoria sin un entendimiento compartido de los mensajes que intercambian (Kuhn, 2005).

La consecución de éste entendimiento *colectivo* se centra en la utilización de una conceptualización común y compartida, descrita de manera explícita, con objeto de evitar inexactitudes o confusiones en la interpretación del significado de los términos y en los procesos de intercambio de información. Estas son las características de las ontologías. Una visión general de este concepto se recoge en la sección 2.1.2 de este capítulo.

En el contexto de la información geoespacial, la interoperabilidad semántica se define como la capacidad de los sistemas y servicios de información geográficos de trabajar juntos sin la necesidad de intervención humana (Harvey *et al.*, 1999). Alcanzar la interoperabilidad semántica es uno de los objetivos más importantes para las comunidades interesadas en la IG. En este contexto, (Egenhofer, 2002) señala la necesidad de definir la denominada *Web Semántica Geoespacial*, basada en un marco que comprenda múltiples ontologías temáticas y espaciales, así como una forma canónica para especificar consultas geoespaciales.

La utilización de ontologías en el dominio geográfico ayuda a definir los significados de los fenómenos contenidos en los geodatos, pudiendo proporcionar la base del entendimiento de la IG. De esta manera, las ontologías contribuyen a hacer efectiva la consecución de una visión común, compartida e interoperable de los fenómenos presentes en la realidad geográfica, eliminando obstáculos para su utilización, consulta y/o asignación de valor añadido.

2.1.2. Ontologías

El origen del término ontología procede del mundo de la filosofía. En este área, las ontologías tienen su origen en la Antigua Grecia (con pensadores como Parménides o Aristóteles), aunque el tratamiento de este concepto avanza hasta el siglo XX (tratado por filósofos como Kant y Ortega y Gasset entre otros). Desde este punto de vista, la ontología se define como una explicación sistemática del ser. Vinculado a este concepto se distingue entre

los términos *esencia*[14] y *existencia*[15]. En la década de los 90 este concepto comienza a ser relevante para la comunidad de la Ingeniería del Conocimiento. Su adopción ha dado origen a una gran variedad de definiciones de este "nuevo" concepto. Estas definiciones proporcionan diferentes y complementarios puntos de vista de la misma realidad. Además, algunas de ellas son independientes del proceso de desarrollo y de su utilización en aplicaciones, mientras que otras son influenciadas por sus procesos de desarrollo.

Una de las definiciones más divulgadas de este concepto es la aportada por (Gruber, 1993). Afirma que una ontología constituye una especificación explícita y formal de una conceptualización compartida. Esta definición tiene una serie de conceptos claves que ayudan a comprenderla. Así, la *conceptualización* se entiende como un modelo abstracto de la realidad, de tal manera, que mediante ésta se identifican los conceptos relevantes de un área. Por *explícita* se entiende que todos sus componentes deben estar definidos explícitamente. Respecto al concepto *formal*, se refiere al hecho de que la ontología debe ser entendible por las máquinas. Por último, *compartida* refleja el hecho de que una ontología debe capturar conocimiento consensuado/aceptado por un grupo o comunidad de expertos, esto es, no debe ser privado o algo individual (Studer *et al*., 1998).

En lo que respecta a su utilización, en los sistemas de información el propósito principal de una ontología es proporcionar una especificación rigurosa para la interpretación de conceptos en los términos en que se almacenan los datos. Esta especificación puede ser utilizada para comprobar y mantener la integridad de una base de datos, para facilitar la integración y funcionamiento entre diversas bases de datos. Además, puede ser utilizada como una base para la implementación de funcionalidad para realizar consultas y presentar información que requiera flexibilidad, diversidad de perspectivas y acceso a datos.

La principal conclusión de la definición del concepto ontología consiste en apuntan a capturar el conocimiento consensual de una manera genérica y que estas pueden ser reutilizadas y compartidas a través de aplicaciones software y por grupos de personas. Asimismo, las ontologías son, normalmente, construidas de forma cooperativa por diferentes grupos de personas con diferente ubicación. Una recopilación de las múltiples definiciones del concepto ontología y la descripción de sus componentes principales aparece en (Gómez-Pérez *et al*., 2003).

Con respecto a la tipología de ontologías, en un proceso de abstracción de la clasificación presentada en la Figura 2.9, la comunidad ontológica distingue dos tipos básicos de ontologías. Por un lado, se encuentran aquellas que son principalmente taxonomías, denominada ***ontologías ligeras***[16]. Y, por otro lado, las ontologías que modelan el dominio de una manera más profunda proporcionando axiomas y restricciones sobre la semántica del dominio, este tipo de ontología recibe el nombre de ***ontologías pesadas***[17]. Asimismo, consecuencia del desarrollo colaborativo y distribuido de ontologías y de las relaciones entre las mismas se distinguen los siguientes tipos de ontologías (Suárez-Figueroa *et al*., 2008):

[14] La *esencia* es aquello que realmente constituye la naturaleza de las cosas, es decir, lo permanente e invariable.
[15] La *existencia* se define como la presencia entre las cosas del mundo real.
[16] *Ontologías ligeras* (*lightweight ontologies*). Este tipo de ontologías encuadran aquellas ontologías que son fundamentalmente taxonomías. Las ontologías ligeras se componen de conceptos, taxonomías de conceptos, relaciones entre conceptos y propiedades que describen conceptos.
[17] *Ontologías pesadas* (*heavyweight ontologies*). Este tipo de ontologías son una extensión de las ontologías ligeras, a las que se les añaden axiomas (sentencias que son siempre ciertas) y restricciones. Estos axiomas y restricciones sirven para completar el significado de los términos que componen las ontologías de forma que se puede explicitar más rigurosamente su semántica.

- ***Ontologías individuales***. Se tiene una ontología individual cuando una ontología no tiene ningún tipo de relación (dependiente o independiente del dominio) con otras ontologías.
- ***Ontologías individuales interconectadas***. Este tipo de ontologías se caracterizan por poseer entre ellas algún tipo de relación ad-hoc dependiente del dominio.
- ***Redes de ontologías***. Son una colección de ontologías relacionadas entre sí mediante diferentes relaciones como correspondencia (*mapping*), modularización, versionado y/o dependencia (Haase *et al*., 2006).

Otra tipología de ontologías, menos abstracta que la propuesta anterior y ampliamente reconocida en el área de la Ingeniería Ontológica, es la compuesta por: ontologías globales o alto nivel (*top level ontologies*), dominio (*domain ontologies*) y aplicación (*application ontologies*).

- Una ***ontología global o de alto nivel*** define los conceptos generales que son independientes de un dominio específico, por ejemplo el concepto espacio o tiempo.
- Una ***ontología de dominio*** define los conceptos que son específicos a un área de conocimiento (por ejemplo, geología, morfología-relieve, vías de comunicación, etc.).
- Una ***ontología de aplicación*** (*application ontologies*) define los conceptos que están especializados dentro de un contexto determinado o un uso específico (por ejemplo, envío de ambulancias, gestión de procesos hidráulicos, etc.).

Al igual que existen diversas definiciones de lo que es una ontología, existen múltiples clasificaciones de estas, ya sea basándose en su estructura interna o en el sujeto de la conceptualización. Cada autor clasifica las ontologías de acuerdo a la definición de ontología que toma como base para su trabajo. Así, actualmente, se tienen, entre otras, las clasificaciones de (Mizoguchi *et al*., 1995), (van Heijst *et al*., 97), (Guarino, 1998) y (Lassila *et al.*, 2001).

Una visión más detallada de este concepto, sus tipologías y otras cuestiones vinculadas con la Ingeniería Ontológica se recopilan en (Gómez-Pérez *et al*., 2003).

2.2. La heterogeneidad en la información geoespacial

Uno de los aspectos principales que se abordan en este trabajo es la heterogeneidad. En un sentido amplio, la heterogeneidad, según la Real Academia Española, se define como la "*mezcla de partes de diversa naturaleza en un todo*". En los sistemas abiertos y distribuidos utilizados en el contexto de los datos geoespaciales, como las Infraestructuras de Datos Espaciales, la heterogeneidad no puede ser evitada, consecuencia de que diferentes actores tienen distintos intereses y costumbres, manejan diversas herramientas y utilizan el conocimiento a diferentes niveles de detalle.

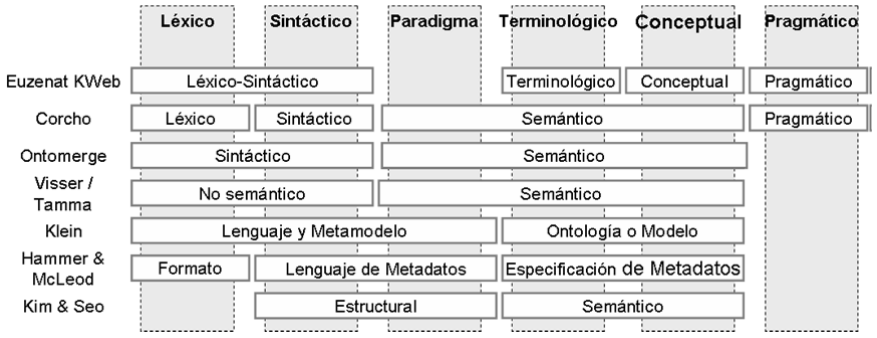


Figura 2.2. Diferentes tipos de heterogeneidad organizados en niveles (Barrasa, 2007)

Desde una perspectiva de alto nivel, la heterogeneidad puede adoptar diferentes formas que en la bibliografía científica han sido tratadas ampliamente y organizadas en diversos tipos o niveles. A modo de ilustración de las variadas clasificaciones de disparidades existentes, cabe citar la recopilación y puesta en conjunto de (Barrasa, 2007), resumida en la Figura 2.2.

En esta sección se describen los principales enfoques y clasificaciones de los diferentes tipos de heterogeneidad contemplados en la bibliografía relacionada con la información geoespacial, bases de datos y modelos de conocimiento/ontologías. Con objeto de ofrecer un primer acercamiento se enmarca este ámbito de estudio dentro de algunos tipos de heterogeneidad recogidos en (Sheth, 1998). La Figura 2.3 muestra gran parte de los distintos tipos de heterogeneidad estudiados en esta sección y delimita el contexto de trabajo.

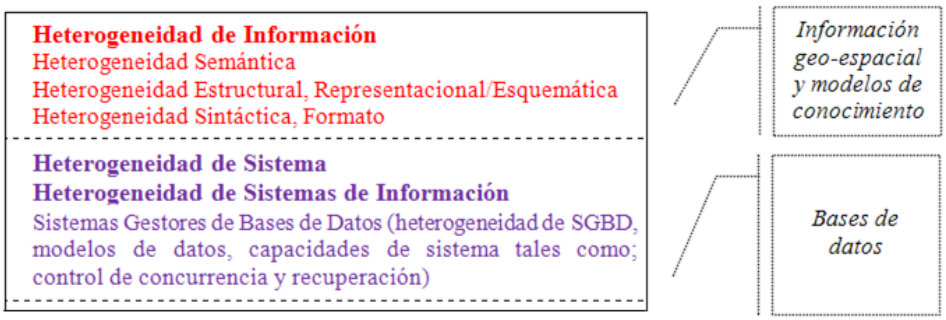


Figura 2.3. Heterogeneidad y contexto. Figura adaptada de Sheth (1998)

Asimismo, con objeto de aportar una visión general, la Figura 2.3 muestra las clasificaciones propuestas en los acercamientos recopilados y agrupa los tipos de heterogeneidad en función de su vinculación con el área de estudio.

2.2.1. Heterogeneidad en información geoespacial

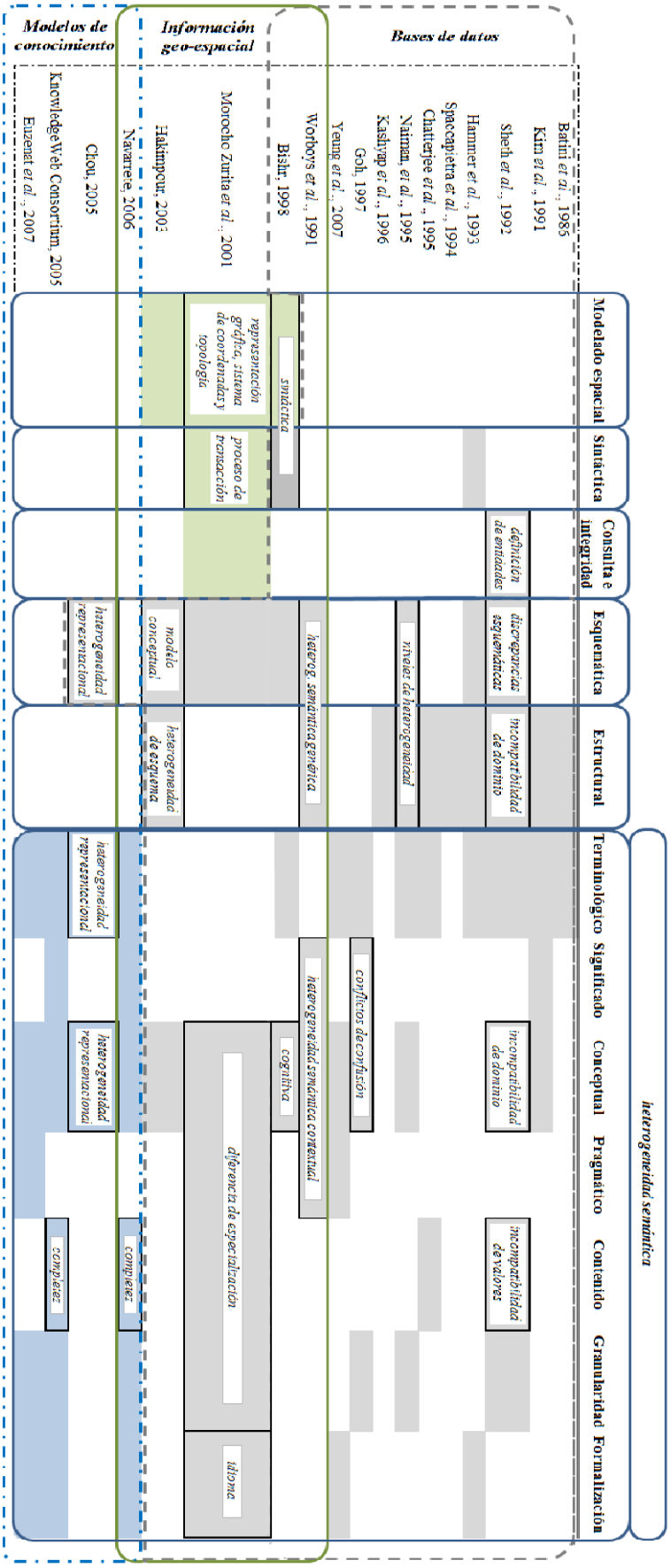
En (Fielding *et al*., 2004) se afirma que la heterogeneidad es debida a visiones específicas del mundo. De esta manera, cualquier diferencia en las fuentes de datos, entre disciplinas, herramientas y/o repositorios de datos puede provocar heterogeneidad en el dominio de la IG (Alonso *et al*., 1994). Estas variadas razones para la heterogeneidad conducen a diferentes tipos de heterogeneidad. A continuación, se presenta un conjunto representativo de diferentes tipos de heterogeneidad comúnmente distinguidos en la bibliografía científica.

Escala 1:25.000 *Escala 1:200.000* *Superposición*

Figura 2.4. Diferente representación geométrica de un mismo fenómeno

La **heterogeneidad en el modelado espacial** (Hakimpour, 2003) o **sintáctica** (Bishr, 1998) está relacionada con la representación geométrica de los fenómenos geográficos en las bases de datos. Así un tipo de fenómeno podría estar representado por un polígono (o un sector de píxeles) en un sistema, mientras que este tipo de fenómeno, en un segundo sistema, podría estar representado por líneas. Un ejemplo de la disparidad representacional comentada aparece en la Figura 2.4.

Tabla 2.1. Tipos de heterogeneidad organizados conforme a sus caracterizaciones



En el contexto de la integración de fuentes heterogéneas de datos espaciales, (Morocho Zurita *et al*., 2001) se plantean diversos problemas referidos a la **heterogeneidad semántica**. Para estos autores, los datos pueden ser interpretados de forma diferente en sus diversos componentes; considerando que el inicio de las bases de datos es diferente, la posibilidad de diferencia en la interpretación para la representación de datos espaciales crece. Partiendo de esta premisa destacan el siguiente listado de problemas:

- *Diferencia en la especialización*. Un productor especializado en demarcación territorial dará más prioridad a los límites y, por tanto, las muestras topográficas serán tomadas con mucha más precisión escalar que otro productor especializado, por ejemplo, usos del suelo.
- *Diferencia en los sistemas de coordenadas*. La ubicación de un punto sobre la Tierra puede expresarse por medio de sus coordenadas utilizando un sistema de coordenadas u otro. Cada sistema de coordenadas comprende un Sistema de Referencia Geodésico, como WGS84 (*World Geodetic System* 1984) y una proyección cartográfica, como UTM (*Universal Tranverse Mercator*) o la utilización directa de la latitud y la longitud. De la misma manera, pueden existir diferencias para la localización de un punto en el espacio al referirse al eje Z (altura) dependiendo de la superficie de referencia
- *Diferencia en representación gráfica*. El seguimiento o no de alguna norma de simbología o representación introduce otro factor de heterogeneidad importante, que puede dar lugar a distintas interpretaciones de la semántica de la información.
- *Diferencia en topología*: Al pretender integrar información procedente de diversas fuentes hay que tener en cuenta el modelo topológico (espagueti, cadena-nodo, grafoplanar, grafo no planar, topología completa,...) y las diferentes relaciones topológicas establecidas entre los objetos (por ejemplo: dentro, incluido en, externo a, etc.).
- *Diferencia en la concepción del objeto a representar*: Dicha concepción dependerá del punto de vista y el contexto de cada productor de información geoespacial. Por ejemplo no es lo mismo el concepto de carretera en Europa que en Sudamérica o en Asia.
- *Diferencia en el idioma de los datos representados*: Dado que la IG no entiende de fronteras ésta debe soportar pluralidad de idioma.

2.2.2. Heterogeneidad en otros tipos de información

Esta sección se centra en describir los tipos de heterogeneidad existentes en bases de datos y en modelos de conocimiento/ontologías relacionadas con la información geográfica.

A) Bases de datos

Desde una perspectiva genérica, según Hull (1997), los Sistemas Gestores de Bases de Datos (SGBD) gestionan y proporcionan acceso a un gran volumen de datos orientados a registros mediante una simple plataforma lógica. Sin embargo, la interoperabilidad entre dos o más sistemas de bases de datos, de manera eficiente, resulta un reto importante consecuencia de la heterogeneidad existente. El tratamiento de la heterogeneidad en este área (bases de datos) presenta abundante literatura, albergando problemas que aparecen en múltiples formas (plataforma *hardware* y *software*, modelos de datos y esquemas utilizados para proporcionar una estructuración lógica a los datos almacenados, tipos de datos e información almacenada, etc.) e inclusive interrelacionando estos problemas con la heterogeneidad recogida en la información geoespacial. A continuación, se recoge un conjunto representativo de diferentes tipos de heterogeneidad comúnmente distinguidos en la bibliografía científica.

Con respecto a la **heterogeneidad sintáctica**, estrechamente relacionada con la heterogeneidad en la información geoespacial, (Bishr, 1998) afirma que una base de datos puede ser implementada conforme a diferentes modelos de datos lógicos y sus subyacentes SGBD, por ejemplo, relacional u orientado a objeto. En estrecha relación a la mencionada heterogeneidad sintáctica se encuentra la **heterogeneidad en el proceso de transacciones** (Morocho Zurita *et*

al., 2001). En este tipo de heterogeneidad los diferentes SGBD o SGBD espaciales, pueden utilizar diferentes mecanismos para el procesamiento de transacciones de información, consecuencia de los diferentes sistemas SIG e IDE. Esto añade, por tanto, características espaciales que deben guardar consistencia. Por este motivo, el problema en el procesamiento de transacciones es subdividido en dos niveles; i) mantenimiento de la consistencia entre datos alfanuméricos y ii) mantenimiento de la consistencia entre datos geográficos.

En la **heterogeneidad esquemática** (Bishr, 1998) o **heterogeneidad en el modelo conceptual** (Hakimpour, 2003) o **heterogeneidad del modelo de datos** (Morocho Zurita *et al*., 2001) o **heterogeneidad representacional** (Chou, 2005) puede considerarse que partes de un mismo dominio o universo de discurso pueden ser representadas en distintos modelos mediante diferentes primitivas. De esta manera, un fenómeno del mundo real puede ser modelado como un objeto en una base de datos y considerado como un atributo en otra. Por consiguiente, no todas las bases de datos en una arquitectura heterogénea están basadas en el mismo modelo de datos. Este hecho genera inconsistencias derivadas de la utilización de diferentes esquemas de bases de datos para el almacenamiento de los datos. Asimismo, las clases de objetos pueden tener diferentes jerarquías de generalización o agregación, aunque describan el mismo fenómeno del mundo real.

Además, en la **heterogeneidad estructural** (Chatterjee *et al*., 1995; Kim *et al*., 1991; Kashyap *et al*., 1996; Sheth *et al*., 1990), de esquema (Hakimpour, 2003) o disparidad estructural (*structural conflicts*) (Spaccapietra *et al*., 1994) puede ocurrir que diferentes sistemas tengan almacenado el mismo nombre para un mismo tipo de fenómeno, pero diferentes atributos o formatos, por lo que se puede deducir que los datos de cada sistema son diferentes y, en ocasiones, la interpretación de la información distinta. Un ejemplo de este tipo de heterogeneidad se muestra en la Tabla 2.2.

Tabla 2.2. Extracto de tablas correspondientes a diferentes bases de datos con esquema dispar
(Los idiomas presentes en la tabla están codificados utilizando ISO 639 (ISO-639-1, 2002))
(* El valor de longitud está expresado en grados decimales)

ID	Nombre	Longitud*	Variante	**Idioma**
188	Ebro, Río	2121,33		es
190	Ebro, Río	76,4	Ebre, l'	ca
74	Miño, Río	53574,76		gl
379	Miño, Río	53834,62		es

ID	**Nombre**	**Vertiente**	**Cuenca**	**Longitud**
1	EBRO	ATLÁNTICA	EBRO	910
5	MIÑO	ATLÁNTICA	MIÑO	310

Dentro de este tipo de heterogeneidad aparece una serie de conflictos estructurales (Batini *et al*., 1986) que incluyen diferentes problemas como resultado de las elecciones en el proceso de modelado o en las restricciones de integridad, tales como:

- *Conflictos de dependencia*. Aparecen cuando un grupo de conceptos están relacionados entre ellos pero con diferentes dependencias en diferentes esquemas. Por ejemplo, en un esquema de zonas de riesgo, la relación entre superficie y contorno es 1:1, una superficie tienen un contorno, descrito por un conjunto de líneas que se cierran. Sin embargo, en un esquema de divisiones administrativas la relación entre superficie y contorno es 1:n, ya que un municipio, por ejemplo, puede tener anejos o adiciones y enclavados o sustracciones, cosa que en el caso de los riesgos no tiene sentido.
- *Conflictos de claves*: Diferentes claves (primarias y/o extranjeras) son asignadas a los mismos conceptos en diferentes esquemas.
- *Conflictos conductuales*: Estos surgen cuando diferentes políticas de inserción/borrado son asociados a la misma clase de objeto en esquemas distintos. Por ejemplo, en un

esquema de una base de datos un departamento puede existir sin que existan empleados, mientras que en otra base de datos el proceso de borrado del último empleado conlleva la eliminación de la propia tabla asociada. Este tipo de conflictos sólo aparecen cuando el modelo de datos permite la representación de propiedades de comportamiento de los objetos.

En el contexto de la integración de fuentes heterogéneas de datos espaciales, (Morocho Zurita *et al*., 2001) plantean, partiendo del estudio realizado por (Sheth *et al*., 1990), diversos tipos de heterogeneidad desde el punto de vista de las bases de datos espaciales.

- *Heterogeneidad en el lenguaje de consulta*. Los diferentes SGBD pueden utilizar diferentes lenguajes de consulta. Por esto, aunque los SGBD estén basados en el modelo relacional[18], uno podría usar SQL[19] y otro un lenguaje basado en cálculo relacional.
- *Heterogeneidad en el procesamiento de consultas***.** Diferentes SGBD podrían utilizar diferentes procesamientos y estrategias de optimización de consultas, debido a la diversidad en la representación de elementos geográficos por dependencia de escalas, sistemas de coordenadas, sistema de proyección, geoides y otros componentes similares.
- *Heterogeneidad en restricciones de integridad*. Los diferentes SGBD hacen cumplir diferentes restricciones de integridad. Las restricciones de integridad velan por la integridad entre datos que representan objetos en el espacio, datos geométricos y datos alfanuméricos.

Por otro lado, se encuentran las heterogeneidades relacionadas con la semántica. Desde el dominio de la lingüística, la semántica se define como las relaciones entre las palabras y las cosas a las que esas palabras se refieren (ter Bekke, 1992). Asimismo, en una aproximación al dominio computacional, la semántica se define como la interpretación matemática de las expresiones de un lenguaje formal (Leeuwen, 1990). En un plano más práctico, la **heterogeneidad semántica** aparece, de forma habitual, como la fuente de la mayoría de los problemas presentes en la información. De forma generalizada, este tipo de heterogeneidad es causada por la representación, significado, interpretación o uso desigual de los datos (es decir, al relacionar los datos con lo que representan) causada por un entendimiento particular del mundo asociado a un determinado contexto (Hakimpour, 2003; Wache *et al*., 2001; Sheth *et al*., 1990; Naiman *et al*., 1995). Según Sheth *et al*. (1990) la heterogeneidad semántica en sistemas de bases de datos es provocada por la dificultad en la identificación, resolución y utilización de esquemas de SGDB que no proporcionan suficiente semántica para interpretar los datos de forma consistente. Este tipo de heterogeneidad es tratada desde diferentes criterios dependiendo de las consideraciones de los autores. A continuación se lista una recopilación de diferentes acercamientos de este tipo de heterogeneidad.

Según Bishr (1998) este tipo de heterogeneidad ocurre como consecuencia de la variación de los modelos entre las diferentes disciplinas y necesidades, incluso aunque los fenómenos geográficos, probablemente, compartan una visión común. Esta heterogeneidad es subdividida en heterogeneidad cognitiva y heterogeneidad designativa.

[18] El *modelo relacional* es un modelo que estructura los datos de manera lógica en lo que denomina relaciones, representadas en forma de tablas (Codd, 1970). A este modelo se llega habitualmente en el área de las bases de datos a través del modelo Entidad-Relación, que trata de modelar el mundo como un conjunto de entidades asociadas entre sí mediante relaciones (Chen, 1975). Una descripción detallada de los componentes del modelo relacional aparece en (Codd, 1970).

[19] *Structured Query Language* o Lenguaje de consulta estructurado es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones sobre las mismas. Una de sus características es el manejo del álgebra y el cálculo relacional permitiendo lanzar consultas con el fin de recuperar -de una forma sencilla- información de interés de una base de datos, así como hacer cambios sobre la misma.

- La *heterogeneidad cognitiva* es frecuente cuando no existe una base común de definiciones para los fenómenos comunes de diferentes catálogos o bases de datos.
- La *heterogeneidad designativa* es debida a fenómenos similares semánticamente, pero que pueden ser nombrados de forma diferente. Las personas se refieren a los mismos datos utilizando su propia terminología y nombres, como consecuencia esto provoca una proliferación de nombres, así como posibles inconsistencias entre los nombres en los componentes de los esquemas (Batini *et al*., 1986). Por ejemplo, "curso de agua" y "río" son dos nombres utilizados por los productores de información para describir el mismo fenómeno geográfico. Este tipo de heterogeneidad acaece también cuando los esquemas de nombrado de la información difieren significativamente (Goh, 1997). Asimismo, esta heterogeneidad o conflictos son identificados y concretados respecto a sinónimos y homónimos en (Naiman, *et al*., 1995; Chatterjee *et al*., 1995; Batini *et al*., 1986; Goh, 1997).

    - Homónimos: Cuando el mismo nombre es utilizado para dos conceptos diferentes, incrementando las inconsistencias a menos que sean detectadas.
    - Sinónimos: Cuando el mismo concepto es descrito por dos o más nombres. A menos que diferentes nombres mejoren el entendimiento de diferentes usuarios, éstos no están justificados.

Ambos tipos de heterogeneidad (cognitiva y designativa) a menudo aparecen a la vez y mezcladas, lo que dificulta aún más los problemas de interpretación de los datos.

Desde la comunidad de bases de datos y en el contexto de la federación de bases de datos (Kim *et al*., 1991) también proponen en su clasificación disparidades a nivel semántico. Estas se refieren al contenido y significado de cada elemento de información. En (Hakimpour, 2003), un sistema puede adoptar un punto de vista sobre un tipo de fenómeno o instancia, mientras otro sistema diferente puede adoptar un punto de vista totalmente diferente. Además, es habitual encontrar diferentes definiciones de un mismo fenómeno.

La propuesta de (Chatterjee *et al*., 1995) señala problemas derivados de la pérdida o incompletitud de la información, errores de codificación o grabación de información (errores tipográficos o variaciones en medida), evolución del dominio (esto ocurre cuando la semántica de los valores en un dominio particular cambian con el paso del tiempo a formas que no son susceptibles de aplicar simples correspondencias entre los "viejos" y los "nuevos" valores) y actualizaciones asíncronas (estas ocurren cuando los elementos de los datos, replicados en bases de datos diferentes, se actualizan en diferentes momentos en el tiempo, llegando a ser inconsistentes).

En el contexto de la federación de bases de datos heterogéneas, (Hammer *et al*., 1993) hablan de un espectro de heterogeneidad basado en cuatro niveles de abstracción:

- *Nivel de lenguaje de metadatos* debido a la utilización de diferentes lenguajes de especificación del modelo conceptual de la base de datos.
- *Nivel de especificación de metadatos* que incluye las diferencias en el modo en que se modela un dominio.
- El nivel denominado "*comparación de objetos*" en el que se agrupan las diferencias terminológicas
- *Nivel de formato de datos* para referirse a las diferentes formas en que los valores atómicos pueden almacenarse en dos modelos diferentes.

Mientras, (Naiman *et al*., 1995) mencionan dos conflictos semánticos; abstracción y niveles de heterogeneidad.

- *Abstracción*, vinculada con la heterogeneidad esquemática, establece una relación de conceptualización (abstracción) entre dos elementos esquemáticos. Éste conflicto es subdividido en generalización y agregación.

    - Generalización, ocurre como consecuencia de que los conceptos y sus instancias asociadas existen a diferentes niveles de abstracción en dos bases de datos distintas.
    - Agregación, ocurre cuando un objeto en una base de datos le corresponde un grupo de objetos en otra base de datos.

- *Niveles de heterogeneidad*; los conceptos correspondientes pueden ser representados a diferentes niveles en cada base de datos y esta relación estructural es expresada por los siguientes pares "objeto, objeto", "objeto, atributo", "objeto, instancia" y otros.

Dentro de los sistemas distribuidos, Worboys *et al*. (1991) afirman, conforme a Sheth *et al*. (1990), que la heterogeneidad semántica existe cuando hay un desacuerdo en el significado, interpretación o uso previsto de los datos. Estos autores (Worboys *et al*., 1991) tratan problemas específicos de la heterogeneidad semántica en un sistema espacial distribuido, identificando dos tipos de heterogeneidad:

- La *heterogeneidad semántica genérica* ocurre cuando diferentes nodos utilizan diferentes modelos genéricos de información espacial.
- La *heterogeneidad semántica contextual* acontece cuando la semántica de los esquemas depende de las condiciones locales en los nodos particulares.

Dos áreas singulares son identificadas por (Goh, 1997) como parte de la heterogeneidad semántica:

- *Conflictos de confusión*. Estos suceden cuando los campos de información parecen tener el mismo significado, pero en realidad son diferentes, por ejemplo, conflictos propios de diferentes contextos temporales.
- *Conflictos de escala*. Ocurren cuando se utilizan diferentes sistemas de referencia para medir un valor. Un ejemplo típico en este tipo de conflictos son los diferentes tipos de moneda.

(Sheth *et al*., 1992) proporcionan una lista exhaustiva de conflictos semánticos organizados en torno al dominio, entidad, valor de los datos, nivel de abstracción y conflictos esquemáticos.

- *Incompatibilidad de dominio*. Esta problemática surge entre dos objetos cuando estos tienen definiciones diferentes aunque con dominios de atributos similares semánticamente. Aquí aparecen problemas como conflictos de nombrado (designativos), conflictos de representación de datos, conflictos de escalado de datos, conflictos de precisión de los datos, conflictos de valores por defecto y los conflictos de restricciones de integridad de atributos.
- *Definición de entidades*. En este caso, las incompatibilidades surgen entre dos objetos cuando los descriptores de la entidad utilizada por los objetos son sólo parcialmente compatibles, incluso cuando el mismo tipo de entidad es modelado. Dentro del mencionado conflicto se encuentran conflictos de identificadores de bases de datos, conflictos de nombrado, conflictos de representación de datos, conflictos de escalado de datos, conflictos de precisión de los datos, conflictos de valores por defecto y los conflictos de restricciones de integridad de atributos.
- *Incompatibilidad de valores de datos*. Esta clase de conflicto cubre aquellas incompatibilidades que aparecen debido a la presencia de valores de los datos en

diferentes bases de datos. En este conflictos aparecen problemas como inconsistencias conocidas, inconsistencias provisionales e inconsistencias admisibles.

- *Niveles de abstracción*. Este conflicto surge cuando dos entidades semánticamente similares son representadas a diferentes niveles de abstracción. Las diferencias en la abstracción pueden surgir debido a los diferentes niveles de generalización en los que una entidad es representada en una base de datos. Éstas también pueden aparecer debido a la agregación utilizadas a nivel de entidad y de atributo. Dentro de este conflicto aparecen los siguientes tipos de problemas: Conflictos de generalización y conflictos de agregación.
- *Discrepancias esquemáticas*. Este tipo de conflictos pueden tener lugar dentro de un mismo modelo y surgen cuando los datos en una base de datos corresponden a metadatos en otra base de datos. En este conflicto se encuentran los conflictos de valores de atributos, conflictos de entidad de atributos y los conflictos de valor de entidad.

En (Yeung *et al*., 2007) la heterogeneidad semántica se define como la inconsistencia entre la representación digital y su correspondiente fenómeno del mundo real dentro de un determinado contexto. Según estos autores esta heterogeneidad se subdivide en heterogeneidad de conceptualización, heterogeneidad de formalización y heterogeneidad de contexto.

- *Heterogeneidad de conceptualización*, también conocida como heterogeneidad de cognición, ocurre cuando diferentes personas abstraen el mismo fenómeno del mundo real utilizando su propia experiencia o desde la perspectiva de sus respectivas disciplinas.
- *Heterogeneidad de formalización*, se produce por el uso de diferentes lenguajes (heterogeneidad de lenguaje) y terminología (heterogeneidad de representación) para describir una conceptualización.
- *Heterogeneidad de contexto*, está relacionada con el cambio de significado de una conceptualización o formalización bajo diferentes condiciones o circunstancias.

B) Modelos de conocimiento/ontologías

En el contexto de la interoperabilidad semántica, descrita en la sección 2.1.1.3, la propuesta de entendimiento *colectivo* de este tipo de interoperabilidad se centra en la utilización de una conceptualización común y compartida, con objeto de evitar inexactitudes o confusiones en la interpretación del significado de los términos, en los procesos de intercambio de información. Estas son las características fundamentales que proponen las ontologías. Una descripción general de este concepto y su vinculación con el dominio geoespacial se recoge en las secciones 2.1.2 y 3.1, respectivamente.

En el contexto de la interoperabilidad semántica de la IG (Navarrete, 2006) identifica tres niveles de heterogeneidad semántica, basados en la clasificación proporcionada por (KnowledgeWeb Consortium, 2005). Estos niveles están estrechamente relacionados con ciertos aspectos ontológicos. A pesar de esto, dichos niveles son recogidos en este apartado como consecuencia de que muchos de los problemas identificados son extrapolables a la heterogeneidad presente en la información geoespacial:

- El *nivel sintáctico* se refiere al hecho de que diferentes ontologías puedan estar expresadas en diferentes lenguajes, como *Web Ontology Language* (OWL) o *Knowledge Interchange Format* (KIF). Desde la perspectiva del dominio geográfico este nivel está relacionado con la heterogeneidad de formatos generada por la heterogeneidad del software existente (Morocho Zurita *et al*., 2001), consecuencia de la utilización de múltiples herramientas (SIG y/o CAD (*Computer Aided Design*)) en la producción de información.

- El *nivel terminológico* comprende los desajustes en el nombre de los conceptos. Ejemplos de desajustes en este nivel son sinonimia, polisemia, diferentes idiomas (español, inglés, etc.) o variaciones sintácticas de la misma palabra (abreviaciones, diferentes formas aceptadas de escritura, prefijos o sufijos, etc.).
- El *nivel conceptual* incluye los desajustes relacionados con el contenido de la ontología. Hay dos tipos de discrepancias conceptuales principales:

    - Por un lado, están las ***diferencias meta-físicas***, referidas a cómo el mundo es "roto en pedazos" (es decir, qué entidades, propiedades y relaciones son representadas en una ontología). Este tipo de diferencias comprende tres subtipos de dimensiones, sobre las que se proporciona una representación gráfica en la Figura 2.5.
        - Cobertura: Diferentes ontologías cubren diferentes porciones del mundo real.
        - Granularidad: Una ontología proporciona una descripción más detallada de los mismos conceptos que otra ontología distinta.
        - Perspectiva: Dos ontologías son el resultado de observar el mundo real desde diferentes puntos de vista, resultado típico en diferentes disciplinas.
    - Y, por otro, las ***diferencias epistémicas***, relacionadas con cómo los humanos entienden estas entidades, es decir, qué afirmaciones son hechas sobre ellas.

A los tres niveles anteriores se le añade un nivel más procedente de (Euzenat *et al*., 2007):

- El *nivel semiótico*, también llamado heterogeneidad semiótica o pragmática, afecta a cómo las entidades son interpretadas por las personas. En efecto, las entidades que tienen exactamente la misma interpretación semántica son, con frecuencia, interpretadas conforme al contexto por los humanos.

Asimismo, dentro de esta subdivisión de la heterogeneidad, en (Chou, 2005) se menciona la **heterogeneidad representacional**. Este tipo de heterogeneidad se centra en la ausencia de un vocabulario controlado y métodos consecuentes para la representación de los datos.

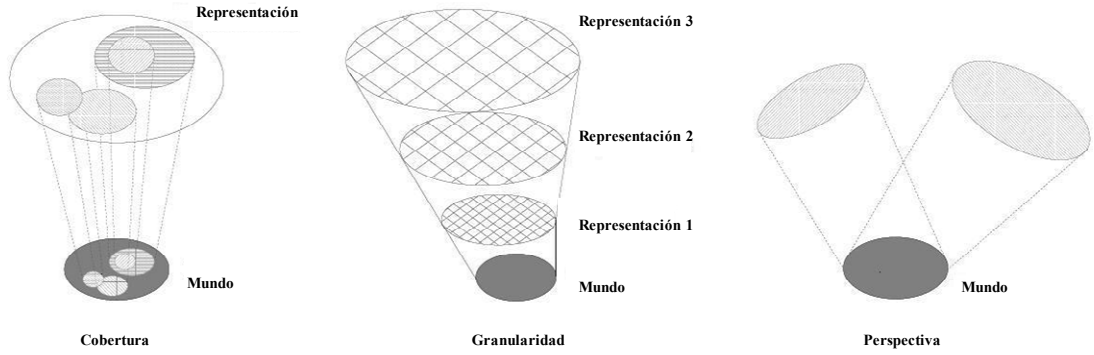


Figura 2.5. Dimensiones de la heterogeneidad a nivel conceptual (KnowledgeWeb Consortium, 2005)

2.3. Una visión general de la problemática de la información geoespacial

La información geoespacial presenta gran diversidad de problemas. En esta sección se abordan cuatro problemas principales que atañen a este tipo de información. Estos problemas se clasifican conforme a los puntos de vista expuestos en (Gómez-Pérez *et al*., 2008). De esta manera, vinculado a la visión global de la información geoespacial aparecen los siguientes problemas:

- Problemas de estructuración de la información geoespacial (sección 2.3.1).
- Problemas de las bases de datos geoespaciales (sección 2.3.2).

Desde un punto de vista más específico y, por tanto, más vinculado a los fenómenos, se recogen los siguientes problemas:

- Problemas relacionados con el factor escala (sección 2.3.3).
- Problemas relacionados con el lenguaje natural (sección 2.3.4).

En la Tabla 2.3Tabla 2.3. Problemática de la información geoespacial relacionada con diferentes tipos de heterogeneidad
se muestra, a modo de resumen, una visión general de los diferentes problemas de la IG vinculados con diversos tipos de heterogeneidad, descritos con anterioridad. Asimismo, en esta tabla se remarca la tipología de heterogeneidad común a un mismo problema con objeto de reflejar una visión de conjunto de una problemática determinada.

2.3.1. Estructuración de la información geoespacial

La información geoespacial es percibida subjetivamente y su contenido depende de las necesidades de aplicaciones concretas. Estas necesidades determinan la forma en que las instancias se agrupan en tipos dentro de un esquema de clasificación particular, consecuencia de que los mismos tipos de fenómenos geográficos son, con frecuencia, abstraídos de forma diferente de una fuente a otra. Así diferentes productores cartográficos crean datos espaciales con diferentes puntos de vista, intereses y necesidades, utilizando diferentes vocabularios y modelos. Este hecho está relacionado con la heterogeneidad esquemática (Bishr, 1998) o heterogeneidad en el modelado conceptual (Hakimpour, 2003), heterogeneidad del modelo de datos (Morocho Zurita *et al*., 2001), la heterogeneidad representacional (Chou, 2005) y terminológica (KnowledgeWeb Consortium, 2005; Navarrete, 2006; Naiman, *et al*., 1995; Chatterjee *et al*., 1995; Batini *et al*., 1986; Goh, 1997; Bishr, 1998; Euzenat *et al*., 2007), descritas en la sección anterior.

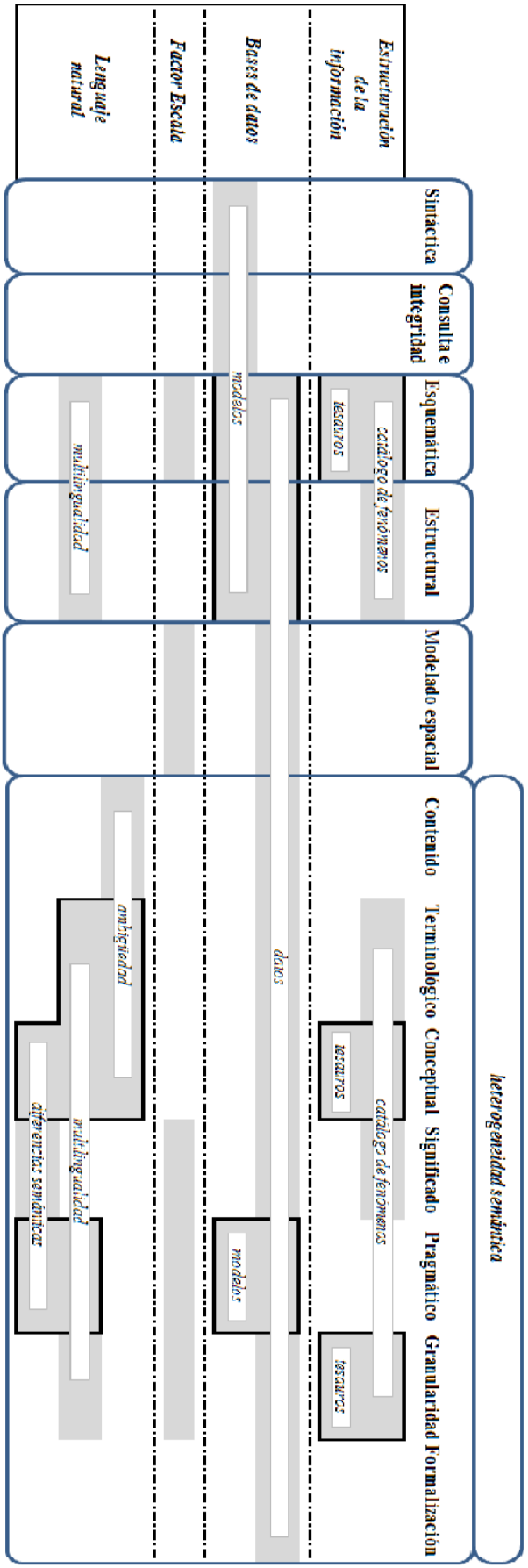
En la actualidad, la forma comúnmente empleada para estructurar nombres, códigos, atributos y otras características, inclusive las referidas a la representación gráfica, asociadas a los fenómenos geográficos es el catálogo de fenómenos (*feature catalogue*). Según ISO 19110 (2005), los fenómenos geográficos son elementos del mundo real asociados con una localización relativa a la Tierra. Sobre dicha localización los datos son capturados, mantenidos y diseminados. Los catálogos de fenómenos definen los tipos de fenómenos, sus operaciones, atributos y asociaciones representadas en los datos geográficos. Estos catálogos promueven la diseminación, compartición y uso de los datos geográficos ya que proporcionan un mejor entendimiento del contenido y significado de los datos. De esta manera, los proveedores y usuarios de datos geoespaciales pueden tener un entendimiento compartido de los tipos de elementos del mundo real representados por los datos y los usuarios pueden juzgar si los datos suministrados se ajustan a sus necesidades y permiten alcanzar sus propósitos.

En la práctica estos catálogos, como se puede comprobar en los extractos presentados a modo de ejemplo en las Figura 2.6 y Figura 2.6, consisten en una lista de fenómenos geográficos agrupados en diferentes clases y donde dichos fenómenos tienen instancias con características comunes (por ejemplo; río, iglesia, carretera, etc.) Cada fenómeno es identificado por un código único, que representa un elemento discreto asociado con sus coordenadas geográficas y temporales.

Además, puede ser representado por un símbolo gráfico concreto (por ejemplo, la catedral de Burgos, el volcán Teide o la ciudad de Madrid). Por otro lado, en ocasiones, este tipo de catálogos pueden contemplar, acompañando a los diferentes tipos de fenómenos, definiciones y diferentes tipos de atributos. Conforme a las características presentadas, desde una perspectiva

general, los catálogos de fenómenos se caracterizan por presentar heterogeneidad esquemática (Bishr, 1998; Hakimpour, 2003), estructural (Chatterjee *et al*., 1995; Hakimpour, 2003;

Tabla 2.3. Problemática de la información geoespacial relacionada con diferentes tipos de heterogeneidad



0602	AUTOPISTA		AUTOPISTA
Nombre:	BCN200_0602L_AUTOPISTA		
Definición:			
Tipo de entidad:			
Criterio de selección:			
Atributos:			
	060201	COMPONENTE	COMPON_0602
		01 EJE CALZADA	
		02 MEDIANA	
		03 RAMAL	
	060202	SITUACIÓN	SITUAC_0602
		01 SUPERFICIAL	
		02 SUBTERRÁNEO	
		03 ELEVADO	
	060203	ESTADO	ESTADO_0602
		01 EN USO	
		02 EN CONSTRUCCION	
	060204	COMPETENCIA	COMPET_0602
		01 AGE	
		02 AUTONÓMICA	
		03 RESTO	
	060205	ACCESO	ACCESO
		01 LIBRE	
		02 DE PEAJE	
	060206	INTERNA	INTERNA
		01 NO INTERNA	
		02 INTERNA	
	060207	ITINERARIO_EUROPEO	ITIN_EUR
		01 NO ITINERARIO EUROPEO	
		02 ITINERARIO EUROPEO	

Figura 2.6. BTN200

0301	RÍO	
ATRIBUTO: 030101	COMPONENTE DE RÍO	COMPON_01
	01 EJE	
	02 MARGEN	
	03 EJE CONEXIÓN	
ATRIBUTO: 030102	RÉGIMEN	REGIME_01
	01 PERMANENTE	
	02 NO PERMANENTE	
ATRIBUTO: 030103	CATEGORÍA DEL RÍO	CATEGO_02
	00 DESCONOCIDA	
	01 PRIMERA	
	02 SEGUNDA	
	03 TERCERA	
	04 CUARTA	

0304	CAUCE ARTIFICIAL	
ATRIBUTO: 030401	COMPONENTE DEL CAUCE ARTIFICI	COMPON_02
	01 EJE	
	02 MARGEN	
	03 EJE CONEXIÓN	

Figura 2.7. BTN25

Fuente: Bases Topográficas Numéricas (BTN), pertenecientes al Instituto Geográfico Nacional, a escala 1:200.000 (BTN200) y 1:25.000 (BTN25) respectivamente

Kashyap *et al*., 1996), terminológica (Navarrete, 2006; Chatterjee *et al*., 1995; Batini *et al*., 1986; Euzenat *et al*., 2007; Goh, 1997), conceptual (Chou, 2005; Yeung *et al*., 2007), de significado (Goh, 1997; Worboys *et al*., 1991) y granularidad (Naiman, *et al*., 1995; Goh, 1997; Euzenat *et al*., 2007).

Otra forma común de organizar la información geoespacial es el uso de tesauros. Sin embargo, éstos no son generalmente utilizados por los productores de IG. Según ISO 5964 (1985); ISO 2788 (1986) un tesauro es vocabulario de un lenguaje de indexación controlado (conjunto controlado de términos extraídos del lenguaje natural y utilizados para representar, de forma breve, los temas de los documentos) organizado formalmente con objeto de hacer explícitas las relaciones, *a priori*, entre conceptos (por ejemplo "más genérico que" o "más específico que").

La construcción de tesauros supone una considerable mejora en la estructuración, consecuencia de la desaparición de la imprecisión y ambigüedad en el uso del lenguaje (motivada por la existencia de sinónimos y polisemias) y del aporte de semántica adicional en sus relaciones (ej.: "Término Genérico", "Término Específico", "Use", etc.) entre los términos (Vilches *et al.,* 2006), haciendo que sean interpretadas de modo inequívoco. Sin embargo, los tesauros no proporcionan una jerarquía explícita, aunque con las especificaciones de los términos más genéricos y más específicos se podría deducir una jerarquía (Lassila *et al*., 2001).

Las características que presentan los tesauros albergan problemas que afectan a la heterogeneidad esquemática (Naiman, *et al*., 1995; Worboys *et al*., 1991; Bishr, 1998), conceptual (Goh, 1997; Chou, 2005; Worboys *et al*., 1991) y de grano (nivel de detalle) (Naiman, *et al*., 1995; Goh, 1997; KnowledgeWeb Consortium, 2005; Euzenat *et al*., 2007).
Dentro de estos recursos, uno de los tesauros más conocido en el contexto de la información geoespacial es el *Getty Thesaurus of Geographic Names* (TGN)[20] (Harping, 1997; Harpring *et* al., 2006). En TGN, como puede apreciarse en la Figura 2.8, cada registro muestra las

[20] *TGN*. http://www.getty.edu/research/conducting_research/vocabularies/tgn/

coordenadas espaciales asociadas, la posición jerárquica, una breve explicación de la región y los nombres de lugares asociados. Además, los nombres para un lugar pueden incluir: nombres en la lengua vernácula, inglés, otros idiomas, nombres históricos. Para más detalles, sobre éste y otros recursos, puede consultarse la recopilación de tesauros geográficos e históricos realizada en (Vilches-Blázquez *et al*., 2007c).

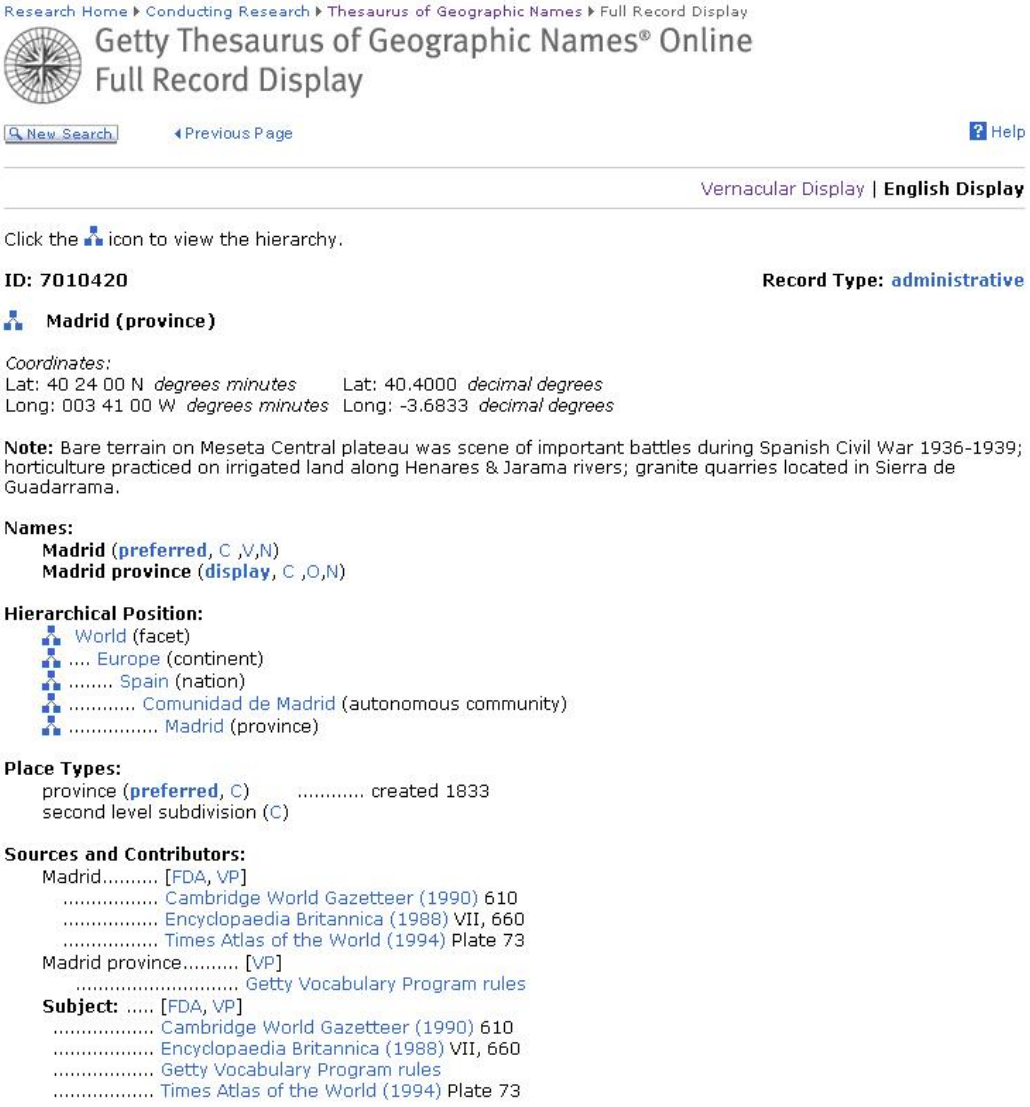


Figura 2.8. Interfaz de usuario del Getty Thesaurus of Geographic Names

En definitiva, las formas más extendidas de estructuración de la información espacial, atendiendo a la clasificación presentada por (Lassila *et al*., 2001), se caracterizan por su simplicidad respecto a otras formas de organización presentes en la Figura 2.9. Este hecho es provocado por (1) la ausencia de una óptima estructuración, ya que, con frecuencia, estas formas se limitan a simples representaciones jerárquicas, más o menos explícitas, determinadas, en ocasiones, por identificadores (ID) o especificaciones del tipo más genérico y/o más específico asociadas a cada fenómeno; (2) la simplicidad y, en ocasiones, ausencia de relaciones entre los términos y (3) la ausencia de una semántica formal claramente definida. Estos factores provocan un impacto negativo sobre la eficiencia de los procesos de búsqueda, acceso y recuperación de información, así como importantes dificultades en procesos de integración semántica de fuentes heterogéneas de IG.

Figura 2.9. Clasificación de tipos de ontologías (Lassila *et al*., 2001)

2.3.2. Bases de datos geoespaciales

El corazón de cualquier SIG o IDE está constituido por un sistema de base de datos, aunque esto no significa que las bases de datos espaciales sean únicamente la tecnología subyacente a estas herramientas. Una descripción detallada de las características de los sistemas de bases de datos espaciales puede ser encontrada en (Güting 1994; Abraham *et al*., 1999; Shekhar *et al.*, 2003; Yeung *et al.*, 2007).

Desde un punto de vista funcional, el rápido crecimiento de la IG ha demandado el desarrollo de SGDB que soporten estos datos espaciales y sean capaces de proporcionar operaciones y funciones de análisis poderosas capaces de ayudar en los procesos de toma de decisiones, proyecciones, predicciones y simulaciones de un amplio abanico de problemas (Lohfink *et al*, 2007). Asimismo, las bases de datos se extienden como una forma común de implementación de las diferentes formas de estructuración de la información geoespacial. A continuación se recoge un listado, no exhaustivo, con las principales características de los SGDB espaciales más extendidos.

- ***Oracle Spatial***[21]. Esta extensión para las bases de datos Oracle incluye capacidades espaciales avanzadas para apoyar el uso de las aplicaciones SIG, servicios basados en localización, etc. Entre sus componentes principales destacan: un esquema (MDSYS) que prescribe el almacenaje, la sintaxis y la semántica de los tipos de datos geométricos apoyados; un mecanismo espacial de indexación; un sistema operador y las funciones para realizar consultas de la sección de interés, consultas de unión espacial, etc.
- ***MySQL Spatial***[22]. Introduce las extensiones espaciales para permitir la generación, almacenamiento y análisis de elementos geográficos. MySQL implementa extensiones espaciales siguiendo la especificación *Simple Features Specifications For SQL*[23] de OGC como un subconjunto del entorno SQL.
- ***PostGIS***[24]. Es un módulo que añade soporte de objetos geográficos a la base de datos objeto-relacional *PostgreSQL*[25], convirtiéndola en una base de datos espacial. Asimismo, implementa metadatos, funciones geométricas y topológicas para el tratamiento de los datos espaciales conforme con los estándares OGC (Baksai *et al.*, 2007). *PostGIS* almacena la información geográfica en una columna del tipo GEOMETRY, donde se almacena la geometría en formato WKB (*Well-Known Binary*), aunque hasta la versión 1.0 se utilizaba el formato WKT (*Well-Known Text*).

[21] *Oracle Spaital.* http://www.oracle.com/technology/products/spatial/index.html
[22] *MySQL Spatial* http://dev.mysql.com/doc/refman/5.1/en/spatial-extensions.html
[23] *Simple Features Specifications For SQL* http://www.opengeospatial.org/standards/sfs
[24] *PostGIS* http://postgis.refractions.net/
[25] *PostgreSQL* http://www.postgresql.org/

- ***SpatiaLite***[26]. Es la extensión espacial para la base de datos relacional *SQLite*[27]. Esta extensión es conforme con algunas de las especificaciones de OGC. Asimismo, soportar los formatos estándares WKB y WKT; implementa funciones espaciales de SQL (por ejemplo: *GeomFromText*(), *Area*(), *PointN*(), etc.); soporta metadatos espaciales; importa y exporta a formato *shapefile*; permite la reproyección de coordenadas; etc.

2.3.2.1. Datos

Desde una óptica general, los datos existen independientemente de los programas de ordenador y cada área de conocimiento utiliza unos tipos de datos para representar su información. Así, los datos pueden variar entre diferentes dominios de aplicación, incluso cuando se refieren a los mismos objetos del mundo real, debido a que las personas tienen, habitualmente, diferentes percepciones e interpretaciones de la realidad, acorde a las actividades que llevan a cabo (Fileto, 2001). Un ejemplo de estas diferencias en el contexto de la información geoespacial se muestra en la Tabla 2.4 (diversidad en la abstracción y representación de fenómenos geográficos provenientes de diferentes fuentes).

A nivel geográfico, consecuencia de los múltiples productores de IG, es bastante común encontrar diferentes bases de datos que describan, al menos parcialmente, el mismo espacio geográfico. Aunque, normalmente los datos son recogidos para propósitos específicos y, por tanto, son muy diferentes de una fuente a otra (Devogele *et al*., 1998). Además, cada base de datos mantiene sus propios supuestos sobre los datos almacenados de manera autónoma e independiente (Souza *et al*., 2006) y estos pueden estar contenidos en múltiples capas de información sobre un área restringida de la superficie de la tierra (Worboys *et al*., 1991).

Junto a los múltiples productores existen otros factores que van a originar uno de los mayores problemas de la IG. Estos factores atañen a la heterogeneidad de los sistemas existentes en términos de modelado de datos, técnicas de codificación de datos, estructuras de almacenamiento, funcionalidades de acceso, pobre documentación y la ausencia de semántica de los datos provocando diversidad de los conjuntos de datos (qué información es almacenada, cómo es representada y estructurada, qué calidad tiene, a qué datos se refiere, qué escala es utilizada, etc.) (Devogele *et al*., 1998). Las diferentes problemáticas causadas por la duplicidad de datos tanto a nivel simple (una base de datos) como a nivel múltiple (en varias bases de datos) se detallan en el Capítulo 8 (sección 8.2) de este trabajo.

Por otro lado, la proliferación de los sistemas distribuidos hace que los datos puedan estar repartidos en múltiples bases de datos, con lo que dichas bases de datos pueden ser almacenadas en sistemas simples (una máquina) o en sistemas múltiples (un conjunto de máquinas), co-localizados o geográficamente distribuidas, pero interconectadas por un sistema de comunicación (Sheth *et al.,* 1990).

A la distribución geográfica de los datos se une la posibilidad de que los datos geoespaciales estén disponibles en varios formatos y sean almacenados en formatos dispares. De esta manera, existen algunos datos que están estructurados conforme a técnicas de modelado de datos bien establecidas (modelos de datos relacionales u orientados a objetos), mientras, otros datos, tales como los mantenidos en varios sistemas de información, hojas de cálculo, ficheros planos o repositorios de Internet, están en formatos propietarios, semi-estructurados o desestructurados (Díaz *et al*., 2008).

[26] *SpatiaLite.* http://www.gaia-gis.it/spatialite-2.2/index.html
[27] *SQLite.* http://www.sqlite.org/

Tabla 2.4. Fenómenos geográficos procedentes de diversas fuentes topográficas (Brodeur *et al.*, 2003)

NTDB[1]	VMap[2]	BC Digital Baseline Mapping (BCDBM)[3]	ON Digital Topographic Database (ONDTD)[4]	BDTQ[5]	Information sur les terres et les eaux pour la province du Nouveau-Brunswick[6]
- Waterbody - Watercourse - Irrigation Canal - Navigable Canal - Flooded area - Reservoir - Liquid depot/dump	- Lake/Pond - Lake subject to inundation - River/stream - Coastline/shoreline	- Coastline - Ditch - Flooded land - Lake - River/stream	- Flooded land - Lake - River/stream	- Canal - Cours d'eau - Lac - Mare	- Canal - Rivière–trait double - Lac (?) - Littoral (?) - Lac de rivière (?)
- Wetland	- Marsh/swamp	- Marsh - Swamp	- Marsh/Fen	- Milieu humide (végétation)	- Marais de canneberge (?) - Marais (?)
- Road - Limited access road	- Road - Car track	- Road	- Accesway - Road	- Voie de communication - Autoroute - Rue - Chemin - Route	- Artère (?) - Route collectrice (?) - Chemin local (?) - Chemin municipal (?) - Chemin d'accès aux ressources naturelles (?) - Route en construction (?) - Rue (?)
- Vegetation	- Trees - Orchard/plantation - Vineyard	- Wooded area - Vineyard - Orchard - Nursery	- Wooded area	- Milieu boisé - Verger (aires désignées)	- Clairière (?) - Bande défrichée (>100m) (?) - Pépinière (?) - Verger (?) - Rangée d'arbres (>100m) (?) - Zone boisée (>2m haut) (?)
- Railroad	- Railroad - Railroad siding/railroad spur	- RailLine	- Rail line	- Voie ferrée	- Chemin de fer (?) - Triage de chemin de fer (?)
- Bridge - Obstacle to air Navigation	- Bridge/overpass/viaduc (?)	- Bridge - Trestle	- Bridge (roadway) - Bridge (railway) - Culvert (roadway) - Culvert (railway)	- Pont - Pont d'étagement	- Pont (?) - Ponceau (petit) (?)

Spatial pictogram descriptions: :0D ; :1D ; :2D ; ?:unknown geometry ; :multiple geometry ; :alternate geometry (see (Bédard, 1999b) and (Brodeur et al., 2000) for more details). 1(Natural Resources Canada, 1996); 2(VMap, 1995); 3(BC Ministry of Environment Lands and Parks (Geographic Data BC), 1992); 4(OBM, 1996); 5(Québec, 2000); 6(New Brunswick, 2000).

Desde la óptica de la heterogeneidad, los problemas que afectan a los datos se circunscriben a los problemas de heterogeneidad esquemática (Bishr, 1998) o heterogeneidad en el modelo conceptual (Hakimpour, 2003), sintáctica (Bishr, 1998), heterogeneidad estructural (Chatterjee *et al*., 1995; Kim *et al*., 1991; Kashyap *et al*., 1996) o de esquema (Hakimpour, 2003) o disparidad estructural (Spaccapietra *et al*., 1994), consecuencia de que diferentes repositorios de datos pueden contener datos relevantes ante una consulta y éstos pueden tener diferentes organizaciones de datos (bases de datos, sistema de ficheros, etc.) formatos y soportar diferentes lenguajes de consulta (Mena *et al*., 2000). Asimismo, se pueden encontrar problemas vinculados a la heterogeneidad de modelado espacial (Hakimpour, 2003; Morocho Zurita *et al*., 2001), conceptual (Goh, 1997; Bishr, 1998; Morocho Zurita *et al*., 2001; Chou, 2005), pragmática (Morocho Zurita *et al*., 2001; Worboys *et al*., 1991), de contenido (Chatterjee *et al*., 1995; Sheth *et al*., 1992), granularidad (Goh, 1997), formalización (Morocho Zurita *et al*., 2001), terminológica (Chatterjee *et al*., 1995; Goh, 1997; Bishr, 1998; Chou, 2005) y de significado (Goh, 1997; Worboys *et al*., 1991).

2.3.2.2. Diversidad de modelos

La misma información, con frecuencia, se almacena utilizando diferentes modelos o esquemas de bases de datos. Además, normalmente, sólo parte del contenido de la información se modela en una fuente de información. Esto se refleja sobre los esquemas de las bases de datos con distintos fundamentos de construcción, paradigmas y poder de expresividad. Un problema añadido es el hecho de que la realidad geográfica es percibida de manera subjetiva. Esto hace que los modelos y, por tanto, su contenido (los datos) dependan de enfoques particulares. Asimismo, la diversidad de modelos y esquemas, combinada con las dificultades intrínsecas para la comunicación y el establecimiento de acuerdos para la representación de los datos en los diferentes dominios de aplicación, generan inconsistencias que hacen difícil el proceso de comunicación e interpretación entre diferentes máquinas y bases de datos, consecuencia de los problemas derivados de la interoperabilidad (Fileto, 2001). De este hecho se extrae la existencia de una total ausencia de homogeneidad en cómo los datos son representados en las bases de datos (Chou, 2005).

Nombre	Idioma Nombre	Variante	Idioma Variante	Grupo	Tipo Entidad	Municipio	Provincia	Comunidad Autónoma	Latitud	Longitud
Isábena, Río	es			HIDRO	Río		Huesca	Aragón	42º 11'	0º 02'
Sotón, Río	es			HIDRO	Río		Huesca	Aragón	42º 01'	-0º 46'
Ésera, Río	es			HIDRO	Río		Huesca	Aragón	42º 06'	0º 15'
Puebla de Alfindén, La	es			POBLA	Población 2		Zaragoza	Aragón	41º 37'	-0º 44'
San Mateo de Gállego	es			POBLA	Población 2		Zaragoza	Aragón	41º 49'	-0º 46'
Zaidín, Canal de	es			HIDRO	Canal		Huesca	Aragón	41º 54'	0º 17'
Épila	es			POBLA	Población 2		Zaragoza	Aragón	41º 36'	-1º 16'
Aragón y Cataluña, Canal de	es	Aragón i Catalunya, Canal d'	cat	HIDRO	Canal		Huesca	Aragón	42º 06'	0º 15'
Brea de Aragón	es			POBLA	Población 2		Zaragoza	Aragón	41º 31'	-1º 35'
Montearagón, Embalse de	es			HIDRO	Embalse		Huesca	Aragón	42º 11'	-0º 21'
Loteta, Embalse de la	es			HIDRO	Embalse		Zaragoza	Aragón	41º 52'	-1º 20'
Ejea de los Caballeros	es			POBLA	Población 2		Zaragoza	Aragón	42º 07'	-1º 08'
Fraga	es			POBLA	Población 2		Huesca	Aragón	41º 31'	0º 21'
Binéfar	es			POBLA	Población 2		Huesca	Aragón	41º 51'	0º 17'
Litera, La/Llitera, La	es/cat			ADMIN	Comarca administrativa		Huesca	Aragón	41º 51'	0º 17'
Almunia de Doña Godina, La	es			POBLA	Población 2		Zaragoza	Aragón	41º 28'	-1º 22'
Aguasvivas, Río	es			HIDRO	Río		Zaragoza	Aragón	41º 02'	-0º 25'
Aragón	es			ADMIN	Comunidad autónoma			Aragón	41º 39'	-0º 52'
Tranquera, Embalse de la	es			HIDRO	Embalse		Zaragoza	Aragón	41º 14'	-1º 48'
Ribagorza, La	es			ADMIN	Comarca administrativa		Huesca	Aragón	42º 06'	0º 28'
Zaragoza	es			POBLA	Capital		Zaragoza	Aragón	41º 39'	-0º 52'
Andorra	es			POBLA	Población 2		Teruel	Aragón	40º 58'	-0º 26'
Virgen, Sierra de la	es			OROGR	Sierra		Zaragoza	Aragón	41º 32'	-1º 47'
Monegros, Los	es			ADMIN	Comarca administrativa		Huesca; Zara	Aragón	41º 47'	-0º 09'
Campo de Daroca	es			ADMIN	Comarca administrativa		Zaragoza	Aragón	41º 06'	-1º 24'
Bajo Aragón	es			ADMIN	Comarca administrativa		Teruel	Aragón	41º 03'	-0º 07'
Bajo Cinca/Baix Cinca	es/cat			ADMIN	Comarca administrativa		Huesca; Zara	Aragón	41º 31'	0º 21'
Cella	es			POBLA	Población 2		Teruel	Aragón	40º 27'	-1º 17'
Campo de Belchite	es			ADMIN	Comarca administrativa		Zaragoza	Aragón	41º 18'	-0º 45'
Canelles, Embalse de	es	Canelles, Pantà de	cat	HIDRO	Embalse		Huesca	Aragón	42º 02'	0º 39'
Ribera Baja del Ebro	es			ADMIN	Comarca administrativa		Zaragoza	Aragón	41º 25'	-0º 29'
Jacetania, La	es			ADMIN	Comarca administrativa		Huesca; Zara	Aragón	42º 34'	-0º 32'
Malditos, Montes	es			OROGR	Sierra		Huesca	Aragón	42º 37'	0º 42'
Puertos de Beceite	es	Port, el	cat	OROGR	Sierra		Teruel	Aragón	40º 48'	0º 16'

Figura 2.10. Extracto de la tabla de la base de datos del Nomenclátor Geográfico Conciso de España

Algunos de los principales problemas originados como consecuencia de la diversidad de modelos en el entorno de las bases de datos se ponen de manifiesto al comprobar la escasa riqueza de los modelos de implementación. Esta queda visible al comprobar que algunas bases de datos tienen únicamente una tabla donde toda la información se almacena, como puede comprobarse en la Figura 2.10. Por otro lado, como puede apreciarse en la Figura 2.11, existen otras bases de datos que son organizadas en diferentes tablas entre las que pueden aparecer de forma explícita o implícita ciertas relaciones y, en ocasiones, sin hacer explícitas las claves primarias y/o las extranjeras (Vilches-Blázquez *et al*., 2007a). No puede obviarse la existencia de bases de datos modeladas conforme a los diferentes niveles de las formas normales (Date, 1993). Estas formas garantizan la ausencia de problemas como la redundancia, ambigüedad y permiten manipular los datos (insertar – modificar – borrar) sin dar lugar a incoherencias.

La disparidad en el proceso de modelado también queda manifiesta a nivel de tabla y de organización de la información contenida en las mismas, es decir, a nivel de instancias. En las diferentes bases de datos tratadas en este trabajo (descritas en el Anexo) se aprecia, en ocasiones, una mezcla arbitraria de información en las diferentes tablas de dichas bases de datos. A modo de ejemplo, en la base de datos perteneciente a BTN25 se pueden encontrar instancias como "Río de Majones o de Fago" y "Perretano Erreka" en una tabla denominada "Rambla" o "Arroyo de las Rozas" en la tabla "Cauce Artificial", mientras en BCN200 dentro de tabla "Ejes de río" aparecen instancias como "Playa de Lairín", "Rambla de las Amoladeras" o "Canal del Regalatxo". Esta disparidad en el proceso de modelado es producto, según Chou (2005), de la organización de los datos en columnas de bases de datos relacionales y, con frecuencia, del entendimiento del significado semántico de cada columna por parte, únicamente, de los creadores de dichas bases de datos. El problema es aún más complicado consecuencia de que los nombres de las columnas no son únicos ni en una ni en diferentes bases de datos. Por tanto, la transparencia, con frecuencia, desaparece, debido a que resulta difícil para los usuarios consultar las diferentes bases de datos sin entender cómo están estructuradas y es imposible para las máquinas consultar las bases de datos sin explicitar de qué columnas recuperar la información.

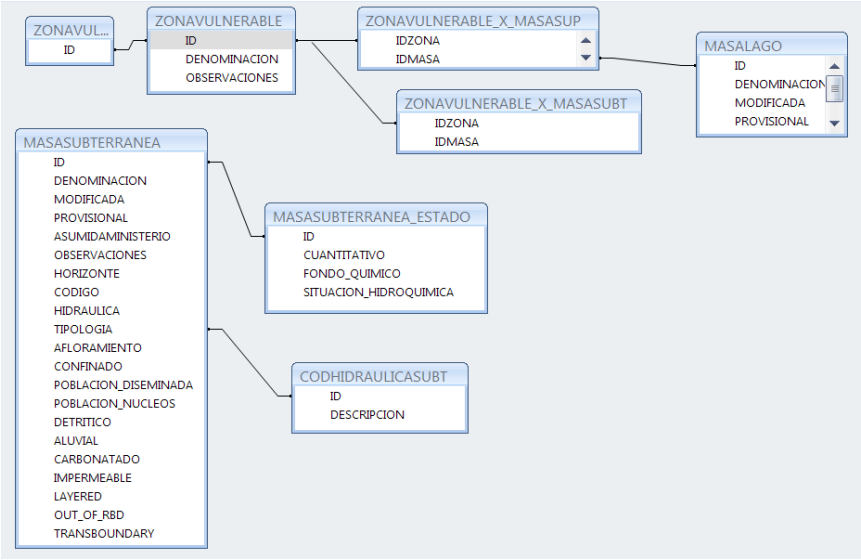


Figura 2.11. Extracto del modelo de la base de datos de la Confederación Hidrográfica del Ebro

Por otro lado, en referencia a la captura de la semántica asociada a los datos, todos los modelos clásicos asociados a bases de datos geográficas han fracasado en el intento de dicha captura, consecuencia de que se construyen registros o relaciones pero éstas no constituyen una unidad semántica atómica.

Además, estas situaciones de múltiples modelos y esquemas combinados con la dificultad para el establecimiento de consensos sobre la representación de los datos en los diferentes dominios de aplicación hace de los datos espaciales, especialmente respecto a otros tipos de datos científicos, una cuestión no trivial respecto al problema de la interoperabilidad (Lemmens *et al*., 2006). Ante esta cuestión, según Hakimpour (2003), un modelo estándar o uniforme para la representación de datos espaciales, tales como los modelos canónicos sugeridos en (Worboys *et al*., 1991), fueron considerados un paso hacia una solución a los problemas de heterogeneidad en modelado de datos. Sin embargo, la baja flexibilidad y la alta complejidad para la administración pueden causar más problemas de los que solventan. Además, es poco probable que los desarrolladores de software estén forzados a utilizar un enfoque de modelado específico, independientemente de lo bueno del enfoque (Goodchild *et al*., 1997). En la actualidad, las especificaciones del OGC, tales como (OpenGIS Consortium, 1999b) o *Geography Markup Language* (GML) (Portele, 2007), y diferentes proyectos en el contexto de EuroGeographics[28] (RISE[29], EuroRegionalMap[30], EuroGlobalMap[31], EuroSpec[32], etc.), INSPIRE (DataSpecification[33]) y otras iniciativas de armonización (INTERLIS[34], CORINE Land Cover[35], GSDBE[36], *Seamless Administrative Boundaries of Europe* (SABE)[37], MeanInGs, etc.) proponen y/o trabajan para solventar problemas de heterogeneidad presentes en el modelado espacial, planteando vías para la armonización de datos espaciales y/o la adopción de un modelo de datos geográfico común o, al menos, un marco para unificar modelos heterogéneos. En este sentido, la armonización del modelado, presentada a los usuarios a través de interfaces uniformes, también requiere una vista común de la interpretación de los datos, es decir, una

[28] *EuroGeographics*. http://www.eurogeographics.org/
[29] *RISE*. http://www.eurogeographics.org/eng/03_RISE.asp
[30] *EuroRegionalMap*. http://www.eurogeographics.org/eng/03_projects_euroregionalmap.asp
[31] *EuroGlobalMap*. http://www.eurogeographics.org/eng/03_projects_EGM_overview.asp
[32] *EuroSpec*. http://www.eurogeographics.org/content/about-us-eurospec
[33] *DataSpecification*. http://www.epsiplus.net/news/inspire_data_specifications_published
[34] *INTERLIS*. http://www.interlis.ch
[35] *CORINE*. Land Cover. http://www.eea.europa.eu/publications/COR0-landcover
[36] *GSDBE*. Georeferenced Soil Database for Europe. http://eusoils.jrc.ec.europa.eu/esbn/GSDE.html
[37] *SABE*. http://directory.eoportal.org/get_announce.php?an_id=4834

interpretación semántica común. Junto con sus otras iniciativas, OGC inició una iniciativa con objeto de resolver problemas semánticos (OpenGIS Consortium, 1999a), a la que se une, en la actualidad, el trabajo iniciado dentro del Comité Técnico 211 de la Organización Internacional para la Estandarización (ISO) (ISO TC/211) respecto al proyecto de norma ISO 19150.

En definitiva, los sistemas geoespaciales distribuidos poseen un gran número de bases de datos desarrolladas de forma independiente. Estas bases de datos contienen diferentes visiones del mundo donde se pueden encontrar diferentes representaciones, esquemas, terminologías así como diferencias semánticas (Bishr, 1997). Esto provoca que las bases de datos puedan ser heterogéneas a muchos niveles. Así, la heterogeneidad esquemática (Bishr, 1998) o heterogeneidad en el modelo conceptual (Hakimpour, 2003), heterogeneidad sintáctica (Bishr, 1998), la heterogeneidad del modelo de datos y en el proceso de transacciones (Morocho Zurita *et al*., 2001), la heterogeneidad estructural (Chatterjee *et al*., 1995; Kim *et al*., 1991; Kashyap *et al*., 1996; Naiman, *et al*., 1995; Worboys *et al*., 1991; Batini *et al*., 1986) o de esquema (Hakimpour, 2003) o disparidad estructural (Spaccapietra *et al*., 1994), los conflictos estructurales (Batini *et al*., 1986), la heterogeneidad en restricciones de integridad (Morocho Zurita *et al*., 2001), la heterogeneidad representacional (Chou, 2005) y la pragmática (Euzenat *et al*., 2007) están presentes entre las disparidades que aparecen en los modelos de base de datos.

2.3.3. Factor escala

La construcción de representaciones múltiples es uno de los problemas clave en aquellos sistemas encargados de la representación de la IG (Brugger *et al*., 1989). Efectivamente, "la planificación jerárquica con diferentes niveles de detalle para diferentes partes de la tarea, parece ser un elemento común en las formas de actuar de los humanos[38]" (Mark, 1989). Por este motivo, los fenómenos del mundo real son representados con diferentes grados de detalle, dependiendo de la escala de las bases de datos geográficas donde son almacenados y representados.

Estrictamente hablando, el factor escala (la ratio entre el tamaño de un objeto de un mapa y su tamaño real) es un concepto cartográfico y, por tanto, este no puede aplicarse para abstraer las representaciones almacenadas en las bases de datos geoespaciales (Devogele *et al*., 1998). Asociado a la escala aparecen conceptos como precisión (grado de detalle con el que se toma una medida), exactitud (relación entre una medida y la realidad que se pretende representar) y resolución (el objeto más pequeño que puede ser representado) (Goodchild, 1991; Müller *et al.*, 1995). En el contexto de las bases de datos geográficas, según Devogele *et al.* (1996), estos conceptos están conectados con la noción de escala comúnmente asociada con un mapa.

La principal característica asociada al factor escala es el hecho de que actúa como filtro en la representación cartográfica, así como en catálogos, diccionarios de información geoespacial, bases de datos y otros sistemas. Por esta razón, es importante considerar información a diferentes escalas (por ejemplo; local, regional y nacional), más aún si no se comprende bien la interacción entre los procesos globales, regionales y locales (Sui *et al*., 1999). Esta consideración múltiple o transescalar de información, también, es consecuencia de que la escala y, por tanto, la variación de la granularidad de la información, afecta a la resolución geométrica y semántica (Vilches Blázquez *et al.*, 2007b), o lo que es lo mismo, a la comprensión de la realidad geográfica, consecuencia de que no todos los fenómenos son significativos a un nivel determinado de *zoom*. Por ejemplo, en el nivel de escala de una ciudad es necesario visualizar los fenómenos de la red de transporte urbano (calles, avenidas, plazas, etc.), sin embargo, esos fenómenos de la red urbana no son significativos para una red de carreteras a nivel nacional (Lacasta *et al*., 2007).

[38] "*hierarchical planning with different levels of detail for different parts of the task, seems to be common in human way-finding*" (Mark, 1989)

Este hecho, también, se evidencia al comprobar las diferencias entre los catálogos y/o diccionarios de datos de los diferentes productores. Un claro ejemplo de esto aparece en dos instituciones españolas de referencia en cuanto a la producción de IG se refiere. El *Institut Cartogràfic de Catalunya* (ICC) posee 49 clases de fenómenos en su catálogo, considerando información a una escala 1:5.000. Por otro lado, el Instituto Geográfico Nacional (IGN), en el catálogo de la Base Cartográfica Numérica – BCN25 a escala 1:25.000, sólo para el tema referido a hidrografía posee unas 40 clases de fenómenos, recogiendo unas 685 clases en su totalidad. De esta manera, queda reflejada la consonancia de esta situación con la descrita en (Bishr, 1997) referente a las diferencias en el proceso de abstracción y al nivel de detalle de la información.

Además, la representación/existencia de diferentes fenómenos a diferentes escalas es un problema añadido en el proceso de generalización cartográfica. Desde el punto de vista de la representación, este proceso se encarga de modificar los datos para producir una representación simplificada o más abstracta con objeto de conseguir legibilidad en el mapa (Lagrange *et al*., 1994). Mientras, desde el punto de vista de la implementación, la generalización aparece porque los conceptos y sus instancias asociadas existen a diferentes niveles de abstracción en dos bases de datos distintas (Naiman, Ouksel, 1995). Estas características del proceso de generalización cartográfica conllevan dificultades en los procesos de superposición tanto gráfica como conceptual entre diferentes fuentes de información. Más detalles sobre el proceso de generalización cartográfica y otros aspectos relacionados con el mismo pueden ser encontrados en (Devogele, *et al*., 1996; Govorov, 1995; Müller *et al*, 1995; Stoter *et al*., 2008; Duchêne *et al*., 2008).

Estas peculiaridades del factor escala conllevan diferentes tipos de heterogeneidad, tales como la heterogeneidad esquemática (Bishr, 1998) o heterogeneidad en el modelo conceptual (Hakimpour, 2003; Naiman, *et al*., 1995; Worboys *et al*., 1991; Yeung *et al*., 2007), la heterogeneidad en el modelado espacial (Hakimpour, 2003) y diferentes aspectos concernientes a la heterogeneidad semántica, tales como la granularidad (Naiman, *et al*., 1995; Goh, 1997; Sheth *et al*., 1992; KnowledgeWeb Consortium, 2005; Euzenat *et al*., 2007), significado (Worboys *et al*., 1991) y pragmática (Worboys *et al*., 1991; Yeung *et al*., 2007).

2.3.4. Lenguaje natural

El Mundo Conceptual es el mundo del lenguaje natural. Las personas al ver y reconocer las cosas podemos nombrarlas y este conjunto de entes y hechos es, con frecuencia, llamado "*universo de discurso*". Este Mundo Conceptual no es suficientemente abstracto para los actuales sistemas que soportan la IG, como son los SIG y las IDE. En este sentido, los sistemas sólo abstraen un conjunto simplificado del Mundo Conceptual que resulta de interés. Este subconjunto es denominado Mundo Geoespacial (OpenGIS Consortium, 1999b). Dentro de este Mundo Geoespacial las categorías geográficas muestran importantes diferencias culturales producto de la subjetividad de las percepciones y las convenciones sociales (Smith *et al*., 1998; Mark *et al*., 2003a, b).

Las comunidades de información (OpenGIS Consortium, 1999a; OpenGIS Consortium, 1999b; Bishr *et al*., 1999) de usuarios y productores de geodatos utilizan vocabularios y semántica propios. Esto provoca:

> *"La IG es un ejemplo de información mantenida públicamente con alto potencial. ..., un número de cuestiones prácticas convierte la explotación de información del sector público en problemática europea. ... La necesidad de traducir y la ausencia de una terminología común ponen una carga extra sobre la firma del contenido europeo"*[39] *(European Commission, 2000a).*

[39] "*Geographic Information is an example of publicly held information with high potential. ..., a number of practical issues make the exploitation of public sector information in Europe problematic. ... The need*

Esta disparidad puede llevar a serios problemas cuando el significado intuitivo de los componentes de la terminología no está claro y el entendimiento de la información proporcionada difiere entre los usuarios (Nowak *et al*., 2005). Situaciones de este tipo son provocadas por la ambigüedad del lenguaje natural, importantes diferencias semánticas y la diversidad provocada por el multilingüismo.

2.3.4.1. Ambigüedad del lenguaje

La terminología es un conjunto estructurado de conceptos y términos de un dominio específico en un determinado idioma (Nowak *et al*., 2005). Ésta es una cuestión inseparable de la diferenciación de fenómenos dentro del mundo debido a la disparidad de consideraciones culturales, a las que se le unen, importantes diferencias históricas, geográficas y sociales. En el dominio geográfico, además, la terminología es utilizada para estructurar (catálogos de fenómenos y tesauros) y almacenar (bases de datos) la información geoespacial. Las influencias a las que está sometida la terminología traen consigo inconsistencias en como los términos son utilizados en las diferentes fuentes.

La selección de un vocabulario adecuado representa un importante reto en términos de interoperabilidad (Heath *et al.*, 2005). Por tanto, los términos de vocabularios controlados (listas controladas, taxonomías, tesauros, etc.) son, con frecuencia, recomendadas para armonizar los datos y metadatos de los sistemas que gestionan información geoespacial, así como para mejorar la calidad de los resultados de las consultas. Sin embargo, a pesar de la ventaja que supone la utilización de vocabularios controlados, ciertos problemas de ambigüedad persisten en el lenguaje natural (Lacasta *et al*., 2007). Esta ambigüedad es, principalmente, causada por las variadas relaciones semánticas entre términos, tales como polisemia (*pluralidad de significados de una palabra o de cualquier signo lingüístico*), sinonimia (*dicho de un vocablo o de una expresión: Que tiene una misma o muy parecida significación que otro*), hiperonimia (*palabra cuyo significado incluye al de otra u otras; p. ej., pájaro respecto a jilguero y gorrión*), homonimia (*dicho de una palabra: Que, siendo igual que otra en la forma, tiene distinta significación; p. ej., Tarifa, ciudad, y tarifa de precios*), hiponimia (*palabra cuyo significado está incluido en el de otra; p. ej., gorrión respecto a pájaro*) y meronimia (*palabra cuyo significado constituye una parte del significado total de otra palabra, denominada ésta holónimo. Por ejemplo, dedo es merónimo de mano y mano es merónimo de brazo; a su vez, brazo es holónimo de mano y mano es holónimo de dedo*). Estos problemas aparecen en muchos conceptos en la IG, debido a que no existe un marco común sobre convenciones de nombrado. Prueba de esto es el hecho de que podemos encontrar diferentes catálogos y/o diccionarios con una gran variedad de conceptos asociados al mismo fenómeno (por ejemplo; "río", "curso fluvial", "curso de agua principal", etc.).

De esta manera, en los actuales sistemas de búsqueda basados en palabras clave, por ejemplo los diferentes servicios de nomenclátor de las variadas IDE existentes, un sinónimo de una determinada palabra clave incluido como parte de una consulta, puede no devolver respuesta, a pesar de que términos diferentes aunque similares puedan ser utilizados para describir información similar en el nivel intensional (por ejemplo, el término "contenido" (*subject*) puede ser modelado como "palabras clave" (*keywords*) en diferentes fuentes de información) así como a nivel extensional (donde dos representaciones diferentes encarnen el mismo valor; "10 de Febrero, 1998" y "2/10/98") (Mena *et al*., 2000). Este hecho está relacionado con la ausencia de criterio para asegurar que la información almacenada en las bases de datos con su correspondiente vocabulario, sea el mismo que utilizan los usuarios en los procesos de búsqueda (Chou, 2005).

Los problemas que afectan a la ambigüedad del lenguaje se centran en la heterogeneidad semántica. Dentro de este tipo de heterogeneidad se distinguen los siguientes tipos de

to translate and the absence of a common terminology put an extra burden on the European content firms. " (European Commission, 2000a)

disparidades; heterogeneidad designativa (Bishr, 1998; Goh, 1997; Naiman, *et al*., 1995; Chatterjee *et al*., 1995; Batini *et al*., 1986), heterogeneidad referida a los puntos de vista (Hakimpour, 2003), heterogeneidad de representación (Yeung *et al*., 2007), heterogeneidad a nivel terminológico (Navarrete, 2006; KnowledgeWeb Consortium, 2005) y heterogeneidad representacional (Chou, 2005).

2.3.4.2. Diferencias semánticas

Las diferencias semánticas en el dominio de la IG son comunes y numerosas, derivadas de las diferentes percepciones o puntos de vista sobre un mismo tipo de fenómeno. Este tipo de diferencias son consecuencia directa de los variados contextos o entornos geográficos, es decir, de la desigual plasmación que cada fenómeno tiene sobre un determinado espacio. El factor contexto genera diferencias e interpretaciones dispares y, por tanto, diferentes abstracciones de los fenómenos geográficos sobre los que se localiza la información y de la que son capturados sus datos (Worboys *et al*., 1991).

En la práctica las diferencias semánticas quedan plasmadas en las definiciones de los diferentes fenómenos. Un ejemplo iterativo es la definición del fenómeno "Río". La Directiva Marco del Agua (European Parliament *et al*., 2000) lo define como "*una masa de agua continental que fluye en su mayor parte sobre la superficie del suelo, pero que puede fluir bajo tierra en parte de su curso*". Para el Instituto Geográfico Nacional es una "*corriente natural de agua de caudal considerable aun en estiaje*". El *Ordnance Survey* (Ordnance Survey, 2001) define a este tipo de fenómeno como "*agua que fluye en un canal definido hacia el mar, un lago u otro río*"[40], Mientras, para una fuente lingüística de referencia, aunque no geográfica, caso del Diccionario de la Real Academia de la Lengua este fenómeno es definido como "*corriente de agua continua y más o menos caudalosa que va a desembocar en otra, en un lago o en el mar*".

Las diferencias semánticas, obviamente, están estrechamente relacionadas con la heterogeneidad semántica de la información geoespacial. Entre los diferentes tipos de heterogeneidad relacionados con este problema destacan los conflictos de confusión (Goh, 1997), la heterogeneidad de contexto (Yeung *et al*., 2007) y la heterogeneidad semiótica o pragmática (Euzenat *et al*., 2007).

2.3.4.3. Multilingüismo

El multilingüismo y, consecuentemente, las bases de datos con información multilingüe conforman un elemento importante en la barrera que sirve de obstáculo a la interoperabilidad de la información geoespacial. Esto puede ser debido, según Aguado de Cea *et al*. (2007), a que, hasta ahora, la representación del multilingüismo no era un objetivo primordial sino que el interés se centraba meramente en su existencia. Es decir, en los recursos y sistemas asociados con la información geoespacial, la consideración del multilingüismo era una propiedad secundaria, más enfocada a la explotación de los mismos que a la conformación del propio recurso.

El uso generalizado de la IG junto a la perspectiva de la globalización está cambiando este panorama y el componente multilingüe comienza a adquirir la importancia que le confiere la explotación actual de la información. Este hecho queda constatado con el surgimiento de proyectos como EuroGeoNames[41].

[40] *Definición del fenómeno río*: "water flowing in a definite channel towards the sea, a lake or into another river" (Ordnance Survey, 2001).

[41] *EuroGeoNames*. Su objetivo es recoger y poner en un servicio plurilingüe de nomenclátor los nombres geográficos –en especial los normalizados por las autoridades competentes- distribuidos a nivel nacional y mantenidos por los organismos de cartografía y catastro de los diferentes Estados miembros de la Unión Europea. http://www.eurogeographics.org/news/eurogeonames

Desde una óptica global, a nivel de servicios, los Sistemas de Información Geográfica (SIG) e IDE, cada vez más, afrontan una perspectiva internacional y multilingüe ante usuarios de diferentes países que acceden a las mismas fuentes de datos (Ostländer *et al*., 2005). Mientras, a nivel de los datos, los diferentes recursos de información geoespacial se enfrentan al denominado "proceso de localización". La localización de un recurso léxico (diferentes nombres asociados a las instancias de fenómenos geográficos) se entiende como el proceso de adaptación-contextualización de ese recurso a una comunidad lingüística y cultural concreta, es decir, a un espacio geográfico determinado. Según Aguado de Cea *et al*. (2007) el multilingüismo y, por tanto, ese proceso de adaptación-contextualización puede darse en tres niveles en los actuales sistemas, concretamente, a nivel de interfaz, representación de conocimiento y datos. A continuación se describen en detalle cada uno de estos niveles.

Interfaz

El multilingüismo de la interfaz, puede o no, conllevar agregación multilingüe al resto de la aplicación. Además, esta puede darse en visualizaciones de mensajes o visualizaciones de contenidos. En las diferentes interfaces de los diversos sistemas que soportan la IG aparecen casos dispares. Por un lado, existen caso como el referido al interfaz del geoportal de la Infraestructura de Datos Espaciales de España (IDEE)[42]. Este geoportal se encuentra en 7 idiomas diferentes (inglés, francés, portugués, gallego, catalán, euskera y castellano), pero las interfaces correspondientes a servicios fundamentales como "búsqueda de nombres geográficos", "visualización de mapas" o "creación de simbología" presentan interfaces sólo bilingües, concretamente, aparecen en español e inglés. Por otro lado, aparecen casos intermedios como los referidos a los interfaces de la Infraestructura de Datos Espaciales de La Rioja[43] donde la información aparece en español e inglés en sus diferentes interfaces, inclusive en servicios tan importantes como los referidos al visualizador (leyendas, *tooltips*, etc.). En último lugar, aparecen casos como el referido al geoportal de la Infraestructura de Datos Espaciales de Andalucía (IDEA)[44], donde la monolingualidad queda patente en sus diferentes interfaces.

Respecto a la visualización del contenido de forma multilingüe, según Aguado de Cea *et al*. (2007), hay dos posibles vías: a) la existencia de una base de conocimiento multilingüe; dentro del dominio geográfico este es el caso de bases de datos como BTN25, cuyas instancias aparecen en 4 de los 5 idiomas oficiales de España (catalán, gallego, euskera y castellano); o b) la base de conocimiento es monolingüe; esta situación, de forma generalizada, se encuentra en las bases de datos de productores cartográficos de Comunidades Autónomas monolingües, por ejemplo, la base de datos referida al dominio hidrográfico de la Junta de Andalucía.

Representación del conocimiento

El multilingüismo en una representación del conocimiento tiene que abordarse en tres niveles diferentes: información, modelado y realización.

I. La información sobre el multilingüismo de una representación del conocimiento forma parte de los metadatos de dicha representación. De esta manera, la información que se refiere a los datos de la representación del conocimiento se encuentra expresada en varios idiomas. En el proceso de almacenamiento de la información sobre multilingüismo existen dos soluciones:
    a. Concepto lingüístico; Dentro de los modelos se dispone el concepto idioma y se asocia con los conceptos de la fuente de información.
    b. b) Modificación del concepto; se incluye dentro del modelo un ítem de metadatos con valor múltiple y ese ítem se representa por medio de tuplas de la lengua y del componente que está en dicha lengua.

[42] *IDEE*. http://www.idee.es
[43] *IDE La Rioja* http://www.iderioja.larioja.org/
[44] *IDEA*. http://www.andaluciajunta.es/IDEAndalucia/IDEA.shtml

II. Modelado, es la representación de los componentes que darán multilingüismo a la representación del conocimiento. Admiten tres posibles soluciones:
    a. Ampliación con información lingüística al metamodelo. En los modelos actuales concernientes a la información geoespacial la consideración de información multilingüe no se tiene en cuenta a nivel de modelo, simplemente ese tipo de información se encuentra, de forma explícita, a nivel de instancias.
    b. Agregación de un modelo de información lingüística que se relacionará con un metamodelo para representar el multilingüismo. Esta propuesta ofrece una importante riqueza de información multilingüe, aunque desde la óptica de la IG, puede estar demasiado centrado en la componente lingüística de la información.
    c. Utilización de un modelo de correspondencias para relacionar información monolingüe y obtener información multilingüe. El gran problema de esta solución es que ningún sistema contempla correspondencias de cardinalidad superior a dos, salvo soluciones *ad hoc* para operaciones concretas. Otra desventaja de esta opción es que habría que repetir la información conceptual en todas las fuentes monolingües.

III. Realización, es la expresión lingüística real del multilingüismo en la representación del conocimiento.

Datos

En el ámbito de las instancias se presenta el último escalón en el que el multilingüismo aparece. Ésta viene condicionada, fundamentalmente, por la propia naturaleza del contexto de los datos y por el modelado de representación del conocimiento considerado para el dominio. A este nivel, consecuencia del surgimiento de iniciativas como la Directiva INSPIRE (Commission of the European Communities, 2007) para la construcción de la Infraestructura de Datos Espaciales de Europa[45] y su transposición y adopción por cada uno de los miembros de la Unión Europea, dos puntos de vista (global y local) influyen sobre los aspectos multilingües de los datos y la necesidad de integrar información:

- Desde un punto de vista global, Europa es un territorio compuesto por un gran número de países y regiones caracterizadas por su propia cultura e idioma. Tras la última ampliación (Mayo de 2004) hay 20 idiomas oficiales europeos en el conjunto de los Estados miembros (Nowak *et al.*, 2005).
- Desde un punto de vista local, en España, la presencia de cinco lenguas oficiales (catalán, gallego, aranes, euskera y castellano), que son según la Constitución Española (Artículo 3) patrimonio lingüístico de España y, por tanto, objeto de especial respecto y protección, genera importantes diferencias en cuanto a producción de información. Especialmente, esta disparidad se presenta entre las fuentes de información (catálogos de fenómenos y bases de datos) de las regiones multilingües frente a las monolingües.

Esta realidad multicultural, caracterizada por entornos distribuidos, provoca, según Nowak *et al*. (2005), diversos problemas que van a aparecer derivados de diferentes tipos de heterogeneidad y que van a dificultar la consecución de la interoperabililidad. En el contexto de la heterogeneidad estructural uno de los problemas es la vaguedad de las definiciones, consecuencia de la utilización de relaciones no taxonómicas para definir un fenómeno. Por ejemplo, en una base de datos espacial de la República Checa puede aparecer un fenómeno llamado "*reka*" para la descripción del ente real "río" y un fenómeno denominado "*potok*" para la descripción de un "río pequeño" o "arroyo". Sin embargo, la definición subjetiva de ríos grandes o pequeños puede ser completamente diferente en cualquier otro contexto cultural e idiomático.

[45] *INSPIRE*. http://www.inspire-geoportal.eu/

Otro conflicto típico es la diferencia en el nivel de la generalización de los fenómenos. Esto ocurre cuando una clase que se define de forma individual en una base de datos puede ser incluida en otra base de datos como parte de una clase más genérica. Por ejemplo, en una base de datos se puede encontrar un grano fino de fenómenos para "cruces de caminos y pequeñas capillas" (en una base de datos polaca "*krzyze i kapliczki przydrozne*") que puede ser incluida en otra base de datos dentro de un fenómeno más general del tipo "monumentos de patrimonio cultural". Esta distinción de grano está basada en la historia cultural de un país, por tanto, la importancia de un tipo específico de fenómeno puede influir en la creación de un fenómeno individual que no se distingue en otros países. Esto también ocurre entre bases de datos que definen, por ejemplo, tres diferentes fenómenos para tres objetos reales ("*rzeka*", "*kanal*" y "*rów odplywowy*"[46]) y otra base de datos distinta define sólo un fenómeno denominado "río" con un atributo tipo para diferenciar varios objetos (por ejemplo, el atributo tipo puede tener tres valores diferentes: "río", "canal" y "sumidero"). Otro ejemplo similar podría ser la existencia de un fenómeno denominado "carretera" frente a la existencia única de fenómenos más genéricos como "redes de transporte interior".

Otros conflictos pueden ocurrir en las diferencias encontradas en los atributos de clase, es decir, cuando dos clases similares semánticamente son caracterizadas por diferentes listas de propiedades. Por ejemplo, un fenómeno denominado "ciudad" y definido por una bases de datos en el contexto español puede incluir como atributos los códigos de Comunidad Autónoma y provincia. Sin embargo, un fenómeno definido en otros países con otra estructura jerárquica en cuanto a división administrativa incluiría sólo un único código para el concepto equivalente a provincia.

Los conflictos de nombrado también están presentes como problemas en la información multilingüe. Así, el concepto "lago" es expresado como "*lake*" en inglés, "*tó*" en húngaro y "*jezioro*" en polaco. Además, hay un segundo nivel asociado a estos conceptos, vinculado con los sinónimos asociados a dichos conceptos en sus respectivos idiomas; por ejemplo "*flod*" o "*elv*" (términos daneses para referirse a "río") o "*tanque*", "*lagoa*" y "*depósitos de àgua*" (términos portugueses para referirse a "estanque").

En el caso opuesto, es decir, el uso de términos iguales o similares para información distinta se introduce homónimos que representan la heterogeneidad conceptual o cognitiva (Kuhn, 2005). Por ejemplo, cuando el término "lago" comprende "lago natural", "estanque de pesca artificial" y "lago inundable" (cuya equivalencia en inglés es "*lake*", "*natural lake*", "*artificial fish-pond*" y "*inundate lake*" y en polaco corresponde a "*jezioro*", "*jezioro*", "*staw rybny*" y "*zalew*" respectivamente).

En un contexto más concreto, como es el referido a las fuentes de información controladas en este trabajo (ver Anexo), la información multilingüe se encuentra desestructurada en las diferentes bases de datos en las que está presente. Esto se produce como consecuencia de que en una misma tabla pueden aparecer instancias pertenecientes a las cinco lenguas oficiales de España. Además, en unos casos el idioma está explícitamente contemplado en un campo de la tabla de la base de datos, mientras, en otros simplemente se encuentran nombres de instancias sin especificación del idioma correspondiente. Asimismo, aparecen instancias duplicadas en diferentes idiomas, por ejemplo, "Río Ebro" y "*Riu Ebre*" sin que aparezcan vínculos entre ellas, por tanto, dichas instancias son consideradas información distinta resultado de su disociación. Esta situación influye en los resultados ofrecidos por los servicios de búsqueda y visualización, incluso a pesar de la creencia de que la visualización de los datos es independiente del idioma.

[46] Términos polacos para referirse a los fenómenos hidrográficos "río", "canal" y "sumidero", respectivamente.

Las características de la mutilingualidad y sus problemas subyacentes conllevan diferentes tipos de disparidad, destacando la heterogeneidad de terminológica (Bishr, 1998; Goh, 1997; Naiman, *et al*., 1995; Chatterjee *et al*., 1995; Batini *et al*., 1986; Navarrete, 2006; KnowledgeWeb Consortium, 2005), heterogeneidad referida a los puntos de vista (Hakimpour, 2003), heterogeneidad de representación (Yeung *et al*., 2007; Chou, 2005), esquemática (Bishr, 1998; Naiman, *et al*., 1995; Worboys *et al*., 1991), estructural (Chatterjee *et al*., 1995; Kim *et al*., 1991), pragmática (Worboys *et al*., 1991; Yeung *et al*., 2007) y de granularidad (Goh, 1997).

2.4. Conclusiones

En este capítulo se ha descrito una serie de conocimientos previos para ayudar a la compresión de diversas cuestiones tratadas en este trabajo. Así, se realiza una breve descripción del concepto interoperabilidad, centrándose en los tres tipos de interoperabilidad más extendidos en la bibliografía científica (sintáctica, estructural y semántica), y se recogen algunas consideraciones básicas relacionadas con las ontologías.

Además, en la sección 2.2, se recoge un análisis de los tipos de heterogeneidad presentes en el contexto de la información geográfica. Con respecto a los tipos de heterogeneidad, estos se analizan en el contexto de la información geográfica y en otros tipos de información (bases de datos y modelos de conocimiento/ontologías) relacionados con el dominio geoespacial. Este análisis recoge una serie de categorizaciones de los distintos tipos de heterogeneidad y su vinculación con diferentes problemas de la información geoespacial, cubriendo así la **carencia de una caracterización exhaustiva del tipo de disparidades** (categorizadas en niveles, por analogía a las propuestas presentadas) **en el contexto de dicha información**.

Finalmente, se describen, desde una perspectiva genérica, diversos problemas presentes en este tipo de información (sección 2.3). Los problemas tratados se refieren a la estructuración de la información (sección 2.3.1), los presentes en bases de datos en torno a los datos y diversidad de modelos (sección 2.3.2), el factor escala (sección 2.3.3) y los relativos a la ambigüedad del lenguaje, diferencias semánticas y multilingüismo (sección 2.3.4). Asimismo, se presenta una **visión general de los diferentes problemas analizados vinculados con los diversos tipos de heterogeneidad**.

Capítulo 3

Estado de la cuestión

En este capítulo se recoge el estado de la cuestión en aquellas áreas directamente relacionadas con este trabajo. Dicho repaso se organiza en torno a cuatro ejes fundamentales:
El primer eje recoge un análisis de las **ontologías** desde su vinculación con el **dominio geográfico** (sección 3.1). Además, en esta sección se lleva a cabo un repaso de un conjunto significativo de ontologías geoespaciales y se pone especial atención en el análisis de diversas ontologías vinculadas al dominio hidrográfico. El segundo eje (sección 3.2) se centra en el análisis del proceso de **integración de información** y, en particular, en cuestiones como su definición, sus diferentes problemas y sus diversas propuestas de modelado existentes. Asimismo, en esta sección se recogen los principales tipos de procesos de integración de información (sintáctico, estructural y semántico) y múltiples propuestas presentadas tanto desde una perspectiva general como desde un enfoque vinculado al dominio geoespacial. El tercer eje trata la cuestión de la **detección de duplicados** (sección 3.3) en el proceso de integración de información. En esta sección se describe el tratamiento de la detección de duplicados en diversas comunidades científicas (estadística, bases de datos e inteligencia artificial). Además, se presentan algunos de las principales propuestas en el contexto de la Web de los Datos y de la información geográfica. El cuarto eje se ocupa de la iniciativa *Linking Open Data* y de las diferentes propuestas que publican información geoespacial conforme a los principios de ***Linked Data*** (sección 3.4). Finalmente, la sección 3.5 recoge las **conclusiones** del capítulo y pone de manifiesto las **limitaciones** de las diferentes propuestas estudiadas y los problemas de investigación abiertos.

3.1 Ontologías geográficas

Esta sección introduce brevemente el concepto geo-ontología, sus componentes principales (secciones 3.1.1 y 3.1.2) y diversas tipologías en las que se clasifican las ontologías geográficas (sección 3.1.3). Finalmente, se aportan diversas propuestas de recursos ontológicos en el ámbito de la información geográfica e hidrográfica (sección 3.1.4).

3.1.1. ¿Qué son las geo-ontologías?

Desde un punto de vista general se puede percibir a la geografía y las ontologías como dos áreas muy lejanas en su objeto y metodología. La geografía es una ciencia práctica que busca entender la realidad concreta del mundo. Las ontologías, por el contrario, pertenecen a una disciplina abstracta interesada por el significado de los conceptos, pero apenas sin considerar las circunstancias particulares de la realidad. Sin embargo, ambos implican sistematización del mundo; mientras que la geografía tiene por objeto representar la estructura física del mundo, las ontologías pretenden estructurar conceptualmente las formas de describir el mundo. Esta

coincidencia implica una relación estrecha entre ambas áreas, ya que no se puede esperar comprender completamente el mundo a menos que entendamos la forma de describirlo. Esta estrecha relación, junto con la *cuestión de moda* de las ontologías en prácticamente todas las áreas de procesamiento de la información, deriva en el surgimiento de las ontologías geográficas o geo-ontologías como una nueva rama de investigación relacionada con la IG. A continuación, se recoge una doble perspectiva del concepto geo-ontología. Por un lado, una visión general y técnica que surge desde el ámbito de ISO/TC211 y, por otro lado, una perspectiva más específica y vinculada con el dominio geoespacial.

Desde una visión general, ISO/TC211 afirma que una ontología en el dominio geográfico se refiere a la representación formal de un fenómeno con un vocabulario subyacente. Esta representación incluye definiciones y axiomas que hacen explícito el significado pretendido y que sirven para describir al fenómeno y a aquellas interrelaciones que puedan ser utilizadas por las aplicaciones software para dar soporte a la compartición, reutilización e integración de información geográfica con cualquier otra fuente de información dentro de un dominio de conocimiento, así como entre varios dominios de conocimiento.

Desde una perspectiva más relacionada con el dominio, (Fonseca *et al*., 2006) afirma que el concepto geo-ontología tiene dos tipos básicos de conceptos. Por un lado, los conceptos correspondientes a los fenómenos físicos en el mundo real. Por otro, los conceptos que corresponden a fenómenos del mundo que han sido creados para representar construcciones institucionales y sociales. De esta manera, los conceptos en una geo-ontología se dirigen hacia los objetos espaciales presentes en el mundo.

Estas visiones se pueden resumir afirmando que una ontología geográfica se diseña para ofrecer un mejor entendimiento de la estructura del mundo geográfico (Smith *et al.*, 1998).

En definitiva, estas definiciones hacen entender que las ontologías no están formadas únicamente por meros conceptos con una cierta organización, lo que no distaría de las formas más habituales de organización de la IG (catálogos de fenómenos, diccionarios de datos y/o tesauros), sino que también se van a definir relaciones, atributos, reglas y axiomas entre conceptos que enriquecen y contribuyen, entre otros aspectos, a formalizar el vocabulario y conocimiento del área, dejando a un lado el simple almacenamiento realizado sobre la parte más específica de la información, es decir, los datos (Vilches *et al*., 2008).

3.1.2. Componentes

En el desarrollo de las geo-ontologías se modelan conceptualizaciones de fenómenos geográficos del mundo real. Estos fenómenos no tienen una simple localización relativa a la Tierra, sino que están intrínsecamente vinculados al espacio, heredando del mismo muchas de sus propiedades estructurales, tales como: mereológicas (la teoría de las relaciones parte-todo), topológicas y geométricas (Smith *et al.*, 1998).

Por otro lado, según Klien *et al*. (2005), las geo-ontologías están formadas, de forma independiente, por conceptos geométricos o geoespaciales. De esta manera, se realiza una distinción entre los conceptos para los fenómenos geoespaciales del mundo real y los conceptos para la representación de los datos (ver
Figura 3.1). Por tanto, estos autores señalan que las representaciones de los fenómenos (punto, línea y polígono) que son necesarias para abstraer el fenómeno del mundo real no son parte de las ontologías geoespaciales, ya que estas tratan con la estructura de implementación de los datos y no con la semántica de un término que alude a un fenómeno del mundo real. Así, por ejemplo, una ciudad puede ser representada como un punto, pero la ciudad del mundo real no tiene ninguna relación ontológica con la estructura representacional de un punto.

La definición de la extensión de una ontología geoespacial resulta complicada, ya que conlleva concretar los conceptos básicos en una conceptualización común. Estos conceptos deben ser suficientemente generales para permitir la anotación de todas las fuentes de información, pero bastante específicos como para lograr definiciones tan significativas como sea posible (Schuster *et al.*, 2001). En consecuencia, las ontologías geoespaciales requieren ser definidas dentro de cierto contexto y para una comunidad de usuarios conocida, es decir, se tienen que lograr subconjuntos manejables y adecuados del conocimiento del dominio geoespacial. En definitiva, la consideración de estos componentes de la información geoespacial es lo que diferencia a las ontologías geográficas del resto de ontologías desarrolladas para otros dominios.

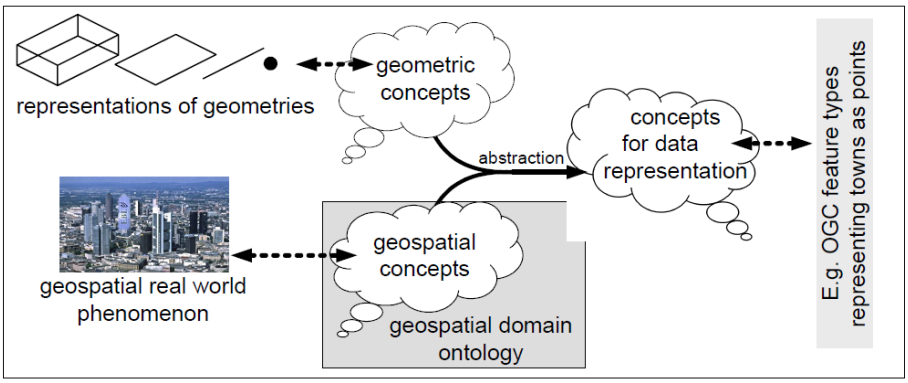


Figura 3.1. Distinción entre los tipos de conceptos que implican las ontologías geoespacial (Klien *et al.*, 2005)

3.1.3. Tipologías de geo-ontolologías

Los tipos de ontologías geográficas no resultan tan prolijos como los existentes desde la perspectiva general de éste concepto. Además, las tipologías de geo-ontologías poseen una estrecha relación con algunos de los tipos generales existentes. En esta sección se recogen, de forma no exhaustiva, diversas tipologías de geo-ontologías.

Según Fonseca *et al.* (2002) las geo-ontologías son divididas en: ontología de dominio fenomenológico (*Phenomenological Domain Ontology* (PDO)) y ontología de dominio de aplicación (*Application Domain Ontology* (ADO)). El primer tipo de ontologías (PDO) captura las diferentes dimensiones y propiedades internas de los fenómenos geográficos. Mientras, el segundo tipo (ADO) afecta a la descripción de tareas y temas específicos que los científicos de la IG utilizan como una fuente de información. En definitiva, PDO está relacionado con como los fenómenos geográficos pueden ser capturados y representados por los sistemas informáticos. Mientras, ADO trata con la descripción del fenómeno en sí, donde encaja en el mundo y como puede ser mejor descrito (ver Figura 3.2).

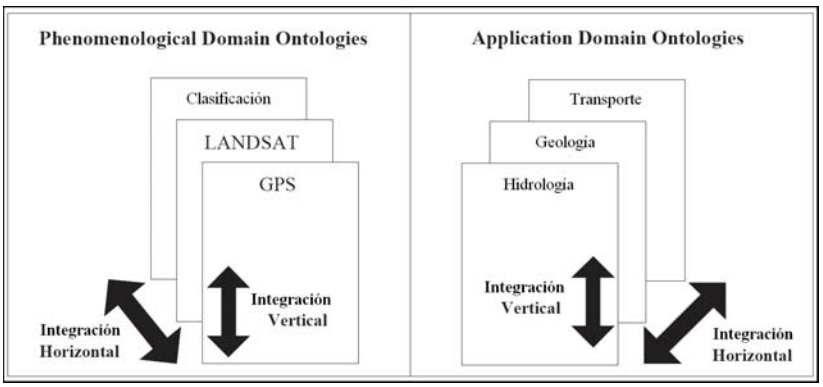


Figura 3.2. Integración vertical y horizontal

Las variadas comunidades de información geoespacial tienen diferentes puntos de vista sobre el mundo. Estas vistas pueden ser formalizadas en diversas ontologías. Por tanto, es necesario acomodar múltiples ontologías para la representación de la realidad geoespacial. Así, en (Fonseca *et al*., 2002) aparecen dos formas de integrar ontologías. La integración vertical está relacionada con la integración dentro de un tema. El otro tipo de integración es denominado integración horizontal e involucra la integración de ontologías de diferentes temas.

En (Lieberman *et al.*, 2007) el *W3C Geospatial Incubator Group* recoge siete categorías diferentes de ontologías geoespaciales. A continuación se describen estas categorías y se listan diferentes ejemplos.

- **Fenómenos geoespaciales** (*Geospatial Features*). Este tipo de ontologías están destinadas a proporcionar una representación formal de los estándares ISO y especificaciones OGC para discernir fenómenos (Modelo de Fenómenos General – MFG). El MFG es el modelo general del que se desarrolla el vocabulario geográfico como un subconjunto restringido. Mientras que la segunda pueda representar adecuadamente una gran proporción de los recursos web, muchas aplicaciones importantes, tales como planificaciones de rutas, observación de la Tierra e interpretación de imágenes requieren una representación más completa. Las sub-ontologías de la ontología de fenómenos incluida en esta categoría cubren geometrías, coberturas (principalmente rasters y grids) y observaciones. Diversas ontologías de este tipo han sido desarrolladas por el *Work Item* sobre Ontologías del Comité ISO/TC211 (NWI211/N2163 "Geographic information — Ontology"). Entre estas ontologías, por ejemplo, aparecen las traducciones de (ISO 19103, 2005) e (ISO 19112, 2003).
- **Tipos de fenómenos** (*Feature Types*). Este tipo de ontología está constituida por los "50 tipos de fenómenos universalmente conocidos y, por tanto, todos estamos de acuerdo". Como un ejemplo de este tipo de ontología podría considerarse la ontología de límites administrativos[1] publicada por el *Ordnance Survey*.
- **Relaciones (geo)espaciales** ((*Geo) spatial Relationship)*. Existen al menos 8 relaciones topológicas 2D comúnmente aceptadas, tales como las relaciones *Region Connection Calculus* (RCC8), pero, probablemente existan otros tipos de relaciones que podrían resultar útiles. Uno de los muchos ejemplos de ontologías de relaciones espaciales[2] ha sido publicado por el Ordnance Survey.
- **Topónimos** (*Places names*). Los nombres de lugares (topónimos) son un medio importante de geolocalización de recursos, al menos de forma aproximada. Sobre este tipo de ontologías, se han realizado algunos trabajos a nivel global y existen algunas taxonomías, pero ninguno de los recursos existentes se puede considerar referencia común para la comunidad. En este contexto, una referencia útil es el trabajo realizado por GeoNames[3].
- **Sistemas de Coordenadas** (*Coordinate Reference Systems*). La utilización de WGS84 como sistema de coordenadas de referencia parece ser suficiente a nivel mundial. Sin embargo, otros sistemas de referencia pueden resultar útiles en el ámbito local o para conseguir diferentes vistas del mundo (por ejemplo, sistemas de coordenadas polares). OGC e ISO proporcionan estándares, como por ejemplo, (ISO 19111, 2003) para modelos y codificación XML de los sistemas de coordenadas. Una primera versión de la ontología de este estándar[4] fue desarrollada por la Universidad de Drexel.
- **Metadatos geoespaciales** (*Geospatial Metadata*). La calidad y origen de los datos son fundamentales para obtener beneficios reales de la información geoespacial. Un

[1] *Administrative ontology*. http://www.ordnancesurvey.co.uk/oswebsite/ontology/v1/AdministrativeGeography.owl
[2] *Spatial relations ontology.* http://www.ordnancesurvey.co.uk/oswebsite/ontology/SpatialRelations.owl
[3] *Geonames*. http://www.geonames.org/ontology/
[4] *ISO 19111 ontology*. http://loki.cae.drexel.edu/%7Ewbs/ontology/2004/09/iso-19111.owl

enfoque general para describir metadatos geográficos se recoge en (ISO 19115, 2003) y las reglas para la codificación XML son especificadas en (ISO 19139, 2007). El *Work Item* sobre Ontologías del Comité ISO/TC211 (NWI211/N2163 "Geographic information — Ontology") ha desarrollado en este contexto una ontología de la norma (ISO 19115, 2003). No obstante, cuestiones fundamentales como la relevancia y validez del tiempo de un recurso web geoespacial (por ejemplo, una imagen de un mapa), así como la representación de tales propiedades en una URI todavía resultan complicadas de resolver, a pesar de los avances en el área del RDF temporal (Gutierrez *et al.*, 2007).

- **Servicios web geoespaciales** ((*Geospatial*) *Web Services*). La evolución de los estándares, como OWL-S[5], ha elevado el listón en la descripción formal de los servicios web. Sin embargo, las elaboraciones son necesarias de la manera en que el contenido vinculado estrechamente a la mayoría de los servicios web geoespaciales afecta a su modelo de proceso y el comportamiento esperado del interfaz. Algunos trabajos sobre modelado de servicios semánticos geoespaciales se encuentran en (Lieberman, 2007).

3.1.4. Propuestas en el dominio geográfico

En la actualidad parece reconocerse ampliamente que las ontologías tienen significativos beneficios para capturar y compartir la semántica de la información geoespacial (Klien *et al.*, 2005) y servir para el desarrollo de sofisticados SIG (por ejemplo, Fonseca *et al*., 2002) y/o IDE (por ejemplo, Lacasta *et al.*, 2005; Lutz *et al.*, 2009). A continuación, se recogen algunos datos (grupos, proyecto, líneas de investigación y ontologías desarrolladas) que reflejan este reconocimiento.

Los trabajos sobre ontologías geoespaciales se realizan en diversos grupos de investigación (por ejemplo: MUSIL, OntoGeo, OntoSpace, Ordnance Survey, SEEK) y proyectos (por ejemplo: SWEET, meanInGs Geobuddies, SWING, etc.). Esta diversidad refleja la variedad de acercamientos y perspectivas sobre ontologías geoespacial dentro de la comunidad de IG. Desde una perspectiva general, la investigación actual está centrada en el modelado de las ontologías geoespacial y la representación adecuada del tiempo y espacio (Frank, 2003; Arpinar *et al*., 2006; Grenon *et al.,* 2004; Tomai *et al.*, 2004), teorías de vaguedad, incertidumbre y granularidad (Bennett & Cristani, 2004; Bennett *et al.*, 2008), ontologías para descubrimiento y recuperación de IG (Hiramatsu *et al*., 2004; Klien *et al.*, 2004; Lemmens *et al.*, 2004; Lutz *et al.*, 2006; Keßler *et al.*, 2009), ontologías para mediación y transformación (Bowers *et al*., 2004; Fonseca *et al.*, 2002) y *ontology grounding* (Kuhn, 2003, 2005; Scheider *et al.*, 2009).

3.1.4.1. Ontologías geográficas

Seguidamente, en este apartado se recogen, de manera no exhaustiva, distintas ontologías de dominio que han sido creadas en el contexto de la información geográfica.

En (Ressler *et al.*, 2007) se recogen un estudio sobre ontologías relevantes para su utilización en anotación razonamiento cualitativo e interoperabilidad en la comunidad de inteligencia geoespacial. En este trabajo se recogen 45 ontologías espaciales y temporales divididas en 11 categorías, compuestas por las 7 categorías del *W3C Geospatial Incubator Group* (descritas con anterioridad) y 4 categorías adicionales para representar completamente el rango de datos geospaciales y temporales (*Geometric ontology, Coverage ontology, Geopolitical ontology y Temporal ontology*)

Ontology for Geography Markup Language (GML3.0) es la representación en OWL del estándar Geography Markup Language versión 3.0 (GML) del Open Geospatial Consortium. GML es el estándar OGC para el transporte y almacenamiento de información geográfica, incluyendo las propiedades y las geometrías de los fenómenos geográficos. Esta ontología surgió en la Universidad de Drexel en 2004 sobre una versión antigua del GML (la 3.0, cuando

[5] *Semantic Markup for Web Services*. http://www.w3.org/Submission/OWL-S/

la estándar para los WFS es la 3.1). La idea de implementar el GML como ontología es que sirviera como ontología base para el desarrollo de otras ontologías que extendiesen las clases y propiedades de estas clases definidas en ella (perfiles de aplicación de GML) para ser utilizadas en al ámbito de sistemas de información que necesitasen explotar tanto la componente geográfica o espacial como las relaciones semánticas de sus datos.

GeoProfile (Yuan *et al*., 2008) es una ontología para describir servicios Web geoespaciales, en especial servicios basados en la localización (LBS), que utiliza OWL-S para describir la semántica del servicio Web en sí y GML para anotar la parte espacial.

Geospatial Resource Description Framework (GRDF) (Alam, Khan, Thuraisingham, 2008) es una ontología en OWL desarrollada por la Universidad de Texas y construida sobre las clases y propiedades definidos en GML. El fin de esta ontología es proponer un lenguaje para el dominio geoespacial, amplio, expresivo y que haga uso de las bondades de la Web semántica. La ventaja más importante que tiene sobre otros lenguajes geoespaciales es la posibilidad de usar inferencia lógica y agregación dinámica de contenidos.

OntoSensor (Russomanno *et al*., 2005) es una ontología en OWL que referencia y extiende la ontología IEEE Suggested Upper Merged Ontology (SUMO) que define conceptos y asociaciones generales. Se basa en parte en SensorML, el estándar OGC para la temática de sensores y que a su vez está basado en GML.

3.1.4.2. Ontologías hidrográficas

En esta sección se recogen, de manera no exhaustiva, distintas ontologías creadas en el contexto de la información hidrográfica o directamente relacionadas con este dominio.

La ontología ***SWEET*** (*Semantic Web for Earth and Environmental Terminology*) de la NASA (Raskin *et al.*, 2005) es una ontología *upper-level* para la ciencia del planeta Tierra. En su versión 2.1 SWEET es altamente modular con 6000 conceptos en 200 ontologías modulares organizadas por temas. Entre esos temas, aparecen con una visión muy general diversas ontologías relacionadas con el dominio hidrográfico.

En (Piasecki *et al*., 2005) se describe el desarrollo de una ***ontología de dominio para la ingeniería y ciencia hidráulica*** a través de la extensión de las ontologías SWEET. En comparación con SWEET, la ontología desarrollada en este trabajo es más pequeña, pero más detallada dentro del dominio hidrológico. De esta manera, las extensiones conllevan el añadido de nuevos términos, el incremento del nivel de detalle para un elemento (clase, relación, atributo) existente o la introducción de nuevos fenómenos basados en los procesos de las ontologías SWEET. Finalmente, el desarrollo de esta ontología sigue los principios de escalabilidad, ortogonalidad, independencia de la aplicación e independencia del lenguaje natural.

La ***ontología de hidrología*** del ***Ordnance Survey*** (Hart *et al.*, 2007), versión 2.0, tiene como propósito describir los fenómenos hidrológicos tierra adentro recopilados por esta institución. La intención de esta ontología es mejorar el uso de los datos recopilados y permitir un procesamiento semi-automático de estos datos. Los fenómenos que aparecen en esta ontología, mayoritariamente, proceden de (Ordnance Survey, 2001). Esta ontología está compuesta por diferentes módulos (*Spatial relations*, *Network relations*, *Language relations*, *Political geography* y *Mereology*) que incluyen sólo aquellas relaciones y conceptos que son relevantes para la ontología hidrológica.

En (López-Pellicer *et al.*, 2007) se describe una ontología hidrográfica desarrollada conforme a un enfoque *bottom-up*. En su proceso de construcción se aplican técnicas de *Formal Concept*

Analysis (FCA) para obtener la jerarquía de conceptos de las instancias contenidas en los repositorios utilizados como fuentes de datos, en este caso el Nomenclátor NOMGEO.
En (Gómez-Pérez *et al.*, 2008) se describe el proceso automático de construcción de una ontología topográfica, denominada ***PhenomenOntology***. En esta ontología, conforme con el catálogo de fenómenos de la Base Cartográfica Numérica a escala 1:25.000 (BCN25) del Instituto Geográfico Nacional, aparecen un importante número de fenómenos hidrográficos. El objetivo de PhenomenOntology es integrar las bases de conocimiento del referido productor de información.

En (Tripathi *et al.*, 2008) se presenta la extensión de la ontología SWEET (*Semantic Web for Earth and Environmental Terminology*) de la NASA para incluir parte del ***dominio hidrogeológico***. Este trabajo parte de la reutilización de diversas ontologías que forman parte de SWEET (*Earth, property, process, units, material_things, substance, phenomena, human_activities*). Asimismo, este trabajo mantiene el diseño modular de las ontologías que componen SWEET para fomentar y asegurar la reutilización de la ontología hidrogeológica en otras áreas de conocimiento interdisciplinar.

La comparación de estas ontologías (ver Tabla 3.1) demuestra que una gran parte de las ontologías se circunscriben principalmente al dominio de la hidrología[6]. La mayoría de las ontologías descritas en este contexto son ontologías modulares vinculadas con la ontología *upper-level* SWEET. Por otro lado, aunque en menor número, también se describen algunas ontologías del dominio hidrográfico[7] o con información vinculada a este dominio. Asimismo, las métricas de estas ontologías poseen grandes diferencias, consecuencia de la fuente de información de procedencia del conocimiento y de la granularidad asociada. En estas ontologías también destaca la diversidad de metodologías utilizadas y que todas ellas son monolingües (inglés y español).

3.2. Integración de información

En esta sección se exponen los elementos fundamentales del proceso de integración de información. Así, en la sección 3.2.1 se presenta el marco teórico de este proceso y su problemática en la sección 3.2.2. En la sección 3.2.3 se identifican los componentes principales de un sistema de integración. Diferentes enfoques para especificar las correspondencias entre datos se recogen en la sección 3.2.4. Además, se listan diversos acercamientos para el proceso de integración (sección 3.2.5) y diversas propuestas generales de integración (sección 3.2.6). Finalmente, en la sección 3.2.7 se describen distintos tipos de integración de información (sintáctica, estructural y semántica).

3.2.1. Marco teórico

En el Diccionario de la Real Academia de la Lengua destacan dos acepciones del concepto integrar:

- *Completar un todo con las partes que faltaban*.
- *Hacer que alguien o algo pase a formar parte de un todo*.

Una percepción similar de este concepto aparece en (Vckovski, 1998). Estos significados, según este autor, difieren ligeramente, consecuencia de que la principal diferencia se encuentra en la

[6] *Hidrología*. Ciencia geográfica que se dedica al estudio de la distribución, espacial y temporal, y las propiedades del agua presente en la atmósfera y en la corteza terrestre. Esto incluye las precipitaciones, la escorrentía, la humedad del suelo, la evapotranspiración y el equilibrio de las masas glaciares.
[7] *Hidrografía*. La hidrografía es una rama de la Geografía que se ocupa de la descripción y estudio sistemático de los diferentes cuerpos de agua planetarios, en especial, de las aguas continentales.

Tabla 3.1. Características de las distintas ontologías hidrográficas

Ontologías	Métricas	Fuentes de información	Metodología	Granularidad	Formalización	Modulos	Idioma
SWEET	4363 clases 541 propiedades 39 atributos	Ciencia del planeta Tierra (NASA)	SWEET	Global	OWL DL (SHOIN(D))	200 ontologías modulares	@en
hydrology (Piasecki *et al.*, 2005)	21 clases 28 propiedades 5 atributos	Ciencia e ingeniería hidráulica	-	-	OWL DL (ALOIF+(D))	Módulo de SWEET	@en
Ordnance Survey	239 clases 74 propiedades 4 atributos	Ordnance Survey Master Map Real World Objects Catalogue	Ordnance Survey's Conceptual ontology authoring method and Logical ontology authoring method	Nacional	OWL DL (SHOIQ(D))	*Spatial relations*, *Network relations*, *Language relations*, *Political geography* y *Mereology*	@en
Hydro	51 clases	Nomenclátor NOMGEO	FCA	Nacional (1:25.000 - 1:5.000)	OWL Full	-	@es
Phenomen Ontology	21 (117) clases - propiedades 114 atributos	BCN25	Ad hoc wrapper	Nacional (1:25.000)	OWL DL (AL (D))	-	@es
Tripathi *et al.*, 2008	-	SWEET	Diversas ontologías de SWEET	Global	OWL DL	Earth, property, process, units, material_things, substance, phenomena, human_activities	@en

autonomía de las partes. Además, afirma que, en general, la integración sólo tiene sentido si el elemento integrado es más que la suma de sus partes; por tanto, el proceso de integración de información procedente de múltiples bases de datos implica algún tipo de interacción entre las partes o componentes de diferentes bases de datos. Estas partes tienen en la mayoría de los casos una cierta autonomía y la interacción no es posible a priori. El proceso de interacción necesita establecer cierta adherencia de las partes mediante algunos tipos de reglas comunes (Vckovski, 1998).

La adquisición de datos automática o semi-automáticamente, el aumento de su volumen, el número de aplicaciones que los usan, la existencia de múltiples repositorios y la necesidad de intercambiar información entre bases de datos heterogéneas provocan el origen de la integración de información. Conforme a esto, el objetivo (Calvanese *et al.*, 1998) de dicho proceso de integración consiste en acceder y transformar datos almacenados en fuentes de información múltiples y, posiblemente, heterogéneas (cada una de ellas puede utilizar distintos modelos de datos, terminologías e incluso distintos esquemas de representación) y proporcionar una visión integrada y coherente de los datos sin necesidad de grandes cambios en las fuentes originales. Este hecho va a originar que la integración de bases de datos heterogéneas consista en la creación de modelos computacionales que ofrezcan un interfaz uniforme de consulta de datos recopilados y almacenados en múltiples bases de datos.

Esta caracterización del proceso de integración induce la existencia de diferentes tipos o niveles de integración (sintáctica, estructural y semántica), en consonancia con los homónimos niveles de interoperabilidad identificados con anterioridad (sección 2.1.1). Asimismo, estos niveles de interoperabilidad e integración son necesarios para proporcionar transparencia a la naturaleza distribuida de las bases de datos. Dicha transparencia hace que no se requiera que los usuarios y aplicaciones conozcan la existencia, localización física, mecanismo de acceso y/o esquemas de las diferentes bases de datos locales componentes.

Desde una óptica más relacionada con el dominio geográfico, la integración de conjuntos de datos geográficos se define como el proceso de establecimiento de relaciones entre instancias de objetos correspondientes en diferentes conjuntos de datos geográficos de un determinado espacio, producidos autónomamente (Uitermark, 1996). De esta manera, el propósito de la integración de conjuntos de datos geográficos es poder compartir información geográfica entre diversas fuentes caracterizadas por su naturaleza distribuida, gestión independiente, su amplio rango de capacidades de consulta y por tener diferentes tipos de interfaces de usuario (Ram *et al.*, 2000; Uitermark *et al*., 1999a).

3.2.2. El problema de la integración

De forma general, según Ziegler *et al.* (2007), la integración de múltiples sistemas de información refleja la combinación de los sistemas seleccionados, conformando una completa, nueva y unificada vista que da a los usuarios la impresión de interactuar con un único sistema de información. En estos casos, a los usuarios se les proporciona una vista lógica homogénea de los datos que están físicamente distribuidos sobre fuentes de datos heterogéneas. Por esto, todos los datos tienen que ser representados utilizando los mismos principios de abstracción. Esta tarea incluye la detección y resolución de conflictos a nivel de lenguajes, esquemas y datos, cuestiones propias de los procesos de integración sintáctica, estructural y semántica, respectivamente.

En el contexto de las bases de datos, la integración de información ha sido reconocida como un problema central de los sistemas actuales (Ullman, 1997). Mientras en las primeras bases de datos la información necesaria estaba auto-contenida, en la actualidad, se extiende, de forma generalizada, la consideración de que hay un gran valor añadido en la toma de información de varias fuentes y en hacer que estas trabajen juntas como un todo. Este hecho provoca que el proceso de integración de información se vea afectado por múltiples problemas, consecuencia

de que la tarea de integración puede depender de; 1) la arquitectura de un sistema de información, 2) el contenido y la funcionalidad de los sistemas componentes, 3) el tipo de información gestionada por los sistemas componentes, 4) los requerimientos sobre la autonomía de los sistemas componentes, 5) el uso previsto de los sistemas de información integrados, 6) los requerimientos de representación y 7) los recursos disponibles (Dittrich *et al.*, 1999).

El elemento base de cualquier proceso de integración de bases de datos heterogéneas son los datos. Este tipo de datos se caracterizan por presentar diferencias en su representación o interpretación, aunque se refieran a una misma realidad (Litwin *et al*., 1986). Desde la óptica de los datos, la integración de información se ciñe al problema de combinar datos residentes en diferentes fuentes, que pueden proporcionar información similar sobre un recurso y facilitar al usuario una visión unificada de estos datos (Lenzerini, 2002). Las dificultades de este proceso residen en la dificultad de identificar y localizar las fuentes de información más adecuadas y disponibles para una necesidad específica de información. De esta manera, según Devogele *et al.* (1998), los mayores problemas del proceso de integración de varias fuentes de datos en un único marco se centran en;

- Desarrollar un entendimiento correcto de la semántica de los datos existentes, es decir, considerar lo que realmente estos significan.
- Establecer una correcta estructura de correlación para evitar la comparación de elementos no comparables (por ejemplo, "manzanas" con "naranjas").
- Elegir una descripción integrada adecuada basada en los objetivos de integración y en las técnicas de conversión y/o procesamiento de los datos disponibles.

Asimismo, la problemática del proceso de integración hace evidentes una serie de requerimientos de usuario, identificados en (Ram *et al*., 2001):

- Acceso a las fuentes de datos distribuidas a través de un único sistema integrado. Los usuarios serían capaces de realizar consultas de manera transparente a la naturaleza distribuida de las fuentes de datos subyacentes.
- Mecanismo para identificar la relevancia de los datos disponibles. Los metadatos geoespaciales son indispensables en el proceso de localización de los datos y ayudan a prevenir el malgasto de tiempo en la descarga de conjuntos de datos irrelevantes y a reducir las transacciones asociadas y los costes de red.

3.2.3. Marco de integración de información

Los componentes principales de un sistema de integración de información (Lenzerini, 2002) son el esquema global, las fuentes y las correspondencias (*mappings*). Formalmente, este tipo de sistemas I es definido, de manera genérica intentando abarcar los distintos enfoques existentes en la literatura, como una tripleta $\langle G \mid S \mid M \rangle$, donde:

G es el esquema global, es decir, un conjunto de símbolos relaciones globales, cada uno con un aridad (*arity*) asociada (el número de sus atributos), más un conjunto de restricciones de integridad expresadas sobre tales símbolos relacionales.

S representa el esquema de la fuente, es decir, un conjunto de símbolos relacionales (disjuntos de G), que constituye una representación relacional de los datos almacenados en las fuentes.

M es la correspondencia (*mapping*) entre G y S, constituida por un conjunto de afirmaciones de la forma

$$q_S \rightarrow q_G,$$
$$q_G \rightarrow q_S$$

donde q_S y q_G son dos consultas de la misma aridad, respectivamente sobre el esquema fuente S, y sobre el esquema global G. Intuitivamente una afirmación $q_S \rightarrow q_G$ especifica que el concepto representado por la consulta q_S sobre las fuentes corresponde al concepto en el esquema global representado por la consulta q_G. De forma similar ocurre en una afirmación del tipo $q_G \rightarrow q_S$.

El esquema de la fuente describe la estructura de las fuentes, donde están los datos. Mientras el esquema global proporciona una vista virtual, integrada y conciliada de las fuentes subyacentes. Mediante las correspondencias se establece la conexión entre los elementos del esquema global y aquellos que pertenecen al esquema de la fuente. En la Figura 3.3, se muestra una representación de la situación generada en el proceso de integración.

3.2.4. Modelado

Uno de los aspectos más importantes en el diseño de un sistema de integración de datos es la especificación de las correspondencias entre los datos en la fuente y estos en el esquema global (ver Figura 3.3). Estas correspondencias determinan cómo las consultas realizadas al sistema son respondidas (Levy, 2000; Lenzerini, 2002).

Desde el campo de la integración de información existen diferentes enfoques para especificar las correspondencias entre los datos en un sistema. Estos acercamientos son Local-as-View, Global-as-View y Global-and-Local-As-View.

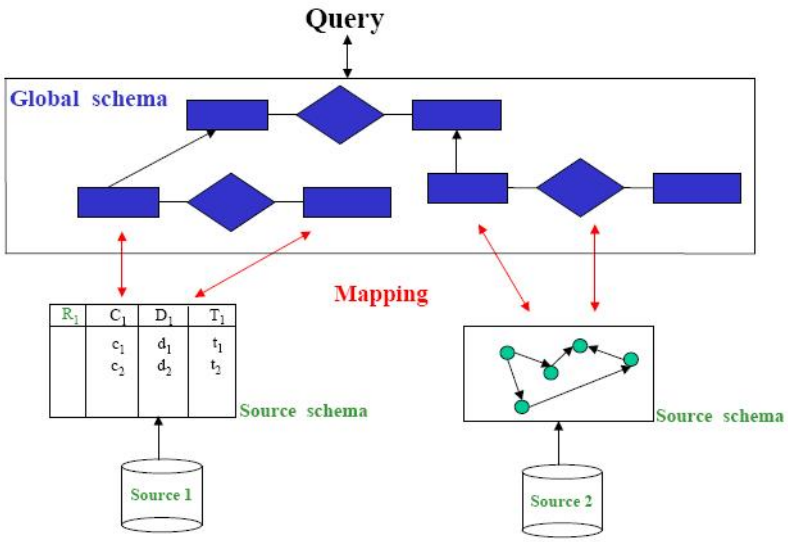


Figura 3.3. Integración de datos (Lenzerini, 2002)

Local as View (LAV)

El enfoque Local-as-View (LAV) (Lenzerini, 2002) ha sido utilizado en diversos trabajos (Levy *et al.*, 1996; Kwok *et al.*, 1996; Duschka *et al.*, 1997a; Duschka *et al.*, 1997b; Friedman *et al.*, 1997; Ives *et al.*, 1999; Lattes *et al.*, 2000; Levy, 2000; Halevy *et al.*, 2006). Este enfoque necesita que el esquema global sea especificado independientemente de la fuente y las relaciones entre el esquema global y las fuentes son establecidas por definición de cada fuente como una vista sobre el esquema global.

En este enfoque las correspondencias de *M* asocian a cada elemento s ∈ *S* una pregunta q_G sobre *G*. En otras palabras, el lenguaje de preguntas $L_{M,S}$ solamente permite expresiones constituidas por un único símbolo del alfabeto A_S. Por lo tanto una correspondencia LAV es un conjunto de aserciones de la forma

$$s \rightarrow q_G$$

una para cada elemento de S.
Desde el punto de vista del modelado, el enfoque LAV está basado en la idea de que el contenido de cada fuente *s* debería ser caracterizado en términos de una vista q_G sobre el esquema global.

Por otro lado, este enfoque favorece la extensibilidad de los sistemas, ya que añadir una nueva fuente simplemente significa enriquecer la correspondencia con una nueva afirmación, sin más cambios.

Global as View (GAV)

El enfoque Global-as-View (GAV) (Lenzerini, 2002) ha sido utilizado en diversas propuestas (Chawathe *et al.*, 1994; Papakonstantinou *et al.*, 1996a; Adali *et al.*, 1996; Florescu *et al.*, 1996; Garcia-Molina *et al.*, 1997; Ullman, 1997; Levy, 2000; Halevy *et al.*, 2006). En este enfoque las correspondencias de *M* asocian a cada elemento $g \in G$ una pregunta q_S sobre *S*. En otras palabras, el lenguaje de preguntas $L_{M,G}$ solamente permite expresiones constituidas por un único símbolo del alfabeto A_G. Por lo tanto una correspondencia GAV es un conjunto de aserciones de la forma

$$g \rightarrow q_S$$

una para cada elemento de G.

Desde el punto de vista del modelado, el enfoque GAV está basado en la idea de que el contenido de cada elemento *g* del esquema global debería ser caracterizado en términos de una vista q_S sobre las fuentes. En algunos casos las correspondencias explícitamente dicen al sistema cómo recuperar los datos cuando un sistema quiere evaluar los diferentes elementos del esquema global. Esta idea es efectiva siempre que los sistemas de integración de datos estén basados en un conjunto de fuentes estables.

Asimismo, el enfoque GAV favorece que al sistema en el procesamiento de consulta llevado a cabo, porque dice al sistema cómo utilizar las fuentes para recuperar datos. Sin embargo, la extensión del sistema con una nueva fuente resulta un problema, consecuencia de que la nueva fuente puede tener impacto en la definición de los diferentes elementos del esquema global, cuyas vistas asociadas necesitan ser redefinidas.

Global and Local as View

Con el objeto de resolver las lagunas generadas por GAV y LAV surge el enfoque Global-and-Local-as-View (GLAV) (Friedman *et al.*, 1999; Calvanese *et al.*, 2001). Este enfoque combina las ventajas de GAV y LAV. Asimismo, las correspondencias en este enfoque son más expresivas y son adecuadas para representar relaciones complejas en los entornos de integración de datos distribuidos. El mencionado poder de expresividad de las correspondencias de *M* asocian a una pregunta q_G otra pregunta q_S.

Comparativa de los diferentes enfoques

Según Levy (2000); Lenzerini (2002); Hernandez *et al.* (2004), la ventaja de GAV es que la descomposición de consultas es fácil, consecuencia de las relaciones en el esquema integrado están ya definidas en términos de las relaciones en el esquema de fuente. Sin embargo, la desventaja es que añadir y borrar fuentes de datos al mediador es complicado, porque el esquema integrado debe ser actualizado para reflejar la presencia o ausencia de las fuentes de datos.

Por el contrario, en el enfoque LAV cada fuente es descrita de forma individual. Las tareas de este sistema resuelven, en tiempo de consulta, cómo las fuentes interactúan y cómo sus datos pueden ser combinados para responder una consulta. Asimismo, el proceso de añadido y/o borrado de las fuentes de datos es fácil. El inconveniente, sin embargo, es que la reformulación es más complicada y, en algunas ocasiones, requiere consultas reiterativas sobre las fuentes. Además, la descomposición de consultas es mucho más complicada en este enfoque.

Finalmente, en (Friedman *et al.*, 1999) es descrito el lenguaje GLAV que combina el poder de expresividad de GAV y LAV y la complejidad de reformulación de consultas es la misma que para LAV.

3.2.5. Nuevas arquitecturas en el contexto de la integración de información

Recientes desarrollos de infraestructuras y algoritmos para sistemas distribuidos junto al incremento del número de fuentes de información heterogéneas aisladas han derivado en una insuficiencia de la arquitectura centralizada introducida para satisfacer las peticiones de integración. Además, en un escenario cambiante continuamente en el que las fuentes pueden ser fusionadas u olvidadas en cualquier momento hace que se presenten otros escenarios de integración de información. De esta manera, en (Ruzzi, 2004) se identifican diferentes acercamientos para el proceso de integración, tales como; **sistemas de gestión de datos *peer-to-peer*** (P2P) (Bernstein *et al.*, 2002; Halevy *et al.*, 2003; Halevy *et al.*, 2006), **integración de datos peer-to-peer** (Calvanese *et al*., 2004; Ziegler *et al.*, 2004) y **otros acercamientos a la integración de información** encaminados a resolver problemas como *what-to-ask* (Calvanese *et al.*, 2004) o los referidos al intercambio de datos (Fagin *et al*., 2003; Fagin *et al*., 2005).

3.2.6. Propuestas generales de integración de información

En (Stuckenschmidt *et al*., 1999) se proponen tres formas diferentes en las que las fuentes de información pueden ser integradas. Estas opciones propuestas son reflejadas, a continuación, en la Figura 3.4.

En la primera forma, para acceder a información heterogénea, normalmente, se necesita una traducción de la representación de una fuente a la representación de otra fuente resultante, caracterizada por un modelo común. Este acercamiento resulta dependiente de la expresividad de las representaciones utilizadas. Las características de este acercamiento lo relacionan estrechamente con el proceso de integración estructural, descrito en detalle en la sección 3.2.7.2. La segunda propuesta de integración, también conocida como meta-información, es utilizada para describir fuentes de información mediante el uso de un metalenguaje. Esto hace que esta propuesta de integración esté vinculada con la integración sintáctica de las fuentes de información, descrita en la sección 3.2.7.1.

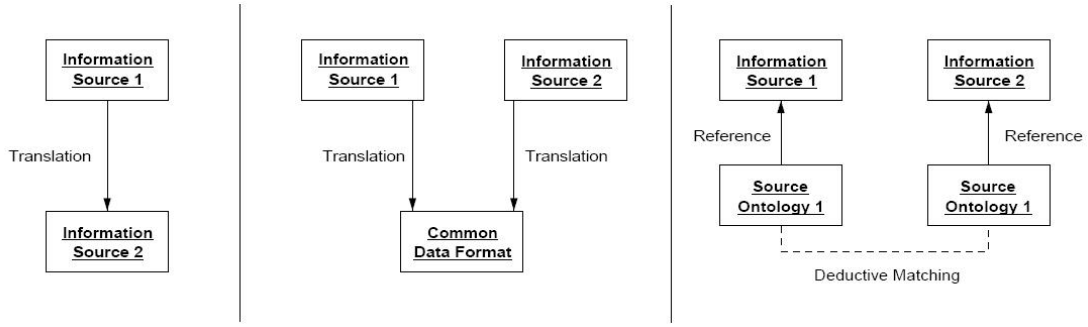


Figura 3.4. Propuestas generales para integrar información (Stuckenschmidt *et al*., 1999)

Finalmente, el tercer enfoque utiliza ontologías como instrumento para conseguir la definición de un vocabulario común y lograr la integración de la información. Las ontologías proporcionan una descripción de las fuentes de información aportando claridad sobre el tipo de información

disponible. Asimismo, éstas permiten el acceso a la información vía conceptos abstractos, incluyendo los conceptos más específicos utilizados desde diferentes fuentes de información. Por tanto, esta propuesta está ligada con el proceso de integración semántica ilustrada en la sección 3.2.7.3.
A estas propuestas generales de integración de información hay que añadirles los diferentes niveles del proceso de integración. Así, según Calvanese *et al*. (1998); Yu *et al*. (2008), se pueden encontrar los siguientes niveles:

- *Nivel conceptual*. Este nivel contiene la información básica del proceso de integración, es decir, los datos. Asimismo, se puede encontrar una descripción formal de los datos, las relaciones y los requerimientos de información con los que la aplicación de integración tiene que tratar.
- *Nivel lógico*. Contiene la descripción de los datos y las consultas de interés al sistema, expresadas en términos de estructuras lógicas típicas gestionadas por los SGBD. En particular, el esquema de una fuente *S* describe la lógica contenida en el esquema de la vista materializada mantenido por el sistema. Colectivamente, el esquema de la fuente y el esquema de la vista materializada conforman lo que se denomina esquema de datos. Obviamente, el esquema de la vista materializada es significativo únicamente en los casos donde los datos integrados son materializados. Por otro lado, resulta menos significativo en los casos de integración virtual completa. Finalmente, el esquema de consulta expresa la información necesaria en el nivel lógico.
- *Nivel físico*. Este nivel se refiere a los datos gestionados por el sistema. En este nivel, habitualmente, se van a materializar *wrappers* y mediadores.
- *Nivel de metadatos*. Aquí se encuentran los repositorios con la información de los metadatos sobre los diferentes sistemas componentes.
- *Nivel de representación*. La necesidad de integrar componentes a nivel de presentación o interfaz resulta evidente de cara al usuario final, por este motivo, son numerosos los ejemplos existentes que en la actualidad se concretan en *mashup*.

3.2.7. Tipos de procesos de integración de información

Diferentes tipos de integración son distinguidas en (Visser, 2005; Visser *et al.*, 2002; Stuckenschmidt *et al*., 2000; Ziegler *et al.*, 2004; Calvanese *et al.*, 1998) con objeto de alcanzar un completo acceso integrado a la información. Conforme a esto, los tipos de integración a tratar, acordes a los objetivos de este trabajo, son; integración sintáctica, estructural y semántica.

Las características de estos procesos de integración (sintáctica, estructural y semántica) están marcadas por los consiguientes tipos de interoperabilidad citados anteriormente. Así, según Jang *et al.* (2007), la integración sintáctica de la información es posible gracias a que los problemas de interoperabilidad de la primera generación de sistemas han sido resueltos por los avances de las tecnologías hardware y software, consiguiendo así la interoperabilidad sintáctica ente sistemas. La integración de esquema, estudiada intensamente desde mediados de 1980, influye en la mayoría de las tecnologías actuales que pretenden tratar con bases de datos heterogéneas y distribuidas de una manera variante. Respecto a la integración semántica, cada vez más, aparecen propuestas basadas en ontologías, consecuencia de que éstas se centran en los problemas de interoperabilidad semántica. A continuación se procederá a la descripción detallada de los diferentes tipos de integración y sus múltiples enfoques de integración propuestos.

3.2.7.1. Integración sintáctica o perspectiva clásica

Tradicionalmente, los sistemas que gestionan la información geoespacial, especialmente los Sistemas de Información Geográfica (SIG), han considerado el proceso de integración de información como una mera superposición, a diferentes niveles, de capas de información. Esta situación se deriva como consecuencia de que los SIG pueden visualizar varios conjuntos de datos dentro de un mapa en base a sus referencias espaciales. Uno de los condicionantes de la

utilización de estos sistemas como "*simples*" almacenes de datos con representación espacial (de escasa utilización como auténticos instrumentos de planificación territorial, toma de decisiones y/o gestión de recursos) (Vilches Blázquez *et al*., 2008) ha sido la utilización de formatos y estándares propietarios. Esto provocó, según Bishr (1998), entre finales de los años 70 y principios de los 80, que las diferentes aplicaciones SIG fueran consideradas islas de información, dado que estaban auto-contenidas en sistemas independientes, donde los datos espaciales eran capturados, almacenados, analizados y visualizados digitalmente. Asimismo, los datos, raramente, eran adquiridos de fuentes digitales dado lo cerrado de los formatos de los ficheros de las diferentes aplicaciones.

La dificultad de integrar información en este contexto se recoge en (Lin *et al*., 2004). La información de carácter temático es recogida en diferentes ficheros de un determinado formato, considerado como un estándar de facto (*shapefile*)[1]. Por otro lado, la información asociada ("eras geológicas" y "tipos de roca") es almacenada en tablas con diferentes esquemas y vocabularios. A este nivel, como puede apreciarse en la Figura 3.5, los SIG permiten la integración de información con formatos similares, es decir, puede visualizarse simultáneamente diferentes capas de información sobre un mapa. Por el contrario, no resulta fácil integrar información de formatos distintos y consultar la información geoespacial a través de un interfaz unificado y mostrar los resultados de las consultas sobre un mapa.

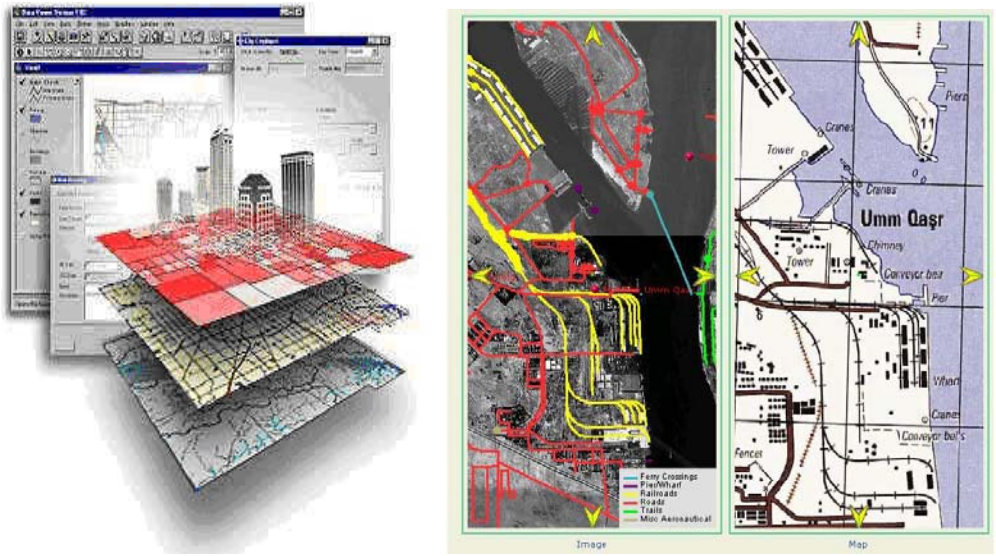


Figura 3.5. Superposición de información en herramientas SIG

A mediados de los 90, el primer paso en la tarea de integración se derivó de los avances en las tecnologías de bases de datos, que permitían alojar datos espaciales en bases de datos relacionales, permitiendo a las organizaciones dar los primeros pasos hacia el SIG corporativo y la eliminación de las islas de datos (ESRI, 2003). Algunos ejemplos son la aparición de Oracle Spatial & Locator, PostgreSQL con la extensión PostGIS y MySQL. A pesar de este primer paso, los SIG existentes en el mercado están caracterizados por formatos propietarios que hace muy complicado conseguir alcanzar interoperabilidad entre sistemas. Asimismo, a finales de los años 90 comienzan a surgir numerosos estándares para el intercambio de datos geográficos entre sistemas (Albrecht, 1999; Shekthar *et al.*, 1999).

[1] *Shapefile*. Formato de archivo propietario, desarrollado por la compañía ESRI, que almacena información vectorial. Este formato se caracteriza por guardar la localización de los elementos geográficos y los atributos asociados a ellos, aunque carece de capacidad para almacenar información topológica. Actualmente este formato se ha convertido en formato estándar de facto para el intercambio de información geográfica entre diferentes SIG (ESRI, 1998).

Los primeros trabajos en el contexto de la integración de datos espaciales en sistemas de bases de datos estuvieron centrados en compartir simples fenómenos espaciales en una base de datos relacional. Los lenguajes de manipulación de datos estándar, como SQL (*Structured Query Language*), comenzaron a adoptar funcionalidades espaciales comunes para albergar operaciones como selecciones espaciales y/o consultas topológicas en sentencias SQL. La llegada de los primeros modelos relacionales capaces de almacenar elementos espaciales e información alfanumérica guió a las bases de datos espaciales (Rigaux *et al.*, 2001) a proporcionar métodos de acceso, algoritmos, modelado de los datos espaciales y extensiones del procesamiento de consulta de los sistemas tradicionales de bases de datos (Díaz *et al*., 2008).

Con el rápido desarrollo de los sistemas de información y el paradigma de las bases de datos distribuidas surge la necesidad de los SIG interoperables. El objetivo de dichos SIG es conseguir intercambio de datos de manera transparente y acceso remoto a los servicios que estas herramientas proporcionan (Bishr, 1998). Asimismo, la puesta en escena de *Open GIS Consortium* (OGC) transformó la arquitectura de los SIG, identificando en el interfaz de los componentes software el ingrediente clave de la tecnología SIG que necesitaba estandarización para conseguir interoperabilidad (Buehler, 2003). El resultado fue una serie de especificaciones para desarrollar interfaces sintácticos que establecen protocolos para intercambiar componentes de información geoespacial, pero sin especificar el significado de sus términos de una forma legible por las máquinas (Kuhn, 2005).

El éxito de la tecnología de los servicios web permite promover la integración de servicios y la interoperabilidad a nivel sintáctico entre fuentes de información distribuidas. El acercamiento de los SIG a la arquitectura orientada a servicios (*Service-Oriented Architecture* - SOA) es representado por el paradigma de las IDE, cuya integración de datos implica una arquitectura similar a la propuesta en la Figura 3.6. Las IDE van a ofrecer la posibilidad de acceder a datos espaciales heterogéneos y distribuidos a través de un conjunto de políticas, reglas comunes y estándares que facilitan la interconexión de información espacial de una manera interoperable (Granell *et al*., 2007).

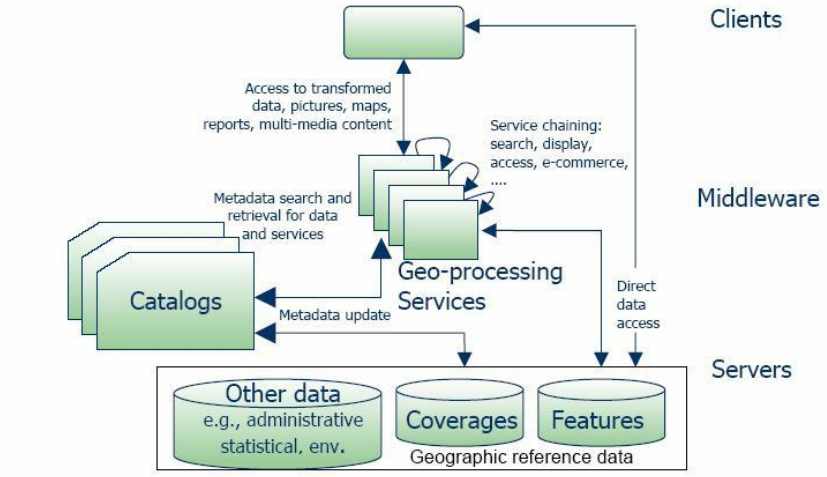


Figura 3.6. Modelo de referencia de la arquitectura de una IDE (Smits, 2002)

El conjunto de actividades políticas y estándares que comprenden las IDE promueven la creación de una infraestructura de información geoespacial para ayudar a las comunidades de usuarios a reunir, compartir, acceder y explotar los recursos de información georreferenciados (Díaz *et al*., 2008). Como puede apreciarse en la Figura 3.6, el enfoque tradicional de las IDE centrado en los datos está cambiando a una visión basada en servicios (Bernard *et al*., 2005), donde los servicios geoespaciales ubicados en el middleware son, cada vez más, utilizados para descubrir y acceder a los datos almacenados en diferentes repositorios. De esta manera, se transformarán los datos en información útil y estará a disposición de los usuarios (Díaz *et al*., 2008).

Mecanismos de integración a nivel sintáctico

Seguidamente, se van a mostrar diferentes propuestas, encaminadas a solventar la heterogeneidad en el proceso de formalización, utilizadas en el contexto de la integración sintáctica, caso del lenguaje XML y de los *gateways*. Asimismo, se expondrá algunas instanciaciones de este tipo de integración para el contexto geográfico a través de las especificaciones OGC y de los estándares ISO/TC 211. Finalmente, de forma breve, se comentará cómo se adaptan los SIG tradicionales al nuevo contexto promovido por la interoperabilidad sintáctica.

***eXtensible Markup Language* (XML)**

XML surge bajo el auspicio de *World Wide Web Consortium* como una evolución de *Standard Generalized Markup Language* (SGML)[2]. El objetivo inicial de este lenguaje era permitir SGML genérico "*para ser servido, recibido y preprocesado en la web de la misma manera que es posible con HTML* (*HyperText Markup Language*) *(también basado en SGML). XML había sido diseñado para facilitar la implementación y para interoperar con SGML y HTML*".

XML es un metalenguaje que puede ser utilizado para definir lenguajes de marcado específicos, como por ejemplo; XHTML[3], MathML[4], SVG[5] o GML en un contexto más relacionado con la información geoespacial. De esta manera, XML permite a los usuarios definir nuevas etiquetas y estructuras para sus propios lenguajes.

Desde la óptica de la integración, este lenguaje, según Halevy *et al.* (2006), impulsó el proceso de integración de datos debido a que este lenguaje ofrecía un formato común para compartir datos entre fuentes de datos. Sin embargo, XML no consideró las cuestiones relacionadas con la integración semántica de la información. Así, las fuentes pueden compartir ficheros XML cuyas etiquetas son completamente insignificantes fuera de una determinada aplicación.

Desde el punto de vista técnico, son múltiples los sistemas de integración desarrollados utilizando XML como modelo de datos (Baru *et al.*, 1999; Ives *et al.*, 2003; Ives *et al.*, 1999; Manolescu *et al.*, 2001; Naughton *et al.*, 2001; Yu *et al.*, 2004; Ives *et al.*, 2001; Papakonstantinou *et al.*, 2003) y los lenguajes de consulta (originalmente, XML-QL (Deutsch *et al.*, 1999) y más tarde XQuery (Chamberlin *et al.*, 2001)).

Geographic Markup Language

GML es la instanciación de XML en el dominio geoespacial (Gaun *et al*., 2003) conforme con ISO 19118 (ISO, 2005c) que define el almacenamiento, transporte, intercambio, procesamiento y transformación de la IG. Este lenguaje está basado en el modelo geográfico definido por OGC; modelo que describe el mundo en tipos de fenómenos. De esta manera, GML proporciona una variedad de tipos de objetos para describir fenómenos geográficos, sistemas de referencia, geometría, topología, tiempo, unidades de medida y valores generalizados.

La especificación GML define la sintaxis del esquema XML, los mecanismos y las convenciones que:

- proporcionan un marco abierto para la definición de los objetos y esquemas de aplicaciones geoespaciales;

[2] *SGML*. http://www.w3.org/MarkUp/SGML/
[3] *XHTML* corresponde a *eXtensible HyperText Markup Language*. http://www.w3.org/TR/xhtml1/
[4] *MathML* son las siglas de *Mathematical Markup Language*. http://www.w3.org/Math/
[5] *Scalable Vector Graphics* corresponde a las iniciales SVG. http://www.w3.org/Graphics/SVG/

- soporta la descripción de los esquemas de aplicaciones geoespaciales para las comunidades de información y los dominios especializados;
- permite la creación y mantenimiento de esquemas de aplicaciones geográficas conectadas y conjuntos de datos;
- soporta el almacenamiento y transporte de esquemas de aplicación y conjuntos de datos;
- aumenta la capacidad de las organizaciones de compartir esquemas de aplicaciones geográficas y la información que ellos describen.

La utilización de GML en la comunidad geográfica está ampliamente extendida. Así lo demuestran la diversidad de trabajos en diversas áreas, como por ejemplo; interoperabilidad y compartición de datos (Rawat *et al.*, 2004; Quak *et al.*, 2002; Murray, 2004; Peng, 2005; Murray *et al.*, 2004), modelos de datos (Valerhaugen, 2004; Steudler, 2006) o servicios web (Chang *et al.*, 2006; Zhang *et al.*, 2001; Vatsavai *et al.*, 2006; Kim *et al.*, 2005; Luo *et al.*, 2004).

Servicios OGC y Estándares ISO/TC 211

Hasta el momento los esfuerzos en los sistemas que soportan los SIG e IDE se han centrado en la implementación de estándares que posibilitan la comunicación entre máquinas, resolviendo los problemas de *sintaxis*. Para ello, se crean los estándares del Comité Técnico 211 del *International Organization for Standardization* (ISO/TC 211), *World Wide Web Consortium* (W3C) y del *Open Geospatial Consortium* (OGC) con los que se resuelven problemas derivados de la diferencia de formatos de información, consiguiendo interoperabilidad a nivel sintáctico.

Las especificaciones y estándares definidos por OGC e ISO/TC 211 se han centrado en la definición de modelos de información para los objetos geográficos basados en fenómenos (instancias de elementos conceptuales en el mundo real o representaciones digitales de los fenómenos) y coberturas (la propiedad que un valor hace disponible para cualquier localización en un dominio espacio-temporal definido), así como diversos aspectos secundarios tales como; representación de información espacial (geometrías), referenciación espacial mediante coordenadas, aspectos temporales y metadatos. De este contexto se extraen los servicios (*Web Map Service, Web Feature Service, Web Coverage Service*) más extendidos en el entorno de las IDE. Estos servicios, descritos a continuación de forma breve, manifiestan de manera evidente que el proceso de integración de información a nivel sintáctico continúa limitándose a una simple superposición de diferentes capas de información.

Web Map Service

La espeficación *Web Map Service* (WMS) (Open GIS Consortium (OGC), 2004; ISO, 2005b) de OGC define un "mapa" para ser una representación de la IG como una imagen digital disponible para su visualización en pantalla. De esta manera, los servicios de mapas proporcionan un simple interfaz web basado en un modelo que soporta reglas de peticiones y respuestas utilizando *Hypertext Transfer Protocol* (HTTP) y *eXtensible Markup Language* (XML) para petición de imágenes de mapas georreferenciadas de una o más bases de datos geoespaciales distribuidas. Una petición WMS define las capas geográficas y el área de interés a procesar. La respuesta a la petición es una o más imágenes de mapas georreferenciadas (en formatos JPEG, PNG u, ocasionalmente, como elementos basados en vectores en *Scalable Vector Graphics* (SVG) o formatos *Web Computer Graphics Metafile* (WebCGM)) que pueden ser visualizadas en un navegador. El interfaz también soporta la capacidad de especificar transparencia a las imágenes para permitir superposición de múltiples capas de información provenientes de múltiples servidores.

Web Feature Service

Web Feature Service (WFS) (Open Geospatial Consortium, 2005; ISO, 2007a) es una especificación de implementación que proporciona interfaces para soportar la transacción, acceso y manipulación de fenómenos geográficos sobre la web. Este tipo de servicios devuelven los verdaderos datos espaciales almacenados en bases de datos geográficas, a diferencia de los

servicios WMS que sólo devuelven una imagen que no puede ser editada y/o analizada espacialmente.

WFS propone interfaces para describir operaciones de manipulación y acceso a los datos sobre fenómenos geográficos utilizando HTTP como protocolo de comunicación. Las operaciones de manipulación de datos incluyen procesos de creación, borrado y actualización de nuevas instancias de fenómenos. Además, estos servicios posibilitan la consulta de fenómenos con información asociada tanto espacial como no espacial.

Para la codificación del proceso de abstracción de los fenómenos geográficos se utiliza *Geography Markup Language* (GML). Asimismo, este tipo de servicios permite recuperar de múltiples fuentes de información datos geoespaciales codificados en GML.

En lo que respecta a la transacción de información, una petición de WFS consta de una descripción de consulta u operaciones de transformación de datos que son aplicadas a uno o más fenómenos. La petición es generada en el cliente y es enviada al servidor WFS utilizando HTTP.

Web Coverage Service

La espeficación *Web Coverage Service* (WCS) (Open Geospatial Consortium, 2008) define un interfaz y operaciones estándar que posibilitan acceso e intercambio interoperable de coberturas[6] geoespaciales. El término "*coberturas de malla*", normalmente, contiene imágenes de satélite, fotografías aéreas digitales, datos de elevación digital y otros fenómenos representados por valores en cada punto de medida.

A diferencia de los servicios WMS, que filtran y representan los datos espaciales para devolver un mapa estático, los servicios WCS proporcionan los datos disponibles junto con su descripciones detalladas, permitiendo consultas complejas contra estos datos y devuelve datos con su semántica original (en lugar de una imagen), que puede ser interpretada, extrapolada, etc. y no sólo representada.

Respecto a los servicios WFS, que devuelven fenómenos geoespaciales discretos, los WCS retornan representaciones de fenómenos cambiantes en el espacio relativo a un dominio espacio-temporal (posiblemente multidimensional) a un rango de propiedades.

Una profusa recopilación con diferentes niveles de detalle (mundial, nacional, regional y local) de diferentes tipos de servicios (WMS[7], WFS[8], WCS[9], etc.) puede ser comprobada en el directorio de servicios IDEE (Infraestructura de Datos Espaciales de España).

Los SIG tradicionales en el contexto de la interoperabilidad sintáctica

Los SIG tradicionales de escritorio, tanto en forma de productos comerciales como los referidos a software libre, se adaptan a este contexto de integración sintáctica de la información. A continuación, sin pretender ser exhaustivos, se mencionan algunos de los SIG más relevantes en el ámbito de los productos comerciales.

[6] *A feature that associates positions within a bounded space (its spatiotemporal domain) to feature attribute values (its range). GIS coverages (including the special case of Earth images) are two- (and sometimes higher-) dimensional metaphors for phenomena found on or near a portion of the Earth's surface. A coverage can consist of a set of features or Feature Collections. Earth images are seen as Grid Coverages that contain features whose geometries are of type "set of cells" or "set of pixels" (surfaces)* (Open Geospatial Consortium, 2006).

[7] http://www.idee.es/CatalogoServicios/CatServ/directorio_servicios.html

[8] http://www.idee.es/CatalogoServicios/CatServ/indexWFS_new.html

[9] http://www.idee.es/CatalogoServicios/CatServ/indexWCS_new.html

- ESRI por medio de su producto de referencia *ArcGIS* permite la publicación de datos en formato KML[10] para compartir e integrar servicios SIG con *Google Earth* y *Google Maps* y soporta la integración de servicios de las especificaciones OGC tales como; WMS, WFS, WCS, etc. para compartir la información con cualquier aplicación.
- Geomedia permite publicar servicios OGC de tipo WMS y WFS y admite integrar información geoespacial siguiendo especificaciones OGC en el entorno *Google Earth*, vía KML.
- Miramon[11] actualmente es conforme a los estándares WMS y WCS.

Desde la óptica del software libre se muestran, sin pretender ser exhaustivos, algunos de los SIG más relevantes a continuación:

- OpenJUMP[12] utiliza estándares como GML, WMS y WFS.
- gvSIG[13] integra en una vista datos tanto locales como remotos a través de un origen WMS, WCS o WFS.
- uDig[14] (*User Friendly Desktop Gis*) en su parte orientada a internet, consume servicios web geoespaciales estándares (WMS, WFS, WCS) y estándares de facto (GeoRSS, KML, tiles).
- Quantum GIS[15] (QGIS) soporta los estándares WMS y WFS de OGC.

3.2.7.2. Integración estructural

Desde mediados de los años 80 en la gestión de bases de datos y otras áreas relacionadas ha habido un creciente interés en la integración de esquemas de múltiples bases de datos autónomas y heterogéneas. La tarea de integración estructural (Stuckenschmidt *et al.*, 2000a; Visser, 2005) o de esquemas (Batini *et al*., 1986; Sheth *et al.*, 1990; Bright *et al*., 1994; Castano *et al.*, 1997; Palopoli *et al*., 2000; Jang *et al.*, 2007) consiste en reformatear las estructuras de los datos a una nueva estructura homogénea. Esto puede ser hecho con la ayuda de un formalismo capaz de construir una fuente de información específica de numerosas fuentes de información. Esta es una clásica tarea de un middleware.

Desde un punto de vista general, las dificultades para llevar a cabo este complejo proceso de integración son comentadas en (Sheth *et al.*, 1990). Entre estos problemas destacan; la incapacidad de los modelos de datos de capturar completamente un estado del mundo real, la necesidad de capturar mucha más información de la que normalmente es tomada en un esquema y, finalmente, las dificultades de sintetizar múltiples visiones del mundo e interpretaciones de los estados del mundo real, considerando, además, que estas pueden cambiar con el paso del tiempo. Esta serie de problemas junto a las características propias del proceso de integración estructural vinculan a este tipo de integración con la heterogeneidad estructural y semántica, descritas en la sección 2.2 de este trabajo.

En el marco de la integración estructural de bases de datos heterogéneas existen dos métodos de integración de información, tales como; integración débil (*lazy integration*) e integración consistente (*eager integration*) (Widom, 1996; Chou, 2005).

[10] *KML* del acrónimo en inglés *Keyhole Markup Language* (lenguaje de marcas de Keyhole), es una gramática XML y un formato de archivo para la creación de modelos y el almacenamiento de funciones geográficas como puntos, líneas, imágenes, polígonos y modelos que se muestran en *Google Earth*, *Google Maps* y otras aplicaciones. Fuente: http://earth.google.es/userguide/v4/ug_kml.html
[11] *Miramon*. http://www.creaf.uab.es/MiraMon/index_es.htm
[12] *OpenJUMP*. http://openjump.org/wiki/show/HomePage
[13] *gvSIG*. http://www.gvsig.gva.es/index.php?id=gvsig&L=0
[14] *uDIG*. http://udig.refractions.net/
[15] *Quantum GIS*. http://www.qgis.org/

- En el proceso de integración débil la información es extraída de las fuentes de datos únicamente cuando al sistema se le presenta una consulta, es decir, la integración se realiza mediante virtualización.
- El proceso de integración consistente la información es extraída de las fuentes de datos y ubicada en un repositorio central por adelantado. En esta integración se crea una vista materializada de los datos.

Existen diferentes acercamientos y/o técnicas dentro del proceso de integración estructural. Así, (Elmagarmid *et al.*, 1999) presenta una taxonomía de soluciones en materia de integración de bases de datos, clasificadas en; esquema global de integración (*data warehouses*), lenguaje de aproximación a múltiples bases de datos (*multidatabases*) y sistemas federados de bases de datos. A estos acercamientos, con el objetivo de completar, de forma exhaustiva, el espectro de propuestas empleadas en el proceso de integración estructural, se le unen los mediadores y wrappers, diversos lenguajes orientados al establecimiento de correspondencias y al proceso de migración y/o *upgrade* y, por último, algunas otras propuestas que por su miscelánea resultan complejas de clasificar. Asimismo, se presentan algunas instanciaciones de las anteriores propuestas para el dominio geográfico.

Propuestas de integración a nivel estructural

Los primeros acercamientos de integración de información centrados en bases de datos, según Mena *et al.* (2000), comprendían las bases de datos múltiples y los sistemas federados. Estos acercamientos tratan de manera consistente la estructuración y manipulación de la información. La integración lógica de los esquemas, llevada a cabo mediante las mencionadas propuestas, conlleva una descripción de los datos subyacentes, utilizada para manejar la heterogeneidad estructural y representacional.

Múltiples bases de datos

Los sistemas de múltiples bases de datos (Chou, 2005; Hurson *et al.*, 1991) proporcionan acceso integrado a múltiples sistemas gestores de bases de datos relacionales (SGDB). Estos sistemas logran cierto grado de transparencia en el acceso a los datos por ocultar que los SGBD componentes son distribuidos. Sin embargo, los sistemas no ocultan al usuario los esquemas de las bases de datos de los SGBD componentes. Estos esquemas exponen al usuario la unión de los esquemas de las bases de datos formantes, donde el nombre de las tablas son renombradas con el nombre de las bases de datos añadidas como prefijos. Por otro lado, estos sistemas de múltiples bases de datos son la implementación más simple de los sistemas débiles.

Los ejemplos de instanciación de este procedimiento de integración estructural no resultan muy numerosos en la bibliografía científica. A pesar de esto, se encuentran casos como MULTIBASE (Landers *et al.*, 1982), The CORDS multidatabase Project (Attaluri *et al.*, 1995) o ADAM (*DAta in Multi-database systems*)[16].

La ausencia de transparencia es la principal desventaja para el uso de la propuesta de estos sistemas de bases de datos múltiples, ya que requieren al usuario fluidez en lenguajes de consulta de bases de datos, caso de SQL, y entendimiento de la distribución de información entre las bases de datos integradas en el sistema para poder plantear consultas. Asimismo, la ausencia de metadatos asociados a las bases de datos componentes dificulta la automatización del procesamiento de consulta y la agregación de resultados, limitando a los sistemas de bases de datos múltiples a llevar a cabo un proceso de integración horizontal (Chou, 2005).

Sistemas federados

Los sistemas de bases de datos federadas (Sheth *et al*., 1990; Blanco *et al.*, 1994; Conrad, 1997; Buccella *et al.*, 2006) son una colección cooperativa y autónoma de bases de datos componentes, caracterizadas por su distribución y heterogeneidad. Estas bases de datos son

[16] http://nordbotten.com/joan/adam.htm

integradas con respecto a su esquema y modelo de datos mediante el establecimiento de una serie de correspondencias entre las vistas de las diferentes bases de datos componentes. Este hecho genera una visión integrada, así como, una integración escalable y combinada de los datos pertenecientes a las diferentes bases de datos componentes.

Este enfoque de bases de datos es similar al acercamiento de bases de datos múltiples comentado con anterioridad, la única diferencia con el enfoque mencionado es que las bases de datos componentes no son únicamente SGBD relacionales.

En lo que respecta al proceso de integración, estos sistemas permiten que cada administrador local de bases de datos defina un subconjunto de datos. Estos subconjuntos son integrados en una o más bases de datos federadas, también denominadas bases de datos virtuales[17]. Asimismo, la información en los sistemas de bases de datos federadas es agregada bajo demanda, es decir, las instancias de estas bases de datos no tienen una existencia materializada. Estas son temporalmente creadas al vuelo (*on the fly*) según petición del usuario. Por otro lado, estos sistemas son proporcionados a los usuarios con un lenguaje de acceso a múltiples bases de datos (normalmente, SQL o algún lenguaje derivado) sin ninguna unificación de la semántica de los datos desde las diferentes fuentes. Los usuarios pueden, por tanto, acceder a las bases de datos federadas como una base de datos centralizada, sin tener que preocuparse por la localización física de los datos o los tipos de SGBD locales.

De forma generalizada, según Chou (2005), este tipo de sistemas está constituido por tres capas, concretamente, las referidas a datos, exportación e integración.

- La capa de los datos está formada por las bases de datos componentes. Cada base de datos tiene un esquema de exportación que define como se accede y se ven los datos para el mundo exterior.
- La capa de exportación permite que los diferentes modelos de datos de las bases de datos componentes sean unificados, de forma consistente, en un modelo de datos canónico[18] (Castellanos *et al*., 1992; Jang *et al.*, 2007) (o común) definido por la capa de integración.
- La capa de integración es una vista unificada de la información disponible en la capa de datos basada en el establecimiento de correspondencias de la capa de exportación. Estas correspondencias pueden ser especificadas tanto por el usuario como por el administrador del sistema federado.

Por otro lado, los sistemas de bases de datos federadas pueden ser fuertemente acoplados y/o débilmente acoplados, dependiendo de quién defina el esquema de integración. Generalmente, se considera que si el esquema es definido por el usuario, entonces el sistema es considerado como un sistema de acoplamiento débil. Por el contrario, si el esquema integrado es definido por el administrador del sistema federado, entonces el sistema es considerado como un sistema fuertemente acoplado. Ambos tipos de acoplamientos presentan una serie de ventajas e inconvenientes (Chou, 2005) con respecto a su transparencia y autonomía. A continuación, se van a enumerar, sin carácter exhaustivo, algunas de ellas:

- El *acoplamiento débil* tiene un esquema integrado creado por el administrador del sistema, quien gestiona la estructura de los esquemas exportados con los administradores de las bases de datos componentes. Por tanto, las bases de datos componentes tienen menos autonomía y cambiar el esquema de las fuentes normalmente requiere generar un nuevo esquema integrado.

[17] Virtual aquí se refiere al hecho de que sólo se crea el esquema de la base de datos federada.
[18] El modelo de datos canónico (*canonical data model)* es un modelo común para todos los esquemas componentes (Castellanos *et al.*, 1992)

En un sistema con un acoplamiento de estas características, los usuarios definen sus propios esquemas integrados basados en los esquemas de exportación proporcionados por el administrador de las bases de datos componentes. Por tanto, los esquemas de las bases de datos componentes pueden ser cambiados fácilmente sin deshacer los esquemas integrados. Una desventaja de este tipo de acoplamiento es que los usuarios son expuestos a la capa de exportación y, además, éstos deben entender la estructura de los esquemas de exportación para crear sus propios esquemas integrados. Además, las cuestiones referidas a actualización están prohibidas, porque diferentes usuarios requieren diferentes políticas de actualización para sus vistas de las bases de datos, lo que podría crear inconsistencias en los datos durante el proceso de actualización.

- La ventaja de un *acoplamiento fuerte* es que las bases de datos remotas permanecen transparentes al usuario, consecuencia de que el administrador gestiona las políticas de actualización y consulta durante la creación del esquema integrado. Por otro lado, las bases de datos componentes en un acoplamiento débil son más autónomas y menos transparentes.

Al igual que ocurre con los diferentes tipos de acoplamientos, identificados con anterioridad, los sistemas de bases de datos federadas también poseen una serie de ventajas y desventajas definidas sobre este enfoque de integración. Las ventajas de un sistema federado son el soporte de consultas ad hoc, el mantenimiento de datos actualizados y el pequeño almacenamiento adicional requerido para soportar los sistemas de bases de datos federados. En cuanto a sus desventajas, éstas incluyen que muchos de estos sistemas están asociados a una integración débil, provocando consecuencias como; grandes retrasos derivados de la velocidad de la red o lentos tiempos de respuesta de los repositorios, sobrecargas asociadas con la traducción entre diferentes esquemas de fuentes y modelos de datos, interferencias de las consultas locales por consultas federadas y ausencia o no disponibilidad de bases de datos componentes.

Las propuestas de bases de datos federadas en el ámbito de la información geoespacial tampoco resultan muy numerosas. A pesar de esto, entre las propuestas instanciadas en este dominio se encuentran (Parent *et al.*, 1998; Laurini, 1998; Devogele *et al.*, 1998; Morocho *et al.*, 2003; Butenuth *et al.*, 2007).

Data Warehouses

En los *data warehouses* (Chaudhuri *et al.*, 1997; Poe *et al.*, 1998; Inmon, 1996; Kimball *et al.*, 2002; Chou, 2005) o almacenes de datos la información de diferentes repositorios es replicada e integrada en un repositorio centralizado que es consultado directamente por el usuario. Esta propuesta de integración estructural es la implementación más común de la integración consistente.

En este tipo de acercamientos, útiles cuando se requieren análisis pesados de gran cantidad de datos, los datos son extraídos de los repositorios de interés y estos son filtrados, traducidos y fusionados. Estos procesos son los pasos previos para la integración de los datos en un esquema de base de datos común especificado por el repositorio central. La separación generada por este acercamiento entre los almacenes de datos y las fuentes de datos evita interferencias entre el procesamiento de los datos y el funcionamiento de los sistemas en el nivel operacional. Esta separación también confiere una gran flexibilidad para la organización y el procesamiento de los datos.

Según Chou (2005), el hecho de almacenar toda la información en un repositorio centralizado tiene ventajas en lo que respecta a la ausencia de los problemas relacionados con las arquitecturas distribuidas, tales como, cuellos de botella, tiempos de respuesta bajos y fuentes de datos no disponibles. Otra de las grandes ventajas de este acercamiento es la ausencia de inconsistencia en lo concerniente a la representación y formatos. Sin embargo, la cantidad de procesamiento requerido para traducir y fusionar la información de múltiples repositorios a un

esquema de base de datos común es extremadamente elevado y no es un proceso que pueda repetirse con frecuencia. Además, los elevados costes asociados al establecimiento de un almacén de estas características hace difícil la actualización de la información y el añadido de nuevos repositorios una vez instalado. La cuestión de la actualización de los datos también debe ser contemplada, ya que los almacenes de datos no acceden a las fuentes de datos originales para responder a las consultas. Asimismo, existen numerosos problemas asociados al diseño e implementación de almacenes de datos (Widom, 1995).

En el contexto de la IG si se tienen en cuenta factores como; la diversidad, el volumen de información manejado por los diferentes productores, las dificultades de los procesos de actualización periódica, la "privacidad" de información, etc., estos factores provocan verdaderas dificultades para llevar a cabo un proceso de integración de este tipo dentro del dominio de la IG. A pesar de ello, existen algunas implementaciones de estos almacenes de datos en el contexto de la información geoespacial, por ejemplo; Health Resources and Services Administration (HRSA)[19] (Cohen *et al.*, 2005), Farm Service Agency (United States Deparment of Agriculture)[20], (Truter, 2003), (Hultgren *et al.*, 2006), GNWT Spatial Data Warehouse Geospatial Portal[21].

Mediadores

Un mediador (Wiederhold, 1992) combina, integra y abstrae la información proporcionada por una o varias fuentes. En esta serie de procesos, llevados a cabo sobre la información, tienen un papel fundamental los *wrappers* (Wiederhold, 1992; Roth *et al.*, 1997; Levy, 2000; Visser *et al*., 2002), pues éstos proporcionan, de forma independiente a los formatos de los ficheros y/o implementaciones de sistemas, el interfaz o canal de comunicación para acceder a las fuentes de datos, recuperar los resultados, mantener el modelo de estructura interna de los datos y traducir su contenido a un esquema común. Asimismo, cada *wrapper* se encarga de un fichero específico o formato de datos. Por ejemplo, pueden existir *wrappers* para fuentes de datos ODBC (*Open DataBase Connectivity*), ficheros XML o ficheros de determinados formatos SIG.

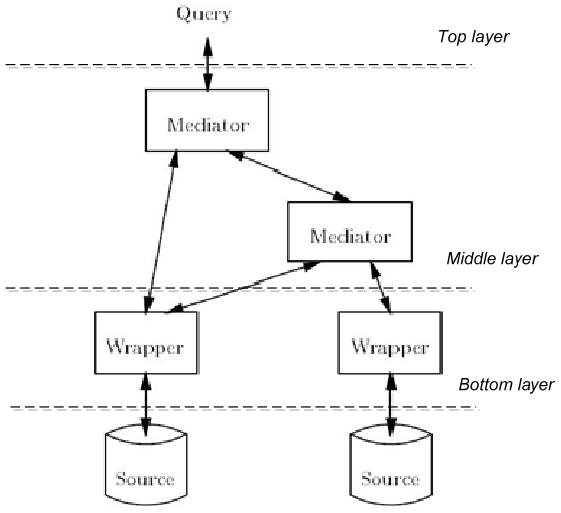


Figura 3.7. Arquitectura común del proceso de integración de información (Ullman, 1997)

Estos acercamientos, con frecuencia, adoptan una arquitectura en tres capas, reflejada en la Figura 3.7. La capa inferior (*bottom layer*) es la base de datos con los *wrappers*. La capa intermedia (*middle layer*) es la consulta soportada por mediadores y la transformación de

[19] *Health Resources and Services Administration* (U.S. Department of Health and Human Services) http://datawarehouse.hrsa.gov/

[20] *Farm Service Agency* (United States Deparment of Agriculture) http://www.fsa.usda.gov/FSA/apfoapp?area=home&subject=maps&topic=gdw

[21] *GNWT Spatial Data Warehouse Geospatial Portal*. http://maps.gnwtgeomatics.nt.ca/portal/index.jsp

resultados entre las fuentes de información *wrapeadas* y las aplicaciones (Guan *et al.*, 2003). Finalmente, la capa superior (*top layer*) es el interfaz para usuarios y aplicaciones. En el proceso de integración de diferentes bases de datos heterogéneas, éstas son *wrapeadas* por diferentes herramientas software que se encargan del proceso de traducción entre el lenguaje local de la fuente, el modelo y sus conceptos, y los conceptos globales compartidos por algunas o todas las fuentes. Los sistemas componentes, llamados mediadores (Wiederhold, 1992) obtienen información de uno o más componentes ubicados bajo éstos, que pueden ser fuentes *wrapeadas* u otros mediadores.

De forma generalizada, la arquitectura de los sistemas mediadores es similar a la propuesta por los sistemas de bases de datos federadas, salvo por el hecho de que los mediadores son diseñados para ser más ligeros y flexibles que los sistemas federados (Chou, 2005). De esta manera, el proceso de integración de información no se ciñe, exclusivamente, a bases de datos, sino que se facilita el soporte para la integración de recursos procedentes de otro tipo de fuentes.

El componente básico de estos sistemas de integración son los datos. Según Ullman (1997), éstos no existen en los mediadores, sino que pueden ser consultados a través del mediador como si los datos estuvieran almacenados, ya que su tarea consiste en ir a las fuentes y encontrar la respuesta a la consulta. A pesar de esto, en ocasiones, los mediadores no pueden ser considerados como tal, sino más bien almacenes de datos. Esta situación resulta de la consideración de un mediador como una vista. De esta manera, un almacén de datos es como una vista materializada, es decir, los almacenes de datos tienen datos que son construidos a partir de los datos en las fuentes. Así, el almacén de datos es consultado directamente sin participación de las fuentes.

Dentro de esta propuesta de integración estructural existen diferentes formas y/o técnicas de instanciación, desarrollando auténticas capacidades de mediación. A continuación se muestra una recopilación de las mismas tanto a nivel general como dentro del dominio geográfico.

Propuestas generales

Dentro del proceso de integración estructural existen multitud de lenguajes y sistemas que llevan a cabo, de forma manifiesta, funciones de mediadores en un sentido amplio. Por tanto, estos lenguajes y sistemas van a llevar a cabo procesos de combinación, integración y abstracción de la información proporcionada por una o varias fuentes. En esta sección se recoge un conjunto de lenguajes que son utilizados para la expresión declarativa de correspondencias entre modelos. Los lenguajes aquí tratados corresponden a los lenguajes operativos. Este tipo de lenguajes permiten la definición de expresiones de condición y transformación sobre los elementos de uno y otro modelo, permitiendo así la definición de correspondencias más complejas y explotables.

Los lenguajes operativos se clasifican conforme al tipo de explotación que se hace de ellos. Así, dentro de los lenguajes operativos se encuentran los lenguajes de integración de información, migración y/o *upgrade* y un grupo de otras propuestas más heterogéneo. Conforme a las características de estos lenguajes y al enfoque de este trabajo, las siguientes secciones se centrarán en realizar una descripción de los diferentes acercamientos relacionados con los lenguajes de integración y migración y/o upgrade.

Lenguajes y sistemas orientados a la integración de información

Los lenguajes operativos orientados a la integración de información permiten explicitar las relaciones existentes entre elementos de uno y otro modelo y además permiten la descripción del tipo de transformaciones necesarias para implementar la correspondencia en cuestión. Este tipo de lenguajes son los más abundantes dentro de los lenguajes operativos. Su misión fundamental es enlazar el modelo mediador global con cada uno de los modelos de las diferentes fuentes de datos o, de forma más concreta, con un *wrapper* del modelo de cada fuente

de datos. Esto provoca que estos lenguajes sean dependientes de *wrappers*. A continuación se recogen los principales lenguajes y sistemas propuestos:

- ***ODL$_I$*** [3]. El *lenguaje ODL$_I$* [3]es una extensión de ODL-ODMG basada en los criterios definidos en (Buneman *et al.*, 1997) para un lenguaje estándar orientado a sistemas de mediación. Se utiliza en el sistema para la integración de información MOMIS (*Mediator EnvirOnment for Multiple Information Sources*) (Bergamaschi *et al.*, 1999; Beneventano *et al.*, 2004).
- ***CARIN***. Este lenguaje se utiliza en el contexto del sistema de integración de información PICSEL (Levy *et al.*, 1998; Goasdoue *et al.*, 2000). CARIN no es propiamente un lenguaje de definición de correspondencias, sino un lenguaje de representación del conocimiento que combina lógica descriptiva con reglas.
- ***OBSERVER: Álgebra Relacional Extendida***. El *sistema OBSERVER* (Mena *et al.*, 2000; Blanco *et al.*, 1999) se propone la definición de correspondencias entre repositorios de datos (modelos origen, generalmente bases de datos relacionales) y ontologías en lógica descriptiva (modelo destino) mediante el uso de álgebra relacional extendida (ERA) (Elmasri *et al.*, 2000).
- ***WOL***. Lenguaje declarativo, creado por las Universidades de Pennsylvania y Berkeley, para especificar correspondencias entre bases de datos y restricciones de integridad, ambas en un entorno de integración de esquemas de datos en un esquema global (Davidson *et al.*, 1998).

Lenguajes orientados a migración y upgrade

Los lenguajes operativos orientados a la migración y/o *upgrade* (Barrasa, 2007) delegan en su mayor parte la definición de las preguntas sobre el modelo origen al propio lenguaje de preguntas de dicho modelo (con frecuencia, modelos relacionales), sin añadirle expresividad en forma de condiciones o transformaciones.

- ***OntomatDBImport***. Este lenguaje desarrollado por la Universidad de Karlsruhe, está basado en XML y permite la definición de correspondencias entre los elementos de un esquema relacional y los de una ontología en RDF. OntomatDBImport (Stojanovic *et al.*, 2002) se utiliza en el módulo Reverse (Volz *et al.*, 2004) del sistema KAON[22], cuyo prototipo fue OntoLift (Volz *et al.*, 2003).
- ***D2R***. Este lenguaje fue desarrollado por la Universidad de Berlín. D2R (Bizer, 2003) está basado en XML y es muy similar a OntomatDBImport, aunque aporta mejoras menores sobre sus antecesores. D2R permite alinear una ontología RDF o derivados con un esquema relacional siguiendo un enfoque GAV.
- ***R$_2$O+***. El *lenguaje R$_2$O+* (Priyatna, 2009), desarrollado por la Universidad Politécnica de Madrid, es una extensión del lenguaje de mappings R$_2$O (Barrasa, 2007). R$_2$O+ permite la explicitación de *mappings* declarativos[23] R$_2$O entre una ontología en RDF(S) y/o OWL y el esquema de varias bases de datos. Este lenguaje es independiente del SGBD, sigue un enfoque GaV y ofrece un conjunto extensible de primitivas propias para la expresión de correspondencias. Este lenguaje se utiliza en el *plugi*
- **ODEMapster++**[24] (Priyatna, 2009) de la herramienta NeOn Toolkit[25]. Este plugin permite dar soporte a bases de datos gestionadas por ORACLE y MySQL y trabajar con múltiples base de datos, debido a que el lenguaje R$_2$O+ contempla en su diseño la explicitación de *mappings* entre una ontología y variadas bases de datos.

[22] *KAON* es un *framework* de gestión de ontologías orientado a aplicaciones de negocio. http://kaon.semanticweb.org/

[23] *Mapping declarativo*. De manera genérica un mapping declarativo puede definirse como una expresión formal que describe la relación entre entidades pertenecientes a dos modelos diferentes (Barrasa, 2007).

[24] http://www.neon-toolkit.org/wiki/ODEMapster

[25] *NeOn toolkit.* Esta herramienta realiza operaciones como la importación/exportación, edición y visualización de ontologías. http://www.neon-toolkit.org/

Propuestas desde la perspectiva geográfica

Dentro del dominio geográfico, según Jang *et al.* (2007), la integración de esquemas supone llevar a cabo un proceso metodológico que pretende integrar estructuras diferentes sobre objetos y representaciones espaciales. El objetivo de este proceso de integración es permitir la reutilización de fuentes de datos existentes en el sentido que propone la interoperabilidad estructural o esquemática. Así, este tipo de integración constituye una metodología práctica y poderosa para aquellos que pretenden integrar diferentes estructuras de objetos y representaciones espaciales para permitir reutilizar las fuentes de datos existentes en el sentido de la interoperabilidad estructural.

Los acercamientos basados en mediadores son un método de integración muy utilizado en los sistemas que soportan la IG. En la actualidad, existe la tendencia (Boulcema *et al.*, 2002; Díaz *et al*., 2008) de utilizar mediadores y *wrappers* en estos sistemas para implementar los interfaces estándar especificados por OGC, dadas las necesidades de intercambio estandarizado, compartición, acceso y procesamiento de información geoespacial.

Las aplicaciones de usuario pueden acceder directamente a las fuentes de datos espaciales como en los sistemas de base de datos tradicionales. A pesar de ello, estas aplicaciones, habitualmente, utilizan un conjunto de interfaces de servicios geoespaciales[26] estándar para descubrir y acceder a fuentes de información en IDE con el objetivo de integrar datos espaciales. Estos interfaces, creados conforme a los estándares de OGC, actúan como *wrappers* para añadir el nivel de abstracción necesario para la integración sintáctica-estructural de datos espaciales. Una visión genérica, ofrecida de manera gráfica en la Figura 3.8, del uso de *wrappers* en el contexto de la IG puede ser comprobada en (Boulcema *et al.*, 2002; Díaz *et al*., 2008).

Geographic Markup Language

La adopción de un modelo de datos geográficos común o, al menos, un marco para unificar modelos heterogéneos constituye un ingrediente para alcanzar la interoperabilidad a nivel estructural de los SIG en un sentido amplio. Un primer intento en esta línea ha sido llevado a cabo por el estándar GML.

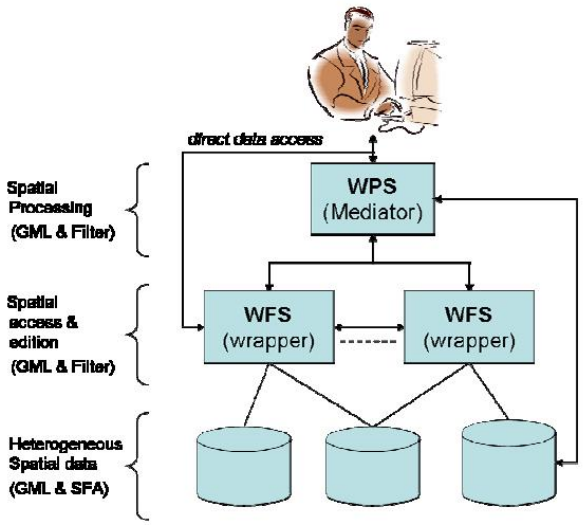


Figura 3.8. Componentes de la integración de datos espaciales en la web (Díaz *et al.*, 2008)

[26] Los interfaces de servicios poseen una sintaxis estándar para el proceso de petición de datos geoespaciales en la web. Ejemplos de este tipo de interfaces son WMS (*Web Map Service*), WFS (*Web Feature Service*) y WCS (*Web Coverage Service*)

Desde el punto de vista de la implementación, se puede decidir almacenar los esquemas de aplicaciones geográficas y la información en GML o, por el contrario, se puede optar por convertir desde algún otro formato de almacenamiento bajo demanda y utilizar GML sólo para el intercambio de información entre sistemas. Generalmente, los productores de IG almacenan sus datos en bases de datos locales y utilizan tecnologías de transmisión de datos para acceder a dicha información. Es aquí donde el estándar GML cumple una doble función; por un lado, ofrece acceso y recuperación de datos espaciales en un formato de intercambio común y, por otro, contribuye a conseguir la interoperabilidad estructural entre los diferentes sistemas de información que manejan a la información geoespacial.

Por otro lado, las implementaciones de bases de datos con extensiones espaciales permiten almacenar directamente datos en formato GML, permitiendo así la interacción con otras fuentes de datos espaciales y no espaciales mediante la utilización de un estándar internacional.

Sistemas de integración

Son diferentes las propuestas que aparecen en el área de integración de datos geoespaciales utilizando un sistema de mediación en el proceso de integración. A continuación, sin pretender ser exhaustivo, se destacan algunos ejemplos de estos sistemas utilizados para integrar información geoespacial.

Hermes (Adali *et al.*, 1995) es un sistema mediator que integra datos multimedia y datos espaciales entre otros datos. El sistema MIX (Gupta *et al.*, 1999) es un sistema mediador, basado en XML, que soporta la integración de datos geoespaciales y algunas operaciones de agregación y agrupación utilizando el enfoque GAV. VirGIS (Essid *et al.*, 2004) es un sistema mediador que utiliza una forma limitada del enfoque LAV para integrar datos de diferentes fuentes geoespaciales. Además, este sistema soporta el establecimiento de correspondencias uno a uno (1:1) entre fuentes y relaciones de dominio. Geongrid (Zaslavsky *et al.*, 2003) describe una arquitectura de servicios de mediación con capacidades de *grid* (GMES) para integrar datos geoespaciales. La arquitectura de GMES se centra en proporcionar los mejores resultados por selección de fuentes con la mejor calidad utilizando un ranking predefinido de fuentes basadas en metadatos de calidad.

Otras propuestas

Mientras las cuestiones y las resoluciones de las metodologías de integración de esquemas han sido investigadas a fondo para las bases de datos convencionales, ha habido un pequeño número de trabajos sobre integración sistemática de esquemas para sistemas de información espacial.

(Laurini, 1998) identifica, de forma detallada, los únicos retos de la integración de esquemas para las bases de datos espaciales incluyendo *matching* para diferentes representaciones espaciales, *matching* de discrepancias geométricas y de límites. Como una versión modificada del procedimiento de (Parent *et al.*, 1998) en (Laurini, 1998) se sugiere un procedimiento de integración de esquemas general para bases de datos espaciales múltiples por inclusión de cuatro procesos más: resolución de conflictos geométricos, alineamiento de límites, continuidad topológica e indexación espacial. Esto implica que la construcción de un sistema de base de datos espacial federado es mucho más complejo que los sistemas de bases de datos federados convencionales debido a las características especiales de la información espacial.

(Devogele *et al.*, 1998) demuestran procedimientos de integración de esquemas más tangibles ya que muestran las cuestiones y sus resoluciones en el proceso de construcción de un esquema integrado utilizando dos bases de datos espaciales con representaciones espaciales diferentes debido a las diferencias de escala. Estos autores utilizan *Unified Modeling Language* (UML) para esbozar el modelo de datos que cada base de datos espacial emplea mientras describe las correlaciones de esquema entre los dos esquemas locales utilizando una declaración de correspondencia inter-esquema.

En (Jang *et al.*, 2007) se propone un acercamiento de modelo canónico basado en los conceptos estandarizados adoptados en la serie ISO 19100 de estándares internacionales para IG. Este marco conceptual es utilizado para lograr una compartición desambiguada de conceptos en el dominio de la información geográfica. Este acercamiento es evaluado en (Jang *et al.*, 2006) mediante la creación de esquemas de aplicación para una aplicación SIG específica donde la información geoespacial tiene un papel crucial. Además, se implementa una codificación independiente de plataforma del esquema de aplicación creado y se desarrolla un prototipo para evaluar el éxito de la aplicación GIS interoperable semánticamente utilizando servicios web.

En (Sotnykova *et al.*, 2005) se propone un acercamiento para integrar esquemas de bases de datos espacio-temporales mediante dos formalismos bien conocidos; modelos conceptuales y lógica descriptiva. El modelo conceptual MADS[27] (MurMur Project, 2002; Parent *et al.*, 2006) pretende describir los datos de aplicación espacio-temporal. Un elemento particular de este modelo en el proceso de integración de datos es que incluye conceptos específicos para describir representaciones múltiples de los datos. Además, permite describir la semántica de los esquemas espacio-temporales tanto iniciales como integrados. La lógica descriptiva es utilizada para expresar los conjuntos de correspondencias inter-esquema. Asimismo, los mecanismos de inferencia de la lógica descriptiva permiten comprobar la compatibilidad de las correspondencias semánticas y proponer diferentes soluciones estructurales para el esquema integrado.

3.2.7.3. Integración semántica

Mientras la integración estructural ha sido ampliamente definida, como se describe en la sección 3.2.7.2, el tratamiento de la perspectiva semántica es un proceso más complicado. Este proceso de integración presenta verdaderos retos (Arroyo *et al.*, 2003; Visser, 2005), producto de la heterogeneidad semántica. Asimismo, este tipo de heterogeneidad es el principal problema a solventar en la integración de datos geoespaciales (Vckovski, 1998). Otros autores de la comunidad SIG llaman a este problema problemas de inconsistencias (Shepherd, 1991). Los problemas generados por la heterogeneidad semántica y las necesidades de integración están conllevando un cambio en el foco de interés del proceso de integración de información, ya que como se refleja en (Sheth, 1998) se ha pasado del interés por la integración sintáctica y estructural a la apuesta por la integración semántica de fuentes heterogéneas.

El proceso de integración semántica (Visser, 2005), directamente relacionado con la interoperabilidad semántica (sección 2.1.1.3), se refiere a los mecanismos que permiten a dos o más sistemas compartir e integrar información desde diferentes fuentes con el fin de superar los problemas que provoca en la información la heterogeneidad semántica. Esta nueva óptica de integración de información aparece como consecuencia de que una ontología proporciona una terminología unificada, completa y coherente de un determinado dominio, que puede ser utilizada de manera consistente, precisa y adecuada en diferentes aplicaciones. Las ontologías tratan de facilitar la compartición de conocimientos, al proporcionar un vocabulario consensuado en un área y al definir, con diferentes grados de formalidad, el significado de los términos y relaciones entre ellos (Gruber, 93). Asimismo, se puede hacer efectiva la consecución de una visión común, compartida e interoperable de los fenómenos presentes en la realidad geográfica, eliminando obstáculos para su utilización, consulta y/o asignación de valor añadido. De esta manera, este tipo de integración es habitualmente utilizado para crear servicios de información agregada que pueden ser utilizados para consultar fuentes de información múltiples de manera transparente (Wache *et al.*, 2001). Así, las ontologías son identificadas como un instrumento útil en el proceso de integración (Grüninger *et al*., 2007), donde, además, éstas pueden ser utilizadas en el proceso de integración semántica mediante su uso para describir las fuentes de información (Visser, 2005).

[27] http://lbdwww.epfl.ch/e/MurMur

En la consideración de integración de información entendida desde el contexto de las bases de datos, el tipo de semántica considerada es, generalmente, la semántica del mundo real que concierne a "*las correspondencias de objetos del mundo computacional o modelo sobre el mundo real* […] y *las cuestiones que involucran a las interpretaciones humanas o los significados y usos de los datos e información*" (Ouksel *et al.*, 1999). En esta concepción, la integración semántica consiste en la tarea de agrupar, combinar o completar datos procedentes de diferentes fuentes considerando la semántica precisa y explícita de los datos para evitar que datos incompatibles semánticamente sean fusionados estructuralmente. Es decir, la integración semántica tiene que asegurar que sólo son fusionados los datos relacionados con los mismos o suficientemente[28] similares fenómenos o conceptos del mundo real (Ziegler *et al.*, 2004).

Integración de información basada en ontologías

En el contexto de la integración semántica, considerando a las ontologías como base de esta propuesta de integración, se van a considerar en este trabajo dos tipos de propuestas de integración de información, concretamente, las referidas a propuestas generalistas y las realizadas en el contexto de la información geoespacial.

Propuestas generales

Existen diversos enfoques propuestos en el marco de la integración de información basada en ontologías. Conforme a la clasificación propuesta en (Wache *et al.*, 2001), las principales propuestas son:

- **Enfoques basados en una única ontología**[29]. Algunas de las principales propuestas desarrolladas en el contexto de este enfoque son: SIMS (Arens *et al.*, 1996), Prometheus (Padgham *et al.*, 2002), Theseus (Barish *et al.*, 2000), MOMIS (Bergamaschi *et al.*, 1999; Beneventano *et al.*, 2004), DartGrid (Chen *et al.*, 2006a; Chen *et al.*, 2006b), Carnot (Singh *et al.*, 1997), InfoSleuth (Bayardo *et al.*, 1997), DOME (Cui *et al.*, 2000; Cui *et al.*, 2001).
- **Enfoques basados en múltiples ontologías**[30]. Dentro de este enfoque se encuentran las siguientes propuestas de integración de información: OBSERVER (Mena *et al.*, 2000), PICSEL (Lattes *et al.*, 2000) y TAMBIS (Goble *et al.*, 2001), TSIMMIS (Chawathe *et al.*, 1994), Information Manifold (Levy *et al.*, 1996), DWQ (Jeusfeld *et al.*, 1998), KnowledgeParser (Contreras *et al.*, 2004).
- **Enfoques híbridos**[31]. Las principales propuestas de este enfoque son: COIN (Siegel *et al.*, 1991; Goh *et al.*, 1994; Goh, 1997; Goh *et al.*, 1999; Firat *et al.*, 2002), MECOTA (Wache *et al.*, 1999) y BUSTER (Visser, 2005).

Estas propuestas proporcionan acercamientos de procesamiento de consultas basado en conceptos. En estos sistemas, las ontologías son utilizadas para proporcionar especificaciones

[28] La cantidad de similitud considerada suficiente depende de la necesidad particular de información y el área de aplicación (Ziegler *et al*., 2004).

[29] Los enfoques de una única ontología utilizan una ontología global para proporcionar un vocabulario compartido para la especificación de la semántica. De esta manera, todas las fuentes de información se relacionan con una ontología global (Wache et al., 2001).

[30] En los enfoques basados en múltiples ontologías cada fuente de información se describe por su propia ontología. En principio, *la ontología de la fuente* puede ser una combinación de varias ontologías, pero no se asume que las diferentes *ontologías de las fuentes* compartan el mismo vocabulario (Wache et al., 2001).

[31] En los enfoques híbridos la semántica de cada fuente se describe por su propia ontología, aunque para hacer comparable las ontologías de las diferentes fuentes se contruye un vocabulario compartido global. Los enfoques híbridos se desarrollaron para superar los inconvenientes generados en los enfoques basdoas en ontologías únicas y múltiples (Wache et al., 2001).

concisas y declarativas de la información semántica (Mena *et al.*, 2000). En la Tabla 3.2 se muestra una comparativa de las diferentes propuestas inspirada en (Xing *et al*., 2007).

Tabla 3.2. Comparativa entre las diferentes propuestas de integración (Xing *et al.*, 2007)

	General Approach		Information Source Wrapping		
	Information Access Approach	Mapping Approach	Information Source Selection	Type of Legacy Data	Wrapper Generation
TSIMMIS	On-demand Access	Local as view	Pre-defined	Structured data DBs	Automatic
Information Manifold	On-demand Access	Global as view	Pre-defined	Structured data DBs	Manual
Carnot	On-demand Access	Global as view	Pre-defined	DBs	Manual
InfoSleuth	On-demand Access	Local as view	Dynamic	DBs	Manual
SIMS	On-demand Access	Global as view	Dynamic	Structured data DBs	Manual
Prometheus +Theseus	On-demand Access	Global as view	Pre-defined	HTML, Images DBs	Semi-automatic
OBSERVER	On-demand Access	Local as view	Pre-defined	DBs, XML bib, HTML	Manual
TAMBIS	On-demand Access	Global as view	Pre-defined	DBs, XML, HTML	Manual
PICSEL	On-demand Access + Data Warehouse	Local as view	Pre-defined	Dbs, XML, HTML	Semi-automatic
DWQ	Data Warehouse	Local as view	Pre-defined	DBs	Manual
Knowledge Parser	Data Warehouse	Local as view	Pre-defined	HTML, pdf, Word, DB	Semi-automatic
Active Ontology	On-demand Access + Data Warehouse	Global as view	Dynamic	Structured data DBs	Manual

Propuestas en el contexto geográfico

Un acercamiento, cada vez más extendido, para capturar la semántica y solventar los problemas causados por la heterogeneidad semántica en el proceso de integración de información geoespacial es el uso de ontologías. A continuación se recogen algunas propuestas de integración semántica de información geoespacial dividas en las siguientes categorías: integración semántica de fuentes geoespaciales, enriquecimiento semántico e integración vía servicios.

Integración semántica de fuentes geoespaciales

Resultan numerosos los trabajos que aparecen en el área de integración semántica de datos geoespaciales. A continuación se describen, ordenadas cronológicamente, algunas de las principales propuestas dentro de esta área.

Los trabajos de (Uitermark *et al*. 1999a; Uitermark, 2001) introducen un acercamiento para integración de datos espaciales entre dos conjuntos de datos por actualización del conjunto de datos más antiguos según el conjunto de datos más nuevo. Los autores definen dos conjuntos de relaciones; un conjunto entre ontología de aplicación y de dominio (relaciones de agregación e igualdad) y el segundo conjunto entre ontologías de aplicación (relaciones de relevancia, relación e igualdad). Las relaciones de similitud son tomadas como conocidas y utilizadas para relacionar los objetos en dos conjuntos de datos. La localización espacial de los objetos es también utilizada para refinar el resultado. El resultado final es utilizado para actualizar los conjuntos de datos más antiguos (Uitermark *et al*., 1999b).

En (Stuckenschmidt *et al.*, 1999; Visser *et al.*, 2002) se propone una visión general de las ontologías formales, como estas pueden ser utilizadas en el proceso de integración de información geoespacial de diferentes catálogos. El uso de ontologías en esta propuesta se

enfoca hacia el establecimiento de traductores semánticos entre las fuentes de datos. De esta manera, la utilización de ontologías, formalizadas con *Ontolingua*[32], proporciona vistas integradas de datos de los diferentes catálogos y procesos de verificación de los datos de *ATKIS-OK-250* (AdV, 1998) con los datos de *Corine Land Cover*[33] (European Environment Agency, 1997-1999) o viceversa.

En los trabajos de (Hakimpour *et al.*, 2001; Hakimpour, 2003) se presenta un acercamiento para integrar esquemas de diferentes comunidades, donde cada comunidad utiliza su propia ontología. Este acercamiento está basado en la fusión de ontologías siguiendo relaciones de similitud entre conceptos de diferentes ontologías. La definición formal de las relaciones de similitud está basada en definiciones intensionales y decide las consecuencias extensionales. Finalmente, la ontología fusionada es utilizada para derivar un esquema integrado, utilizado como esquema global en un sistema de base de datos federada.

En (Fonseca *et al.*, 2002) el objetivo es establecer una arquitectura SIG que permita la integración de información geográfica, teniendo en cuenta su valor semántico e independencia de representación, de una forma homogénea y flexible. La solución propuesta en este trabajo es un sistema de información geográfica guiado por ontologías (ODGIS) (Fonseca *et al.*, 1999) que actúa como un integrador del sistema independiente del modelo. En ODGIS, una ontología es un componente más, como puede serlo la base de datos, que coopera para cumplir con los objetivos del sistema. Asimismo, la utilización de términos de WordNet para resolver cuestiones de conexión entre fuentes y el establecimiento de correspondencias entre múltiples ontologías proporcionan un elevado nivel de integración. En definitiva, se lleva a cabo la integración semántica de imágenes aéreas, SIG y la posibilidad de añadir otros tipos de información (diferentes sensores), contribuyendo a la mejora del modelado geoespacial.

En (Cruz *et al.*, 2003) se define un modelo basado en XML para representar relaciones de equivalencia 1-1 y 1-n entre clases en una dominio jerárquico. Además, se proporciona un simple mecanismo de consulta que considera correspondencias para integrar datos de dos conjuntos de datos basado en XPath. Sin embargo, las correspondencias tiene que ser establecidas de forma manual y sólo se definen tipos simples de relaciones.

Lin *et al.* (2004) describen un sistema prototipo, en el contexto del proyecto GEON[34], para integrar mapas geológicos de diferentes superficies geológicas. El objetivo es integrar conjuntos de datos geológicos disponibles para proporcionar una web basada en mapas geológicos interactivos para encontrar la localización de determinadas propiedades, respecto a eras geológicas, tipos de rocas o una combinación de estas. Los conjuntos de datos geológicos son registrados a través de ontologías para ayudar en el proceso de integración y consulta de estos conjuntos de datos heterogéneos. El sistema propuesto está compuesto de tres componentes; un repositorio de ontologías (eras geológicas y tipos de rocas), donde vocabularios de fuentes heterogéneas se hacen compatibles y donde, además, el usuario puede crear y guardar ontologías en OWL para que éstas pueden ser utilizadas por el sistema; el registro de conjuntos de datos (mapas geológicos) y las aplicaciones que soportan ontologías. Por otro lado, las correspondencias entre conjuntos de datos y ontologías son establecidas por aplicaciones para explorar y extraer información de los conjuntos de datos.

BUSTER (Visser, 2005) es un sistema completo para integrar y consultar información heterogénea de diferentes conjuntos de datos geoespaciales. Dicho sistema soporta consultas de tipo concepto-localización y proporciona servicios de integración a nivel sintáctico, estructural y semántico. A nivel sintáctico este sistema utiliza *wrappers*, mediadores a nivel estructural y

[32] *Ontolingua*. http://www.ksl.stanford.edu/software/ontolingua/
[33] *Corine Land Cover*. http://reports.eea.europa.eu/COR0-landcover/en
[34] *GEON*. http://www.geongrid.org

para el nivel semántico emplea motores de reglas de transformación de contexto (*CTR-Engines*) y clasificadores (*mappers*).
El proceso de integración de conjuntos de datos topográficos realizado en (Uitermark, *et al*., 2005) establece relaciones entre las instancias de objetos correspondientes de diferentes y autónomos conjuntos de datos topográficos del mismo espacio geográfico, considerando las diferencias en contenido y abstracción. Para ello, desarrollan un marco conceptual cuyos componentes principales son las ontologías y los conjuntos de reglas contempladas. Asimismo, introducen un modelo de referencia al que pertenece, únicamente, la combinación de conjuntos de datos topográficos que van a ser integrados.

Arpinar *et al.* (2006) describen el proceso de desarrollo manual de una ontología y modela las fuentes utilizando la ontología desarrollada manualmente en un marco titulado Semántica Geoespacial y Analítica (GSA).

En (Paul *et al*., 2006) el objetivo principal es integrar diversos repositorios espaciales para aplicaciones geográficas utilizando una arquitectura orientada a servicios (SOA) para descubrir y recuperar información geoespacial. La arquitectura utiliza una ontología central como información de metadatos que actúa como agente de servicios. Asimismo, el sistema está compuesto de una ontología de dominio y de las ontologías de aplicación proveedoras de servicios que necesitan adoptar la ontología central.

En (Roman *et al.*, 2007) se describe un marco semántico para los servicios geoespaciales, llamado SWING. Ese marco muestra la viabilidad de utilizar un razonador basado en ontologías para superar la heterogeneidad presente en los nombres de las capas y atributos de los diferentes servicios geoespaciales. Esta propuesta es similar al acercamiento de los mediadores ya que estos utilizan una forma simple de descripción lógica para codificar las reglas.

Buccella *et al.* (2007) introducen un proceso de integración basado en dos conjuntos principales de tareas. Por un lado, las tareas no lógicas apuntan a encontrar similitudes basadas en análisis sintácticos y estructurales de los datos geográficos. Por otro lado, las tareas lógicas son utilizadas para calcular inferencias de la semántica de los datos mediante el uso de ontologías. En el contexto de esta tarea tiene gran importancia la búsqueda de correspondencias entre dos ontologías geográficas. Estas correspondencias son utilizadas para generar una ontología global integrada que contiene la estructura de los datos de un sistema federado. En definitiva, esta propuesta es a nivel de esquema, es decir, las instancias de los conceptos no son tenidas en cuenta sino sólo la estructura de los datos. Desde estos datos se construyen las ontologías formales.

Green *et al.* (2008) presentan un prototipo que lleva a cabo la integración de datos de diferentes organizaciones de una manera explícita, produciendo una ontología fusionada que resuelve las cuestiones de integración. Este trabajo ilustra, en el contexto de la contaminación de aguas, diversos aspectos del mencionado prototipo, tales como:

- La descripción del recubrimiento de los dominios por parte de los datos a través de las ontologías de dominio.
- El establecimiento de correspondencias entre las ontologías de dominio y las bases de datos físicas se lleva a cabo mediante una ontología de datos.
- La producción de una ontología fusionada (ontología de aplicación) describe la aplicación por documentación de tareas y combinación de ontologías de dominio que, de forma indirecta, conecta los datos.
- La ejecución de las consultas semánticas incluye elementos espaciales sobre fuentes de datos múltiples utilizando el vocabulario de la ontología de aplicación.

Gómez-Pérez *et al.* (2008) integran diferentes fuentes de información con el objetivo de construir un vocabulario común de referencia y, de esta manera, gestionar la enorme cantidad de información gestionada por el Instituto Geográfico Nacional. Para ello, se crea automáticamente

una ontología, denominada *PhenomenOntology*, utilizando la semántica de los catálogos seleccionados y, posteriormente, se procede a descubrir correspondencias de manera automática con el objetivo de relacionar los conceptos de la ontología con las instancias de las bases de datos.

Enriquecimiento semántico

Algunas otras propuestas necesitan añadir la especificación de metadatos o explicación del contexto de los datos que se están integrando. De esta manera, los metadatos van a ser utilizados para describir el contenido de los datos subyacentes de una manera fácil de entender, mientras el contexto es una descripción más compleja que especifica el dominio de las fuentes de datos. A continuación se muestran algunas de las propuestas que siguen este enfoque, aunque no de forma exhaustiva dado que no es el foco central en este trabajo.

En (Souza *et al.*, 2006) se plantea la utilización de la información de contexto para ayudar a mejorar la integración de datos geoespaciales en aspectos como: (1) la generación de correspondencias inter-esquema, ya que ayuda a determinar el significado de los términos; (2) la respuesta a consultas resulta más relevante; (3) los conflictos geoespaciales específicos son mejor solventados conforme al contexto de la formulación de consultas, operaciones espaciales necesarias y nivel de detalle previsto. Estos autores desarrollan una ontología para representación del contexto relacionad con los sistemas de integración de datos geoespaciales, ya que el contexto tiene gran importancia cuando proporciona a los usuarios respuestas más significativas.

Ram *et al.* (2001) propone un modelo semántico espacio-temporal para definir el contenido de la información geoespacial de las fuentes. Asimismo, emplea una ontología de resolución de conflictos para resolver la heterogeneidad semántica y usa metadatos geoespaciales, tal y como especifica *Federal Geographic Data Committee* (FGDC, 2001), para ayudar a los usuarios a evaluar la utilidad de las fuentes de datos disponibles y para generar, de manera eficiente, planes de consulta. El prototipo del sistema, denominado *GeoCom*, está basado en web con el objeto de ayudar a integrar datos geoespaciales heterogéneos distribuidos de forma autónoma.

En (King *et al.*, 2005) se propone un marco ontológico para resolver los problemas de heterogeneidad semántica en los datos de las zonas costeras. Este marco proporciona la capacidad de conectar a los usuarios al conocimiento mediante visualización integrada de la información. Además proporciona facilidad en los procesos de búsqueda y respuesta a consultas y recoge información a diferentes niveles de granularidad. Este trabajo aporta el desarrollo del Sistema Integrado de Observación del Océano (*Integrated Ocean Observing System* (IOOS)) que cuenta con variadas fuentes de información. Esta variedad de fuentes hace importante la representación del contexto de cada fuente a través de los metadatos. Con respecto a las ontologías, éstas son utilizadas en la tarea de integración de datos por la descripción de la semántica (contexto) de las fuentes de información. Asimismo, la construcción de ontologías en este acercamiento guía al middleware, lo que permite la transformación de datos e información heterogéneos en conocimiento dada su utilización inteligente de recursos relevantes.

Integración vía servicios

Desde la perspectiva geográfica la necesidad de añadir semántica a las especificaciones sintácticas de servicios fue reconocida por OGC desde sus inicios. Este hecho llevó a la visión de servicios interoperables dentro o a través de las comunidades de información mediante el proceso de traducción semántica (OpenGIS Consortium 1999a). Conforme a esta visión, surge el proceso de integración semántica a través de servicios web.

Según Ziegler *et al.* (2004), la integración de información a través de servicios web lleva a cabo el proceso de integración mediante componentes software que soportan el proceso de interacción máquina a máquina sobre una red. El proceso de comunicación entre máquinas está basado en el intercambio de mensajes XML transportados por los protocolos de internet.

Dependiendo de la funcionalidad de integración ofrecida, los servicios web representan una propuesta uniforme de acceso a los datos o un interfaz de acceso común a los datos para llevar a cabo, posteriormente, una integración manual o basada en aplicaciones. A continuación se van a reflejan algunas propuestas basadas en esta línea de trabajo.

En (Navarrete, 2006) se define un marco semántico cuyo núcleo es una ontología que representa los conceptos temáticos en un repositorio de conjuntos de datos, así como las relaciones entre dichos conceptos. La ontología se construye mediante un proceso de fusión (*merging*) de las ontologías de aplicación de los conjuntos de datos que se han insertado en el repositorio. En este trabajo se propone un método semiautomático de fusión, para el que se desarrollan tres algoritmos diferentes con el objetivo de generar una lista de sugerencias de operaciones de mapeado que un experto podrá aceptar o modificar. Además, este marco semántico permite la definición de servicios semánticos que van más allá de las funcionalidades que los actuales catálogos de información geográfica ofrecen. En concreto, uno de los tres servicios semánticos definidos se centra en la integración en un nuevo conjunto de datos de información temática proveniente de diversas fuentes. Finalmente, el marco semántico y sus servicios son utilizados en un sistema de indexación y recuperación de elementos multimedia geo-referenciados (imágenes estáticas y secuencias de vídeo) a partir de su contenido geográfico temático.

Tanasescu *et al.* (2007); Gugliotta *et al.* (2008) proponen un acercamiento para la integración de datos geoespaciales basado en Servicios Web Semánticos (*Semantic Web Services* - SWS) denominado *eMerges*[35]. Esta propuesta permite la representación unificada y manipulación de fuentes de datos espaciales heterogéneas. *eMerges* lleva a cabo el proceso de integración por mediación del legado de fuentes de datos espaciales a ontologías espaciales de alto nivel a través de SWS y por representación de cada acción dependiente del contexto del objeto. Este acercamiento es aplicado en el contexto de la gestión de emergencias.

En (Lutz *et al.*, 2009) se superan los problemas causados por la heterogeneidad semántica en tres niveles (metadatos, esquemas y al nivel de contenido de los datos). Para ello, los autores presentan un método basado en ontologías y razonamiento lógico, que mejora el descubrimiento, recuperación, interpretación e integración de datos geográficos en las infraestructuras de datos espaciales. Este trabajo se ilustra con ejemplos de los dominios geológico e hidrológico.

3.3. Detección de duplicados

Un aspecto importante de la integración de información es la capacidad de integrar datos de la misma entidad (denominado fenómeno geográfico en el contexto de la información geoespacial) recogidos en diferentes fuentes de información.

La diversidad de fuentes existentes con información parcialmente superpuesta (desarrolladas por diferentes organizaciones, bajo diferentes premisas y requerimientos, y utilizando diversos modelos de datos y/o significados) hace que una misma instancia pueda aparecer en múltiples ocasiones en diferentes fuentes de información. Además, ésta puede ser estructurada bajo diversas formas y/o pueden variar los valores registrados, prevaleciendo cierto grado de heterogeneidad. Una descripción detallada de los diferentes tipos de heterogeneidad que afectan a la información geoespacial se describe en el Capítulo 2 (sección 2.2). Esta multiplicidad y heterogeneidad hace necesario un proceso de detección de duplicados de instancia para realizar una auténtica integración de las fuentes de información.

El objetivo de la detección de instancias duplicadas es identificar registros o entidades (también denominadas instancias o recursos), con representaciones múltiples, referidas a la misma

[35] *eMerges*. http://irs-test.open.ac.uk/sgis-dev/

entidad del mundo real, incluso si los registros no son idénticos. Esta identificación, dado que las entidades a resolver pueden estar localizadas en entornos distribuidos y ser fuentes heterogéneas, puede realizarse en la misma o en diferentes fuentes de información.

Según Weis *et al*. (2008) la detección de duplicados tiene dos importantes aspectos a considerar: Primero, las múltiples representaciones normalmente no son idénticas, ya que pueden existir diferencias, por ejemplo: errores ortográficos, alteraciones o ausencia de valores, etc. Segundo, la detección de duplicados es una operación que requiere un gran procesamiento computacional, ya que requiere la comparación de cada posible pareja de duplicados utilizando medidas de similitud complejas.

A continuación, se describe el tratamiento de la detección de duplicados desde diversas ramas de la ciencia de la información (estadística, bases de datos e inteligencia artificial) en la sección 3.3.1. En las secciones 3.3.2 y 3.3.3 se recopilan diversos enfoques de la detección de duplicados en el contexto de la Web de los Datos y en la información geoespacial, respectivamente.

3.3.1. Un problema con múltiples tratamientos

El problema de la detección de duplicados, ampliamente estudiado, se trata en diversas ramas de la ciencia de la información. A continuación se describen diversas propuestas agrupadas en torno a las principales áreas que tratan esta problemática. Finalmente, en esta sección se recogen algunos acercamientos en el dominio de la información geoespacial.

En la **comunidad estadística** existe una extensa línea de investigación (más de cinco décadas) basada en la vinculación de instancias. Esta línea de trabajo es conocida como vinculación de registros probabilística (*probabilistic record linkage*) (Fellegi *et al*., 1969), vinculación de registros (*record linkage*) o problemas de emparejamiento de registros (*record matching problems*) (Newcombe, 1967; Newcombe *et al.*, 1959; Tepping, 1968).
Desde el campo estadístico, la vinculación de registros se define como la metodología para reunir los registros correspondientes desde dos o más ficheros o encontrar duplicados dentro de las filas (Winkler, 1999).

El concepto *record linkage* proviene del área de la salud pública. Este concepto se utilizó cuando se juntaron ficheros de pacientes utilizando nombres, fechas de nacimiento y otra información. Las ideas de *record linkage*, inicialmente, fueron originadas por (Newcombe *et al.*, 1959; Newcombe, 1967) quien introdujo curiosos ratios de frecuencias y las reglas de decisión para definir emparejamientos y desemparejamientos. Con posterioridad, (Fellegi *et al.*, 1969) proporcionan fundamentos matemáticos formales acerca de *record linkage*.

La teoría está basada en el uso de registros que se refieren a la misma entidad teniendo un número de características en común. Si a y b son elementos de poblaciones A y B y algunos de los elementos son comunes a A y B, entonces dos conjuntos disjuntos pueden ser creados. El primer conjunto, M, es el conjunto de elementos que representan entidades idénticas y el segundo conjunto, U, es el conjunto de elementos que representan entidades diferentes. Si $\alpha(a)$ y $\beta(b)$ se refieren a registros de las bases de datos A y B respectivamente, y cada registro tiene k características, entonces un vector de comparación γ es definido para que contenga el acuerdo codificado y el desacuerdo sobre cada característica.

$$\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], \dots \gamma^k[\alpha(a), \beta(b)]\}$$

Esta línea de investigación, fundamentalmente, ha estado basada en el trabajo de (Fellegi *et al*., 1969). Este trabajo formula el emparejamiento de entidades (*entity matching*) como un problema de clasificación, donde el objetivo básico es clasificar pares de entidades como emparejadas o no emparejadas. Estos autores (Fellegi *et al*., 1969) proponen utilizar en gran

parte métodos no supervisados para esta tarea. Estos métodos estarían basados en una representación de pares de elementos que serían manualmente designados y que contendrían alguna extensión de problemas específicos. Estas propuestas han sido ampliamente adoptadas en diversos trabajos, tales como (Jaro, 1989; 1995; Winkler, 1999; Larsen, 1999; Belin *et al.*, 1997).

Dentro de la **comunidad de bases de datos** este problema es referido como identificación de instancias (*instance identification*) (Wang *et al.*, 1989), fusión-limpieza (*merge-purge*) (Hernández *et al.*, 1998), identificación de objetos (*object identification*) (Tejada *et al.*, 2001), deduplicación de datos (*data deduplication*) (Sarawagi *et al.*, 2002) y/o resolución de entidades (*entity resolution*) (Menestrina *et al.*, 2006). Este problema de investigación se deriva de la necesidad de vincular registros cuando los registros o filas de diferentes bases de datos necesitan ser fusionados o unidos. En este proceso cada base de datos puede tener instancias duplicadas del mismo registro. Esto podría hacer los datos inconsistentes o sucios (Hernández *et al.*, 1998).

En éste área, son diferentes los estudios (*surveys*) existentes (Gu *et al.*, 2003; Winkler, 2006; Batini *et al.*, 2006; Elmagarmid *et al.*, 2007; Herzog *et al.*, 2007; Bleiholder *et al.*, 2008). Estos trabajos, de forma general, se centran en los problemas de emparejamiento de registros, es decir, en aquellos problemas que tratan sobre qué registros representan a la misma entidad y cuáles no. A continuación, se realiza una breve descripción de algunos de estos trabajos.

En (Elmagarmid *et al.*, 2007) se presenta un estudio de los acercamientos y algoritmos existentes para mejorar la eficiencia de la detección de duplicados en bases de datos. Las diferentes propuestas recogen funciones de similitud en términos de emparejamiento de atributos y registros a través de similitud de cadenas (*strings*).

(Herzog *et al.*, 2007) proporcionan una amplia visión general sobre algoritmos probabilísticos para el cálculo de similitud de entrada con un fuerte enfoque en la identificación de personas y direcciones en grandes conjuntos de datos.

En (Bleiholder *et al.*, 2008) la detección de duplicados se trata en el amplio contexto de la integración de información, presentándose como el paso intermedio entre la fusión de datos y el *mapping* de esquemas. En este trabajo se enfatiza en la necesidad de funciones de similitud con propiedades efectivas y eficientes que sean capaces de tratar con grandes conjuntos de datos.

En la **comunidad de Inteligencia Artificial**, el mismo problema es denominado como *database hardening* (Rahm *et al.*, 2000), emparejamiento de nombres (*name matching*) (Bilenko *et al.*, 2003) y emparejamiento de entidades (*entity matching*) (Köpcke *et al.*, (2009). Asimismo, los nombres de resolución de correferencia (*coreference resolution*), incertidumbre de identidad (*identity uncertainty*), detección de duplicados (*duplícate detection*) se utilizan para referirse a esta misma tarea. Más recientemente, dentro de la Web de los Datos (*Web of Data*)[36] se refiere este problema como reconciliación de referencias (*reference reconciliation*) (Dong *et al.*, 2005) y correferencia (co-reference) (Jaffri *et al*., 2008; Glaser *et al.*, 2009).

La explosión en el número de fuentes de información expuestas en la Web de los Datos, como datos RDF, ha derivado también en la explosión en el número de URI (*Uniform Resource Identifiers*) utilizadas para identificar diferentes recursos de no información[37], tales como personas, lugares o cosas abstractas. Por tanto, resulta frecuente encontrar diferentes repositorios de datos con información de recursos idénticos (Jaffri *et al*., 2007). Este problema, según Jaffri *et al*. (2008), puede ocurrir en dos formas:

[36] *Web of Data.* http://www.w3.org/DesignIssues/LinkedData.html

[37] *Non-information resources*. Para más detalles sobre este concepto revisar el siguiente tutorial. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

- Por un lado, cuando una única URI identifica más de un recurso, por ejemplo, existen diversos recursos (topónimos) con el mismo nombre pero estos recursos (topónimos) son localizados en diferentes lugares.
- Por otro, cuando múltiples URI identifican el mismo recurso (por ejemplo; España) y tiene diferentes URI que depende de la fuente de procedencia. Así, en DBpedia aparece como http://dbpedia.org/resource/Spain, mientras en GeoNames tiene la siguiente URI "http://sws.geonames.org/2510769".

Este hecho está requiriendo afrontar problemas estrechamente relacionados con aquellos presentes en la integración de información tradicional.

Por otro lado, fuera de estas tres grandes comunidades (estadística, bases de datos e inteligencia artificial) existen una serie de métodos de detección de duplicados propuestos para un amplio rango de dominios de aplicación, incluyendo, por ejemplo: data warehouses (Ananthakrishna *et al*., 2002) sistemas de gestión de flujo de datos (*data stream*) (Deng *et al*., 2006), redes de sensores (Jefferey *et al*., 2006), datos XML (Weis *et al*., 2005) y *ontology-matching* (Kalfoglou *et al.*, 2003; Noy, 2004; Choi *et al.*, 2006), aunque sólo una pequeña parte de las propuestas de *ontology-matching* (por ejemplo: Castano *et al.*, 2008; Qin *et al.*, 2007; Isaac *et al.*, 2007; Castano *et al.*, 2008) se centran en el ámbito de las instancias. Cada acercamiento persigue diferentes objetivos y, con frecuencia, por naturaleza siendo específicas de un determinado dominio. Asimismo, existen varias herramientas comerciales especializadas en la detección de duplicados, como por ejemplo: FUZZY![38], IBM *Entity Analytic Solutions* (EAS)[39] y *Trillium software system*[40].

3.3.2. Co-referenica en la Web de los Datos

En esta sección, se describen, de forma no exhaustiva, algunos de los principales acercamientos que tratan el problema de la correferencia en el contexto de la Web de los Datos.

En (Jaffri *et al*., 2007; Glaser *et al.*, 2009) se proporciona una visión general de la especificación, implementación, interacción y experiencias de la utilización del servicio de consistencia de referencias (*Consistent Reference Service* – CRS). Este servicio se utiliza para facilitar una rigurosa gestión de la identidad de las URI y su co-referencia. Asimismo, este servicio permite interactuar entre múltiples fuentes de la nube de *Linked Data*[41]. En este trabajo también se recogen las principales diferencias entre la forma de gestionar múltiples URI para un mismo recurso de CRS y la práctica emergente de utilizar la propiedad *owl:sameAs* para identificar URI duplicadas.

(Nikolov *et al.*, 2008) presentan un acercamiento que utiliza los mecanismos de propagación de creencias *Dempster-Shafer* para mejorar la calidad de la integración de datos, concretamente, la correferencia de individuos. Los autores utilizan esta técnica para refinar los resultados de la resolución de la correferencia producidos por las técnicas de similitud de *strings*.

En (Salvadores *et al.*, 2009) se presenta LinksB2N, un algoritmo para descubrir superposición de información en repositorios de datos RDF y realizar la integración de datos sin intervención humana sobre conjuntos de datos que parcialmente comparte el mismo dominio. LinksB2N identifica los recursos RDF equivalentes de diferentes conjuntos de datos con diferente grado de confianza. Este algoritmo se basa en un enfoque que utiliza técnicas de agrupación (*clustering*) para analizar la distribución de los objetos únicos que contienen información superpuesta en grafos de datos diferentes.

[38] *FUZZY!*. http://www.fazi.de/
[39] *EAS*. http://www-306.ibm.com/software/data/db2/eas/
[40] *Trillium*. http://www.trilliumsoftware.com/de/content/products/index.asp
[41] http://linkeddata.org/

Silk – Link Discovery Framework (Volz *et al.*, 2009) es una herramienta para encontrar relaciones entre entidades dentro de diferentes fuentes de datos. *Silk* cuenta con un lenguaje declarativo para especificar qué tipos de links RDF se descubrirán entre las fuentes de datos, así como las condiciones que las entidades deben cumplir para ser interrelacionadas. Las condiciones para establecer relaciones pueden basarse en diferentes métricas de similitud y pueden considerar el grafo en torno a la entidad. Además, *Silk* accede a las fuentes de datos a través del protocolo SPARQL, pudiendo ser utilizado sin tener que replicar los conjuntos de datos localmente. Una descripción detallada de este *framework* aparece en su página web[42].

LIMES - *Link Discovery Framework for Metric Spaces* (Cyrille *et al.*, 2010) implementa una propuesta rápida y sin pérdidas para el descubrimiento de enlaces a gran escala. Esta propuesta está basada en las características de los espacios métricos, pero proporciona un lenguaje menos expresivo para especificar las condiciones de enlace. Una descripción detallada de LIMES aparece en la página web del proyecto[43].

3.3.3. Detección de duplicados en la información geoespacial

Centrándose en el contexto de la información geoespacial, la detección de duplicados es vista como una parte de la integración de información combinando múltiples fuentes heterogéneas. A continuación, se recogen de forma no exhaustiva algunas de los trabajos en este área.

En (Schwering, 2008) se presenta un *survey* centrado en la similitud de datos geoespaciales, enfatizándose la importancia de modelos de representación espacial diferente para interpretar la similitud semántica que en otros trabajos se incluye bajo el término de información de contexto geoespacial (por ejemplo, Baumgartner *et al.*, 2009).

En (Sehgal *et al.*, 2006) se introduce un método para resolver la localización de entidades en datos geoespaciales utilizando elementos espaciales y no espaciales a través de conjuntos de datos. El acercamiento propuesto en este trabajo aprende a combinar diferentes elementos (nombre, coordenadas espaciales, tipos de localización e información demográfica) para realizar una resolución precisa. Además, en este trabajo se presentan los resultados experimentales de la combinación de elementos espaciales y no espaciales (por ejemplo, localización-nombre, localización-tipo, etc.).

(Beeri *et al.*, 2004) presentan algoritmos que consideran únicamente la localización de los objetos y encuentran conjuntos de objetos para su fusión con elevados valores de las medidas de *precision* y *recall*. Además, en este trabajo se desarrolla un nuevo acercamiento basado en un modelo probabilístico y diferentes algoritmos que utilizan este modelo. La mejora del método de pesos normalizados (*normalized-weights method*) logra los mejores resultados bajo todas las circunstancias consideradas en este trabajo. Este método combina el modelo probabilístico con un algoritmo de normalización e información sobre el solapamiento entre conjuntos de datos.

En (Bakillah *et al.*, 2006) se propone un modelo de similitud semántica basado en ontolologías. Este modelo propuesto se utiliza para establecer *mappings* semánticos entre cubos de datos geoespaciales. El modelo propuesto integra varios tipos de componentes de similitud adaptados para diferentes niveles jerárquicos de dimensiones en bases de datos multidimensionales y también integra similitud entre elementos de conceptos. Este trabajo ha sido aplicado a un conjunto de especificaciones de diferentes inventarios de bosques en Canadá.

(Rodríguez *et al.*, 2008) propone utilizar la semántica para "reparar" bases de datos espaciales y restricciones de integridad espacial. Ésta semántica de reparación define un conjunto de instancias de bases de datos que restablecen la consistencia mientras permanezcan vinculadas

[42] http://www4.wiwiss.fu-berlin.de/bizer/silk/
[43] http://aksw.org/Projects/limes

con la instancia original. Asimismo, en este trabajo se propone la noción de respuesta consistente a una consulta conjuntiva espacial. Esto introduce la idea de tolerancia inconsistente en el dominio espacial, cambiando el objetivo de la consistencia de una base de datos espacial a la consistencia de respuestas.

Las propuestas presentadas, tanto en el contexto de la Web de los Datos como en la información geográfica, ponen de manifiesto la ausencia de una combinación de enfoques, es decir, no se trata la detección de duplicados mediante la consideración de las características de la información geoespacial en la Web de los Datos. Entre las propuestas descritas, las pertenecientes al ámbito de la información geográfica se centran en la localización y, mayoritariamente, en bases de datos espaciales. Mientras, los enfoques en el contexto de la Web de los Datos se centran en métricas de similitud entre cadenas (*strings*). En este contexto, únicamente, el *framework Silk* considera entre sus métricas un aspecto propio de la información geográfica (distancia entre dos puntos).

3.4. Linked Data

La Web de Datos enlazados, del inglés ***Web of Linked Data***, supone un nuevo paradigma que pretende explotar la Web como un espacio global de información en el que la navegación se realiza a través de datos estructurados enlazados (*Linked Data*) en lugar de realizarse a través de documentos. Se pasa así de una Web basada en documentos HTML (*HyperText Markup Language*), en la que el usuario es el destinatario de la información publicada, a una Web de Datos Enlazados que están expresados en RDF, en la que sistemas y agentes software pueden explotar estos datos de forma automática (recopilándolos, agregándolos, interpretándolos, publicándolos, etc.) utilizando vocabularios consensuados y ontologías.

3.4.1. Linking Open Data

La iniciativa *Linking Open Data* (LOD) tiene como objetivo ofrecer conjuntos de datos abiertos en la Web y establecer enlaces entre ellos. De esta manera, los pilares de esta iniciativa son:

- La utilización de RDF para publicar la información estructurada en la Web.
- El uso de enlaces (relaciones RDF) para establecer relaciones entre las diferentes fuentes de datos.

Según Bizer *et al*. (2008) esta Web de datos podrá ser explorada por medio de navegadores de datos enlazados, de la misma manera que se explora la Web de documentos por medio de navegadores HTML. La diferencia es que en lugar de seguir enlaces entre páginas HTML, los navegadores de datos enlazados permitirán a los usuarios navegar entre distintas fuentes de datos siguiendo enlaces RDF. Esta Web también podrá ser recorrida por los motores de búsqueda como lo hacen los actuales, con la diferencia de que las capacidades de búsqueda serán mucho mayores, incluso similares a las que proveen las bases de datos relacionales, debido a que los resultados no son sólo páginas, sino datos estructurados.

En la actualidad se están creando diversas aplicaciones para la Web de los Datos (Bizer *et al*., 2008) como navegadores (por ejemplo: Tabulator[44], Disco[45], OpenLink[46] o Zitgist[47]),

[44] http://www.w3.org/2005/ajar/tab
[45] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/
[46] http://demo.openlinksw.com/DAV/JS/rdfbrowser/index.html
[47] http://dataviewer.zitgist.com/

crawlers o motores de búsqueda (Falcons[48], Sindice[49], Sig.ma[50], Swoogle[51] o Watson[52]) entre otras.

3.4.2. Los principios de la Web de Datos Enlazados

El término *Linked Data*, propuesto por Tim Berners-Lee, se refiere a las prácticas recomendadas para exponer, compartir y enlazar datos RDF a través de URI en la Web Semántica (Bizer *et al*., 2008).

Los cuatro principios de diseño (Berners-Lee, 2006) en los que se basa la Web de Datos Enlazados son:

1. Utilizar URI (*Uniform Resource Identifier*) como nombres únicos para los recursos
2. Utilizar el protocolo HTTP para nombrar y resolver la ubicación de los datos identificados mediante esas URI
3. Representar los datos en RDF y utilizar SPARQL[53] como lenguaje de consulta de dichos datos
4. Incluir enlaces a otras URI para permitir la localización de más datos enlazados

Técnicamente, *Linked Data* se trata del uso del lenguaje RDF y del protocolo HTTP para publicar datos estructurados en la Web y conectar datos entre diferentes fuentes de datos, permitiendo que los datos de una fuente sean conectados a los datos en otra fuente (Bizer *et al.*, 2008). Esta manera de publicar los datos permite que sean legibles por las máquinas y los significados sean explícitamente definidos (Bizer *et al.*, 2009).

Entre las diferentes reglas y buenas prácticas para la publicación de datos enlazados en la Web (Bizer *et al*., 2007; Ayers *et al*., 2008) se recoge una serie de recomendaciones. Entre ellas se destacan:

- La asignación de una URI para la identificación de un recurso debe contener nombres apropiados para que otros usuarios puedan enlazar de una manera sencilla y con confianza sus recursos al recurso en cuestión.
- Se debe utilizar HTTP URI debido a que el esquema http:// es el único esquema soportado por las herramientas y la infraestructura actual de la Web. El espacio de nombres (*namespace*) HTTP utilizado en la URI debe estar bajo el control del usuario, es decir, que en la medida de lo posible se debe usar un dominio propio.
- Para un mismo recurso de información es recomendable la definición de tres URI. Una URI para el recurso como tal, una URI para un recurso de información relacionada que sea visualizado por los navegadores HTTP actuales y una URI para un recurso de información relacionada en RDF que pueda ser visualizado por los navegadores RDF.
- Para facilitar el procesamiento de la información se aconseja la reutilización de vocabularios conocidos y extendidos en la Web Semántica. Entre estos vocabularios destacan: Friend Of A Friend[54] (FOAF) para la descripción de personas y sus relaciones, Dublin Core[55] (DC) para describir propiedades genéricas de recursos, SIOC[56] para describir comunidades on-line, VoID[57] para describir conjuntos de

[48] http://iws.seu.edu.cn/services/falcons/objectsearch/index.jsp
[49] http://sindice.com/
[50] http://sig.ma
[51] http://swoogle.umbc.edu/
[52] http://watson.kmi.open.ac.uk/WatsonWUI/
[53] http://www.w3.org/TR/rdf-sparql-query/
[54] http://xmlns.com/foaf/spec/
[55] http://dublincore.org/documents/dcmes-xml/
[56] http://sioc-project.org/

datos enlazados o Creative Commons[58](CC) para la descripción de términos de licencia. Una lista más extensa de los vocabularios más usados se encuentra en la web de la iniciativa Linking Open Data[59].

Más detalles sobre estas y otras reglas y las mejores prácticas emergentes de la comunidad *Linked Data* para la publicación de datos enlazados en la Web se describen en (Berners-Lee, T., 2006, Ayers *et al*., 2008; Bizer *et al*., 2007).

3.4.3. Linked Data geoespacial

La cantidad de datos enlazados publicados en la Web de los Datos ha experimentado un enorme crecimiento en los últimos años. La lista de recursos ya disponibles en *Linked Data* crece día a día[60]. El mayor auge hasta ahora se ha producido en el contexto de la publicación de datos del sector público, siendo referencia los gobiernos del Reino Unido[61] y de Estados Unidos[62]. Asimismo, el fenómeno de *Linked Data* se está extendiendo a otros sectores, entre los que destacan los medios de comunicación, infraestructuras y logística, el ámbito universitario y científico. En este contexto, el hecho de que países de habla inglesa sean los que han tomado la iniciativa origina que la gran mayoría de datos disponibles en *Linked Data* estén en inglés. Una visión general del conjunto de datos (95 *datasets*) correspondientes a la propuesta del *Open Linking Data Project* se muestra en la Figura 3.9.

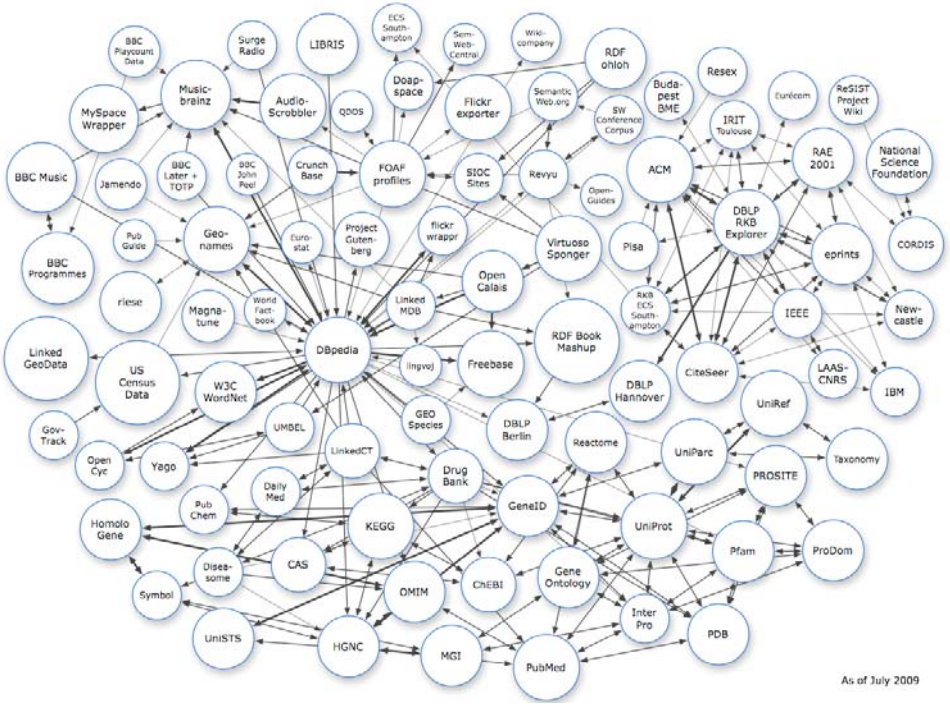


Figura 3.9. Conjunto de datos pertenecientes a *Linking Open Data* (Julio 2009)[63]

En el ámbito de los datos geográficos se han iniciado diferentes esfuerzos en esta dirección. Una de las primeras iniciativas es la propuesta del concepto ***datos geo-enlazados*** (OGC, 2004). Este concepto está vinculado a los datos relacionados geográficamente donde la geometría no está

[57] http://semanticweb.org/wiki/VoiD
[58] http://creativecommons.org/ns
[59] ttp://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/CommonVocabularies
[60] http://linkeddata.org/data-sets
[61] http://data.gov.uk/
[62] http://www.data.gov/
[63] http://richard.cyganiak.de/2007/10/lod/imagemap.html

directamente almacenada con los atributos de los datos. En este caso se utiliza un identificador geográfico que se refiere a un elemento geométrico en un conjunto de datos geoespaciales separado.

Asimismo, diversas fuentes de datos con conocimiento geográfico han comenzando a surgir dentro del contexto de la iniciativa Linking Open Data. Desde un punto de vista genérico, un buen exponente es DBpedia.

- **DBpedia**. Recoge el esfuerzo de una comunidad por extraer información estructurada de Wikipedia, relacionando esta información con otros conjuntos de datos y haciéndola disponible en la web. Entre la información que recoge DBpedia aparece un importante volumen de información con carácter geográfico.

Por otro lado, recientemente han comenzado a surgir diversas fuentes estrechamente vinculadas al dominio geográfico. Estos son los casos de GeoNames, LinkedGeoData y Ordnance Survey.

- **GeoNames**. Recoge información de datos geográficos, tales como nombres de lugares, poblaciones, etc. de diversas fuentes[64]. Asimismo, los usuarios pueden, de forma manual, editar, corregir y añadir nuevos topónimos.
- **LinkedGeoData**. Esta propuesta utiliza la información recopilada por el proyecto OpenStreetMap[65] y la hace disponible como una base de conocimientos de RDF de acuerdo a los principios de *Linked Data*. Además, esta información se enlaza con los datos de DBpedia (Auer *et al.*, 2009).
- **Ordnance Survey**. Proporciona la información sobre las unidades administrativas y distritos censales (*voting regions*) de Gran Bretaña en formato *Linked Data*. Estos datos incluyen los nombres, código censal y área de las diferentes regiones británicas. Las características de este trabajo se describe en (Goodwin *et al*., 2009). En (Correndo *et al.*, 2010; Omitola *et al.,* 2010) se recoge el proceso de publicación de la información del sector público del Reino Unido utilizando como soporte la información publicada por el Ordnance Survey.

La comparación de estas propuestas, recogida en la Tabla 3.3, demuestra que, hasta la fecha, únicamente se utiliza el elemento puntual (representados por un par de coordenadas (latitud, longitud) para la representación de la información geográfica en el conjunto de datos geoespaciales presentes en la nube de *Linked Data*.

Asimismo, se observa que uno de los *triple store* más utilizados es Virtuoso y que se usan diversas herramientas para la generación de RDF (Triplify[66] y D2R[67]). Con respecto a las ontologías utilizadas, se puede comprobar que éstas son diversas, aunque como denominador común todas las propuestas utilizan el vocabulario de WGS84 Geo Positioning[68]. Finalmente, los dominios de la información publicada también son diferentes (unidades administrativas, puntos de interés, topónimos e información general) y el idioma inglés es predominante, aunque DBpedia y GeoNames proporcionan diferentes idiomas, por ejemplo: español, italiano, francés, etc. En la Tabla 3.3 los idiomas aparecen codificados utilizando ISO 639-1 (ISO 639-1, 2002).

[64] Un listado detallado de las fuentes utilizadas por GeoNames aparece en http://www.geonames.org/data-sources.html
[65] http://www.openstreetmap.org/
[66] http://triplify.org/Overview
[67] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
[68] http://www.w3.org/2003/01/geo/wgs84_pos

Tabla 3.3. Características de los diferentes *Linked Data* geoespaciales

Proveedor	Origen de los datos	*Triple store*	Información	Fuentes	generación RDF	Patrón URI	Vocabulario	Geometría	Idioma
Ordnance survey	Ordnance survey	Talis	Unidades Administrativas	Un nomenclátor (1:50,000) y un nomenclátor administrativo para Gran Bretaña	-	http://data.ordnancesurvey.co.uk/id/.+	Spatial Relations ontology, Administrative Geography ontology, WGS84 Geo Positioning, FOAF, OWL, y Gazetteer ontology	Punto	en
Linked GeoData	OpenStreetMap	Virtuoso	Puntos de interés	Base de datos OpenStreetMap (paradas de autobús, señales de tráfico, correos)	Triplify	http://linkedgeodata.org/triplify/+	LGD ontology, WGS84 Geo Positioning	Punto	en
GeoNames	GeoNames	RDF dump	Topónimos	Fuentes de información utilizadas por GeoNames	-	http://sws.geonames.org/id/	GeoNames ontology, WGS84 Geo Positioning	Punto	en, es, de, ca, nb, it, da, fr.
DBpedia	Wikipedia	Virtuoso	Información general	Wikipedia	D2R Server	http://dbpedia.org/resource/+	DBpedia ontology, WGS84 Geo Positioning	Punto	92 lang.

La transformación y publicación de estas diversas fuentes de información geoespacial conforme a los principios de *Linked Data* añade una nueva dimensión a la Web de los Datos. De esta manera, la información geoespacial puede ser recuperada y enlazada a niveles de granularidad sin precedentes.

3.5. Conclusiones

En este capítulo se han tratado algunas nociones sobre las geo-ontologías y se han descrito y comparado algunas propuestas de recursos ontológicos geoespaciales (sección 3.1). Esta sección permite comprobar que a pesar de la existencia de gran diversidad de fuentes con conocimiento geoespacial **no son frecuentes los recursos ontológicos con conceptualizaciones globales de la realidad de los fenómenos geográficos**. Por tanto, estos recursos no pueden ser utilizados como recursos de referencia en la comunidad geográfica para conseguir la interoperabilidad semántica de la información. En el dominio de la información hidrográfica, los diferentes recursos ontológicos se circunscriben principalmente al dominio de la hidrología y se relacionan con la ontología *upper-level* SWEET. Por otro lado, aunque en menor número, también se describen algunas ontologías del dominio hidrográfico o con información vinculada a este dominio. **Estas ontologías recogen una visión particular y monolingüe** (mayoritariamente en inglés) **del dominio, puesto que son producto de la transformación de determinadas fuentes de conocimiento**. Este hecho pone de manifiesto la **ausencia de una ontología global y multilingüe de fenómenos hidrográficos que recoja el conocimiento sobre los fenómenos propios del contexto nacional español**.

Además, se ha comprobado que existen múltiples propuestas para la integración de información en sus diversos enfoques: sintáctica, estructural y semántica (sección 3.2). Centrándose en el marco de la integración semántica, derivado del objetivo de este trabajo, se han destacado los principales enfoques propuestos basados en ontologías y se mencionan algunas propuestas de este tipo en el contexto de la información geoespacial. Sin embargo, estas propuestas **no resultan fácilmente replicables, no es fácil conocer sus características y no se pueden llevar a cabo, de manera óptima, las tareas de planificación necesarias**. Por otra parte, a pesar de estas múltiples propuestas, hasta el momento, **no se han desarrollado metodologías que se puedan aplicar en este área de integración de información**. Además, a pesar de los diversos enfoques de integración basados en ontologías existentes dentro del dominio geoespacial, **no existe en la bibliografía del área una caracterización detallada de los problemas en el ámbito de las instancias durante el proceso de integración de información de múltiples fuentes de datos geoespaciales**.
Asimismo, a pesar de la existencia de múltiples lenguajes y herramientas para la transformación de bases de datos relacionales a RDF se ha detectado una **ausencia de propuestas que permitan tratar y transformar, de manera específica, la información geométrica recogida en las bases de datos geoespaciales a RDF**.

Con respecto a la detección de duplicados, la sección 3.3 recoge la existencia de diversos estudios que recopilan distintas técnicas y métodos. Estos trabajos, fundamentalmente, aparecen en la comunidad de bases de datos. Sin embargo, **no existen trabajos que recopilen, de forma exhaustiva, los métodos de detección de duplicados considerando las características de la información geoespacial**.

En el contexto de las Web de los Datos (sección 3.3.2) aparecen algunos trabajos que desarrollan un análisis sistemático de la co-referencia en la información y que presentan diversos enfoques para solventar éste problema. No obstante, resultan escasos los trabajos centrados en el estudio de la detección de duplicados durante el proceso de integración de múltiples conjuntos de datos geoespacial. Asimismo, se pone de relieve que **son realmente escasas y limitadas las propuestas de detección de duplicados entre conjuntos de datos RDF que consideren las características propias de la información geográfica**.

Finalmente, entre las propuestas de detección de duplicados en el contexto geoespacial (sección 3.3.3) existen enfoques que consideran información espacial y no espacial. Sin embargo, **resultan escasos los trabajos cuyo enfoque esté basado en ontologías**, así como tampoco **existe ningún enfoque que recoja explícitamente información temática propia de las diferentes realidades de los dominios que componen la información geográfica**.

La sección 3.4 presenta la iniciativa *Linking Open Data,* recoge sus principios y algunas de sus principales reglas y buenas prácticas. Asimismo, describe diferentes propuestas que publican información geoespacial conforme a los principios de *Linked Data*. Sin embargo, **no existe una iniciativa abierta destinada al enriquecimiento de la Web de Datos con datos geoespaciales del territorio nacional español**. Asimismo, el hecho de que países de habla inglesa sean los que han tomado la iniciativa origina que la gran mayoría de datos disponibles en la nube de *Linked Data* estén en inglés. Esto evidencia la **pobre presencia en *Linked Data* de datos en español**.

Por otro lado, entre las propuestas destinadas a la publicación de datos geoespaciales únicamente se representa la posición espacial de los fenómenos a través de elementos geométricos de tipo puntual (latitud y longitud). Por tanto, **no se realiza tratamiento de las características de las geométricas complejas vinculadas a este tipo de información**.

Capítulo 4

Planteamiento

En este capítulo se describen los objetivos de este trabajo y se ofrece una visión genérica de las principales aportaciones al estado de la cuestión, relacionándolas con los problemas de investigación abiertos identificados en la sección 3.5. Por último, se establecen los supuestos, hipótesis y limitaciones de este trabajo.

4.1. Objetivos

El objetivo general de esta tesis es **integrar información de bases de datos heterogéneas en el contexto de la información geográfica**. Para lograr este objetivo y proponer soluciones para superar las limitaciones identificadas en el estado de la cuestión se proponen una serie de objetivos específicos, divididos en objetivos científicos y tecnológicos:

Objetivos científicos

O1. Proponer una **metodología para la integración, basada en ontologías, de información de bases de datos heterogéneas en el dominio geográfico**.
O2. **Identificar** y **armonizar** los diferentes **tipos de heterogeneidad** existentes en diferentes contextos de la información geográfica e **interrelacionarlos con los principales problemas** de esta información.
O3. Refinar los problemas existentes en el contexto de la **integración de instancias procedentes de bases de datos geoespaciales heterogéneas**.

Objetivos tecnológicos

O4. **Integrar información de bases de datos heterogéneas** para proporcionar una visión armonizada de la realidad recogida en los datos geoespaciales.
O5. **Publicar información geoespacial** del territorio nacional español **conforme a los requerimientos de *Linked Data***.
O6. Proponer **métodos para la detección de instancias duplicadas** en conjuntos de datos heterogéneos.

4.2. Contribuciones previstas al estado de la cuestión

En este trabajo se pretende dar solución a los problemas identificados en la sección 3.5. En este sentido, esta tesis aportará avances al estado de la cuestión mediante resultados evaluables conforme a los objetivos planteados con anterioridad.

Con respecto al primer objetivo identificado, los avances propuestos son;

C1. **Identificación de las fases, actividades y tareas** a realizar en la **propuesta metodológica** para la integración de información basada en ontologías de bases de datos heterogéneas. Asimismo, se identificarán los participantes, técnicas a utilizar y las entradas y salidas asociadas a cada tarea.

Conforme al objetivo de identificación y armonización de las diferentes clasificaciones de problemas existentes en el contexto de la integración de información geoespacial se presenta el siguiente avance:

C2. **Recopilación exhaustiva de las diferentes clasificaciones de tipos de heterogeneidad vinculadas con la información geográfica**. Esta clasificación recoge los problemas presentes en el ámbito de las bases de datos, la información geoespacial y los modelos de conocimiento.

C3. Un **refinamiento de los problemas de integración** que se pueden detectar en bases de datos geoespaciales heterogéneas en el contexto de la información hidrográfica. Estos problemas son abstraídos de forma general para el dominio de la información geográfica y, en ciertos casos, manifiestamente dependientes del dominio hidrográfico. La clasificación de estos problemas sigue un enfoque basado en capas (léxica, sintáctica, semántica y pragmática) que aparece en diferentes trabajos sobre interoperabilidad semántica (Euzenat, 2001; Corcho, 2005; Barrasa, 2007).

Con respecto al objetivo referido a la integración de información de bases de datos heterogéneas con la finalidad de proporcionar una visión armonizada de la realidad este trabajo propone el siguiente avance:

C4. **Creación de recursos ontológicos de dominio y multilingües en el área de la información hidrográfica**. Estos recursos están centrados en fenómenos representables cartográficamente, con objeto de unificar las diversas fuentes y modelos existentes en la comunidad geográfica a diferentes niveles, tales como los propuestos por Alexandria Digital Library Feature Type Thesaurus, Glosario Hidrológico Internacional de la UNESCO, GeoNames, EuroGlobalMap, Directiva Marco del Agua, Base Topográfica Numérica - BTN25, etc.

C5. **Generación de siete conjuntos de datos RDF**, a través de la creación de *wrappers* entre los recursos ontológicos creados y las diversas fuentes de información (bases de datos).

En cuanto al objetivo de publicación de información geoespacial del territorio nacional español conforme a los requerimientos de *Linked Data* este trabajo realizará el siguiente avance:

C6. Publicación de información hidrográfica y enriquecimiento de la Web de los Datos con información geoespacial en español.

C7. Tratamiento e inclusión de las características geométricas de la información geoespacial a través de la incorporación de *Geography Markup Language* (GML) en el RDF generado.

Finalmente, respecto al objetivo referido a la proposición de métodos para la detección y fusión de instancias duplicadas en conjuntos de datos heterogéneos este trabajo plantea el siguiente avance:

C8. **Proposición de heurísticas para la detección de instancias duplicadas en conjuntos de datos RDF heterogéneos y establecimiento de relaciones** (*owl:sameAs*) **entre dichas instancias**. Las heurísticas desarrolladas, de forma individual, serán

consideradas independientes y dependientes del dominio geográfico y estarán clasificadas en distintas capas (léxica, sintáctica, semántica y pragmática), conformes con el enfoque mencionado con anterioridad. Asimismo, desde una visión incremental se clasificarán en dos grandes grupos: heurísticas de localización y temáticas.

4.3. Premisas, hipótesis y limitaciones de la solución propuesta

El trabajo descrito en este documento está basado en un conjunto de premisas y limitaciones que contribuyen a contextualizar el trabajo realizado. Las principales aportaciones se resumen mediante una serie de hipótesis de partida, cuya verificación se realiza en el Capítulo 9 (sección 9.5) de este trabajo.

4.3.1. Premisas

En el caso de las **premisas**, éstas ayudan a explicar las decisiones tomadas para el desarrollo de las soluciones adoptadas con la finalidad de alcanzar los objetivos propuestos en este trabajo.

P1. La heterogeneidad de las fuentes de información consecuencia de la gran diversidad de productores, intereses y necesidades de información geoespacial conlleva disparidad de tipos de fenómenos. A esto se une el hecho de que muchas de las fuentes de información no cuentan con definiciones clarificadoras de las consideraciones semánticas de cada fenómeno hidrográfico o presentan manifiestas diferencias semánticas para el mismo tipo de fenómeno.

P2. El factor escala actúa como filtro, tanto en la representación gráfica de la realidad geográfica como en los tipos de fenómenos a considerar en las diversas fuentes de información (catálogos de fenómenos, diccionarios de datos, tesauros, nomenclátores, etc.)

P3. La presencia de un mismo fenómeno en diferentes idiomas en los sistemas de información actuales hace que sea considerado como un conjunto de fenómenos distintos (sólo por tener otro nombre en otro idioma), consecuencia del tratamiento sintáctico de la información.

P4. Cada una de las bases de datos con las que se evaluará este trabajo presenta distintos tipos de fenómenos, granularidad, estructura (organización interna), relaciones, claves y formatos.

P5. La existencia de duplicidades en las diferentes fuentes de información normalmente se deriva, entre otros muchos factores, de: (a) la implementación de topología sobre la representación geométrica de la realidad, (b) coincidencia de la información consecuencia del tratamiento a diferentes escalas, (c) cuestiones propias del modelo de implementación de base de datos, etc.

P6. Se utiliza el concepto instancia, dentro del dominio geográfico, para referirse a un individuo u ocurrencia que representa un fenómeno concreto del mundo real asociado a la superficie de la Tierra. Además, cada individuo tiene una representación geométrica y un conjunto de características asociadas.

P7. Se utilizan los términos información geográfica e información geoespacial como sinónimos. Ésta relación entre conceptos se establece considerando que el término "Información Geográfica" (IG) es la expresión clásica para referirse a información georreferenciada. Con posterioridad, el *Open Geospatial Consortium* introdujo una distinción alternativa:

- Información Geográfica. Información representada gráficamente (pantalla/papel) en un Sistema de Coordenadas referido a la Tierra.
- Información Geoespacial. Información que no tiene porqué representarse.

4.3.2. Hipótesis

Las **hipótesis** "*son nuestros ojos cuando tratamos de encarar los problemas de manera científica. Por medio de ellas se puede escudriñar el desorden que constituye un problema y considerar las posibilidades de orden*" (Hodnett, 1955). Las hipótesis de partida de este trabajo son:

H1. La propuesta de una metodología para la integración, basada en ontologías, de información de bases de datos heterogéneas en el dominio geográfico genera un acercamiento replicable, transparente, finito, basado en prácticas existentes, independiente de herramientas y conforme con diversos principios INSPIRE.

H2. El modelado, formalización en lógica de descripciones y la implementación en OWL de *hydrOntology* contribuye a la construcción de un marco de referencia común y compartido. Este marco ayudará a solventar los problemas de estructuración y heterogeneidad presentes en la información geográfica del dominio hidrográfico.

H3. El proceso de generación de conjuntos de datos RDF (bases de conocimiento) conlleva la solución de diversos problemas de las bases de datos originales y una primera depuración de instancias duplicadas.

H4. Es posible tratar la información de la geometría asociada a un fenómeno geográfico e incorporar esta información en tripletas RDF.

H5. La publicación de los diversos conjuntos de datos RDF generados, conforme a los principios de *Linked Data*, enriquecen con información del territorio nacional español la Web de los Datos e incrementan la presencia del español.

H6. Es posible detectar instancias duplicadas de conjuntos de datos RDF procedentes de bases de datos heterogéneas y resolver estas duplicidades mediante la definición y aplicación de heurísticas que combinan aspectos independientes y dependientes del dominio geográfico.

4.3.3. Limitaciones

Finalmente, las **limitaciones** definen los límites de la contribución de este trabajo, por otro lado, permite determinar futuros objetivos de investigación.

L1. A pesar de la miscelánea temática existente en los variados catálogos y diccionarios de datos de los diferentes productores de IG y, más concretamente, dentro del dominio hidrográfico y, por tanto, en sus bases de datos respectivas, este trabajo se centra, única y exclusivamente, en fenómenos puramente hidrográficos, específicamente, en fenómenos continentales. Por tanto, quedan fuera de este estudio fenómenos como "presa", "dique", "molino de agua", "ensenada", etc.

L2. Las bases de datos que se integran en la experimentación se limitan a aquellas procedentes de productores europeos y nacionales de información geoespacial (instituciones tanto a nivel nacional como de Comunidades Autónomas y Confederaciones hidrográficas).

L3. En lo que respecta a las heurísticas, éstas aunque serán desarrolladas desde una óptica generalista serán de aplicación, de manera concreta, al dominio hidrográfico y, de forma

general, al dominio geográfico. Por tanto, resultará excepcional que las heurísticas resultantes puedan ser de aplicación global a cualquier dominio.

L4. Con respecto a las limitaciones tecnológicas, se puede afirmar que este trabajo no se centrará en problemas de planificación para la generación de planes de consulta para el acceso a bases de datos heterogéneas, como se realiza en otros enfoques, ni proporcionará avances significativos en el área de generación de *wrappers*, ya que se reutilizarán tecnologías existentes.

L5.En cuanto a las limitaciones científicas, este trabajo no propondrá nuevos modelos de reescritura de consultas utilizando las correspondencias descritas entre bases de datos y ontologías.

L6. Los recursos ontológicos desarrollados contienen información perteneciente al dominio de la información hidrográfica caracterizada por ser representable cartográficamente. Asimismo, *hydrOntology*, ontología en torno a la que se desarrolla la red de ontologías, considera información en español, catalán, inglés y francés. El resto de recursos ontológicos que conforman dicha red únicamente contienen información en inglés.

L7. La metodología de integración de datos geográficos propuesta en este trabajo está destinada, únicamente, a la integración basada en ontologías de información de bases de datos geográficas heterogéneas.

Capítulo 5

Metodología para la integración basada en ontologías de información de bases de datos heterogéneas en el dominio geográfico

El propósito de esta metodología para la integración de datos geográficos procedentes de bases de datos geoespaciales heterogéneas es el de dar soporte a aquellos desarrolladores/organizaciones que deseen ofrecer una visión integrada de la realidad abstraída y almacenada en diferentes fuentes de información.

Este capítulo presenta una metodología para la integración basada en ontologías de información de bases de datos heterogéneas para el dominio geográfico. En la sección 5.1 se define lo que se entiende por metodología, así como sus conceptos asociados. Los principios de diseño considerados para la definición de esta propuesta metodológica se presentan en la sección 5.2. En la sección 5.3 se describen los procesos seguidos para definir esta metodología y, además, se presenta una visión general de las actividades y tareas propuestas. Los diversos participantes implicados en el proceso de integración se describen en la sección 5.4. La sección 5.5 presenta una descripción de alto nivel del contexto en que se lleva a cabo la instanciación de esta metodología. Asimismo, se detallan las diferentes actividades y tareas propuestas. Finalmente, se recogen las conclusiones a este capítulo en la sección 5.6.

5.1. Definición de metodología

De forma generalizada, los términos metodología, método, técnica, proceso, actividad, etc. son utilizados en la literatura de forma indistinta (Hodge, 2000). Con objeto de clarificar el uso de estos términos y aplicarlos a la temática tratada en este trabajo, se adoptarán las definiciones propuestas por el IEEE (*The Institute of Electrical and Electronics Engineers*).

El IEEE define **metodología** como "*una serie integrada y exhaustiva de técnicas o métodos que crean una teoría general de sistemas de cómo una clase de trabajo intensivo pensado debería ser llevada a cabo*[1]" (IEEE, 1990). Las metodologías son ampliamente utilizadas en Ingeniería del Software (Downs *et al*., 1998; Pressmann, 2000; Kruchten, 2000; MÉTRICA v.3), Ingeniería del Conocimiento (Waterman, 1986; Gómez-Pérez *et al*., 1997; Schreiber *et al*.,

[1] "*A comprehensive, integrated series of techniques or methods creating a general systems theory of how a class of thought-intensive work ought to be performed*" (IEEE, 1990).

1999) o construcción de ontologías (Fernández-López *et al.*, 1999; Staab *et al.*, 2001; Gómez-Pérez *et al.*, 2003; Pinto *et al.*, 2004; Suárez-Figueroa *et al.*, 2008). La definición previa de metodología menciona que métodos y técnicas son partes de la metodología. Un **método** es un conjunto de "*procesos o procedimientos ordenados utilizados en la ingeniería de un producto o transformación de un servicio*[2]" (IEEE, 1990). Una **técnica** es "*un procedimiento directivo y técnico utilizado para lograr un objetivo dado*[3]" (IEEE, 1990b).

En (de Hoog, 1998) se tratan las relaciones entre metodologías y métodos. Según este autor, las metodologías y métodos no son lo mismo porque "las metodologías se refieren al conocimiento sobre los métodos". Las metodologías establecen "qué", "quién" y "cuándo" una actividad dada debería ser llevada a cabo. En (Greenwood, 1973) también se investiga las diferencias entre métodos y técnicas. Un método es un procedimiento general mientras una técnica es una aplicación específica de un método y la forma en que el método es ejecutado. Normalmente, existen varias técnicas para aplicar un método dado.

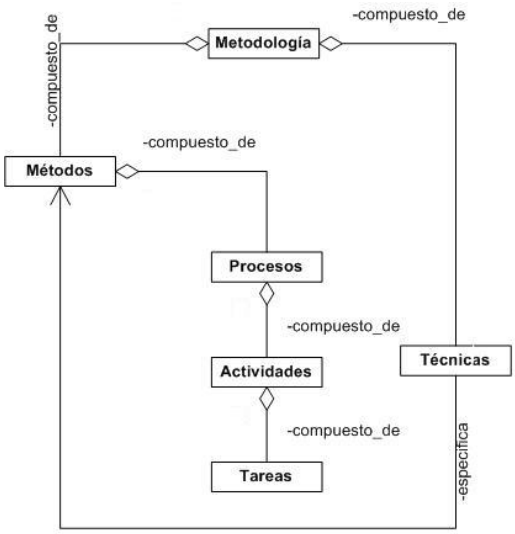


Figura 5.1. Representación gráfica de las relaciones terminológicas en las metodologías (Gómez-Pérez *et al*., 2003).

De la definición de metodología, expuesta anteriormente, se deduce que métodos y técnicas están estrechamente relacionados, consecuencia de que ambos son utilizados para llevar a cabo las tareas presentes en los diferentes procesos de los que consta una metodología. El IEEE define un **proceso** (IEEE, 1995) como "una *función que debe ser llevada a cabo en el ciclo de vida del software. Un proceso está compuesto de actividades*[4]". Una **actividad** (IEEE, 1995) es "*una tarea componente de un proceso*[5]". Otra definición de actividad (IEEE, 1997) afirma que es un cuerpo definido de trabajo para ser efectuado, incluyendo su información de entrada y salida requerida. En definitiva, una tarea es la unidad de trabajo más ínfima sujeta a la responsabilidad de gestión. Así, "*una tarea* (IEEE, 1995) *es un trabajo bien definido asignado a uno o más miembros del proyecto. Las tareas relacionadas, normalmente, son agrupadas para formar actividades*[6]".

[2] "*Orderly processes or procedures used in the engineering of a product or performing a service*" (IEEE, 1990).
[3] "*A technical and managerial procedure used to achieve a given objective*" (IEEE, 1990b).
[4] "*A function that must be performed in the software life cycle. A process is composed of activities*".
[5] "*A constituent task of a process*"
[6] "*A task* (IEEE, 1995) *is a well-defined work assignment for one or more project members. Related tasks are usually grouped to form activities*".

Las relaciones entre las definiciones mencionadas son resumidas en la Figura 5.1, donde se puede comprobar que una metodología está compuesta de métodos y técnicas. Los métodos están compuestos de procesos y son detallados con técnicas. Los procesos están compuestos de actividades. Finalmente, las actividades están formadas por grupos de tareas.

5.2. Principios de diseño

Los principios de diseño aplicados al desarrollo de esta metodología de integración de información fueron extraídos de las condiciones que, según Paradela (2001), cada metodología debería satisfacer. Estas condiciones son clasificadas en condiciones necesarias (también llamadas condiciones formales) y condiciones suficientes (también denominadas condiciones materiales). Las condiciones necesarias son independientes del dominio de aplicación de las metodologías y son comunes a cualquier metodología. Las condiciones suficientes son específicas del dominio al que se aplica la metodología: en este caso el dominio se ciñe a la integración basada en ontologías de información de bases de datos heterogéneas en el dominio geográfico.

Las **condiciones necesarias** de cualquier metodología definidas en (Paradela, 2001) son las siguientes:

- *Completitud*. La metodología debe considerar todos los casos presentados, a pesar de quién presente el caso y su grado de dificultad, tipología, etc.
- *Eficiencia*. La metodología debe ser eficiente, independientemente de la persona que la utilice.
- *Efectividad*. La metodología debe permitir resolver de manera adecuada todos los casos que, teniendo solución, sean presentados.
- *Consistencia*. La metodología debe ofrecer los mismos resultados para los mismos problemas, independientemente de quién lo lleve a cabo.
- *Cuestiones sobre la esencia*. La metodología debe considerar el "qué", "quién", "por qué", "cuándo", "dónde" y "cómo" de cada tarea.
- *Conjunto de reglas metodológicas generales*. La metodología debe considerar las reglas metodológicas generales del Método Cartesiano de Descartes, es decir, *evidencia* (nunca aceptar algo como verdad que no pueda conocer de forma evidente como tal), *análisis* (dividir cada una de los problemas a examen en tantas partes como sea posible y como pueda resultar necesario para su solución más adecuada), *síntesis* (realizar reflexiones comenzando por los objetos más simples y fáciles de conocer y subir poco a poco al conocimiento de los más complejos) y *examen* (hacer enumeraciones y revisiones completas y generales para poder asegurar que nada fue omitido).
- *Finitud*. El número de elementos que componen la metodología y el número de tareas debe ser finito y éstas deberían consumir un corto periodo de tiempo.
- *Discernimiento*. La metodología debe estar compuesta de un número pequeño de componentes estructurales, funcionales y representacionales.
- *Entorno*. La metodología debe ser clasificada en uno de los dos grupos principales: científica o tecnológica.
- *Transparencia*. La metodología debe ser como una caja blanca, permitiendo conocer en cada momento cuál es la tarea que está siendo procesada, qué está siendo llevado a cabo, quién lo está llevando a cabo, etc.

Las **condiciones suficientes** de esta metodología de integración de datos geográficos están fundamentadas en una serie de condiciones específicas propias de los diferentes componentes que conforman este acercamiento metodológico.

- *Depuración o curación de bases de datos*. La metodología debe estar orientada a producir mejoras en las bases de datos implicadas en el proceso de integración, así

como a obtener pautas y lecciones de buenas prácticas tanto en el proceso de construcción de bases de datos como en la integración de fuentes de datos heterogéneas.

- *Dominio*. La metodología se propone para la integración de información de bases de datos heterogéneas aplicada al dominio de la información geográfica. Esta metodología hace un énfasis especial en la detección de duplicados de datos (instancias) en los diferentes conjuntos de datos utilizados para dicho proceso de integración.
- *Usabilidad*. La metodología debe ser fácil de entender y aprender. Además, el esfuerzo necesario para utilizar la metodología debe ser mínimo con el fin de facilitar su éxito y generalizar su uso.
- *Inspirada en prácticas existentes*. La metodología debe estar fundada en diferentes propuestas metodológicas existentes.
- *Independiente de herramientas*. De forma general, una metodología debe ser independiente de las tecnologías existentes.
- *Gran escala*. Esta metodología debe tratar con diferentes tipos de recursos de información (catálogos, tesauros, diccionarios, ontologías, bases de datos, etc.) con objeto de llevar el proceso de integración a diferentes niveles.
- *Flexibilidad*. Esta metodología debe ofrecer flexibilidad para ciertas tareas dependientes del propósito y/o del conocimiento de los expertos.
- *Almacenamiento, disponibilidad y mantenimiento de datos espaciales*. La metodología debe ser consecuente con los principios de INSPIRE. Esto conlleva que las fuentes de información, entre otros aspectos, sean almacenadas y mantenidas al nivel más adecuado, es decir, por sus propios productores.
- *Interoperabilidad*. Esta metodología debe permitir que este concepto no sea únicamente visto como una propiedad computacional, sino que esta propuesta metodológica debe contribuir a la interoperabilidad estructural y semántica.

5.3. Proceso de desarrollo. Descripción de alto nivel

Esta sección describe, desde una perspectiva general, el conjunto de actividades y tareas que componen la metodología de integración de datos geográficos. Este conjunto de actividades se basan en las principales actividades identificadas por el proceso de desarrollo de software (IEEE, 1999).

El proceso de desarrollo de la metodología se conformó a través de los siguientes pasos:

- Análisis de alto nivel de las características que particularizan al dominio de trabajo.
- Selección de procesos relevantes pertenecientes a otras áreas de investigación.
- Identificación de las principales tareas, en los procesos relevantes elegidos, para seleccionar y analizar las tareas utilizadas en la mayoría de procesos de cada área.
- Abstracción y adaptación de las diferentes tareas para cubrir las necesidades del proceso de integración de información basada en ontologías de dominio.
- Análisis de las dependencias de las tareas para definir en consecuencia el orden de las mismas.

Un primer resultado de la aplicación de este proceso de desarrollo es la identificación del conjunto de actividades que componen la metodología de integración. Estas actividades están formadas por: estudio de viabilidad (actividad 1), análisis de fuentes (actividad 2), integración estructural (actividad 3), integración semántica (actividad 4), publicación de datos integrados (actividad 5), depuración / post-procesamiento (actividad 6) y modificaciones sobre fuentes de información (actividad 7). A estas actividades se le une un proceso horizontal de mantenimiento. Dentro de estas actividades, las asociadas al proceso de integración estructural (actividad 3) y a la publicación de datos integrados (actividad 5) son opcionales y su justificación aparece en los apartados 5.5.3 y 5.5.5 respectivamente. Este hecho es el que provoca la diferenciación en el color y en la trama de la línea respecto al resto de actividades.

Las actividades propuestas, como se refleja en la Figura 5.2, tienen influencias de la ingeniería del software, ingeniería de conocimientos, ingeniería ontológica, *ontology matching, Linking Open Data* e ingeniería de datos. Una descripción detallada de las diversas actividades y sus influencias se lleva a cabo en la sección 5.5 de este trabajo.

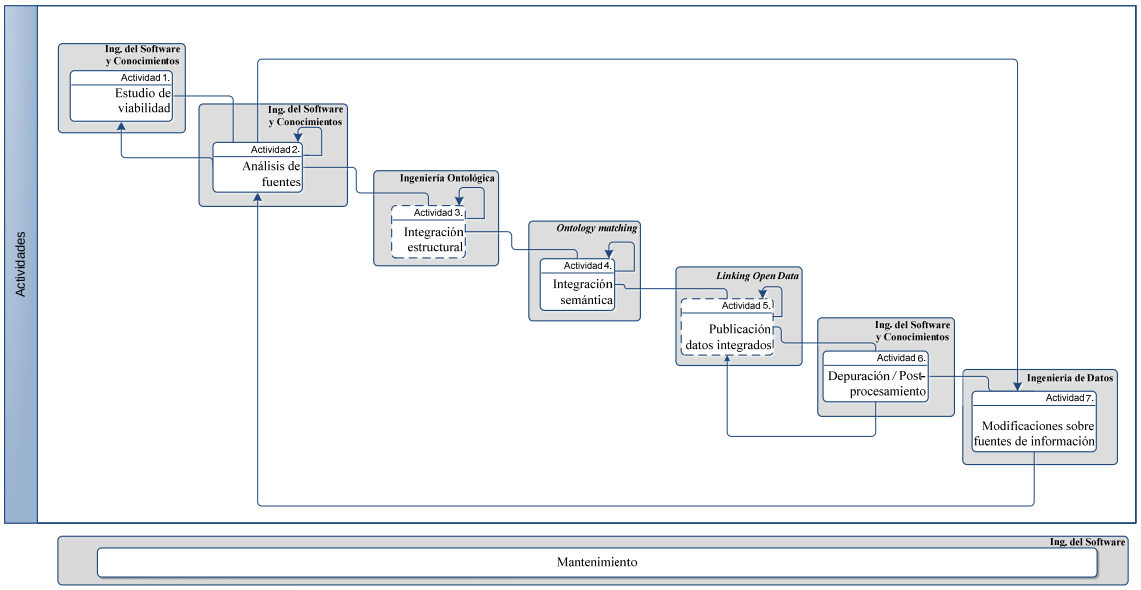


Figura 5.2. Visión general de las diversas actividades de la propuesta metodológica y sus respectivas influencias.

Esta metodología propone el desarrollo de sus actividades a través de un ciclo de vida iterativo. Este ciclo de vida se caracteriza por el desarrollo de iteraciones (es decir, la repetición del conjunto de actividades propuestas en la metodología) y tras cada iteración, con la colaboración del cliente/experto, se evalúa el producto corrigiéndolo o proponiendo mejoras. Al final de cada iteración se le entrega al cliente una versión mejorada o con mayores funcionalidades del producto. Estas iteraciones se repetirán hasta obtener un producto que satisfaga al cliente. De este modo, se proporciona una primera versión de las diferentes actividades y, posteriormente, con la retroalimentación de los expertos/cliente se procede a incrementar el refinamiento de los resultados obtenidos en las mismas. En la Figura 5.3 se muestra el proceso propuesto por este ciclo de vida.

El conjunto de estas actividades se descompone en una serie de tareas reflejadas en la Figura 5.4. A continuación se realiza una enumeración de las tareas que componen cada actividad y se comentan algunas de sus características generales. Una descripción detallada de las diferentes tareas que componen esta metodología se recoge en la sección 5.5 de este trabajo.

La actividad referida al estudio de viabilidad (actividad 1) está compuesta por las siguientes tareas: Delimitar / definir el dominio de trabajo (tarea 1.1), planteamiento de la problemática (tarea 1.2), identificación del objeto (tarea 1.3), identificación de posibles mejoras (tarea 1.4), nivel de formalización del dominio (tarea 1.5), recopilación y descripción de fuentes de información (tarea 1.6), consideración del multilingüismo (tarea 1.7), realización de prototipos (tarea 1.8) y, finalmente, decisión sobre la viabilidad /evaluación (tarea 1.9). Las tareas 1.3 y 1.4, dadas sus características, son o pueden ser llevadas a cabo en procesos paralelos. Una situación similar ocurre con el conjunto de tareas 1.5, 1.6 y 1.7. De estas tres últimas, dependiendo de las características del dominio y de la información contenida en las fuentes de información, las tareas 1.5 y 1.7 pueden ser consideradas opcionales.

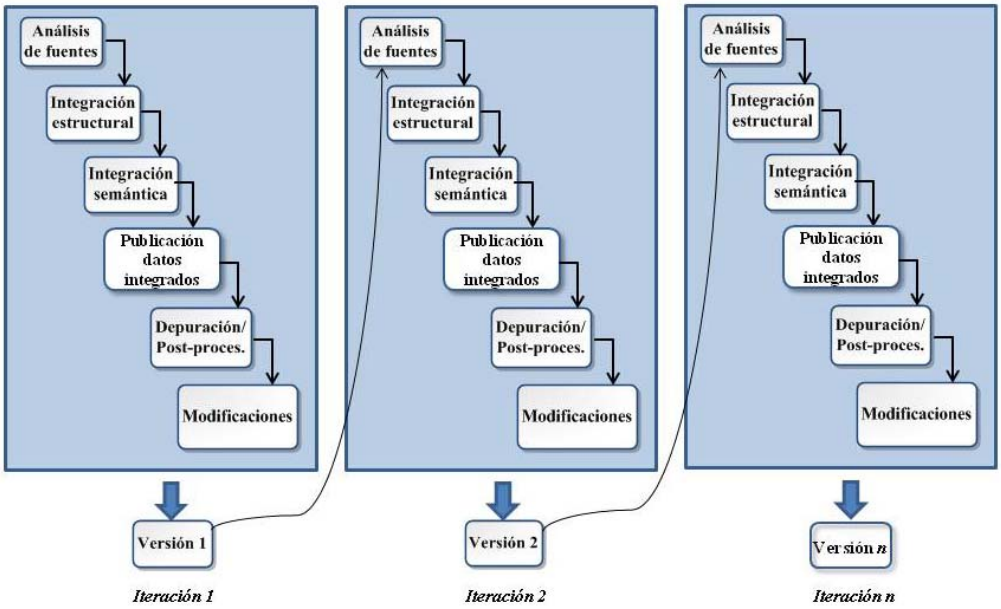


Figura 5.3. Ciclo de vida iterativo

El análisis de fuentes (actividad 2) está formado por dos bloques de tareas. Por un lado, aquellas que van dirigidas a describir (descripción de esquemas (tarea 2.1)) y comparar (comparativa entre esquemas (tarea 2.2)) los diferentes esquemas de las fuentes que componen el proceso de integración. Y, por otro, las tareas destinadas a describir (descripción de datos (tarea 2.3)) y comparar (comparativa entre datos (tarea 2.4)) los datos. Estos dos grupos de tareas pueden ser llevadas a cabo en paralelo dadas las características de las mismas. Finalmente, se lleva a cabo un proceso de revisión basada en la retroalimentación de los expertos (tarea 2.5).

La actividad de integración estructural (actividad 3), como se comentó con anterioridad, es una actividad opcional. Esta actividad está constituida por la tarea referida al desarrollo de ontologías (actividad 3.1); cuyo proceso está marcado por las necesidades de integración y, de ahí, la propuesta de cuatro enfoques distintos dentro de esta tarea; y una tarea de evaluación (tarea 3.2).

La integración semántica (actividad 4) está compuesta por cinco tareas: La identificación de datos (por ejemplo, a través de URI) (tarea 4.1), creación de *mappings* (tarea 4.2), generación de instancias (por ejemplo, en formato RDF) (tarea 4.3), revisión y retroalimentación de expertos (tarea 4.4) y, por último, una evaluación (tarea 4.5). El conjunto de estas tareas lleva un proceso de desarrollo secuencial.

La actividad de publicación de datos integrados (actividad 5), propuesta como opcional en esta metodología, está constituida por las siguientes tareas: descripción de los conjuntos de datos (metadatos) (tarea 5.1), verificación de datos publicados (tarea 5.2) y visualización de la información publicada (tarea 5.3).

La actividad de depuración / post-procesamiento (actividad 6) está compuesta de las siguientes tareas: detección de duplicados (tarea 6.1), desarrollo de heurísticas (tarea 6.2), creación de enlaces entre conjuntos de datos (tarea 6.3), evaluación de heurísticas (tarea 6.4) y descubrimiento de datos integrados (tarea 6.5).

Finalmente, las actividades de modificación sobre las fuentes de información (actividad 7) se compone de una tarea de recopilación y priorización de eventos (tarea 7.1) y de la presentación de propuestas de cambios (tarea 7.2).

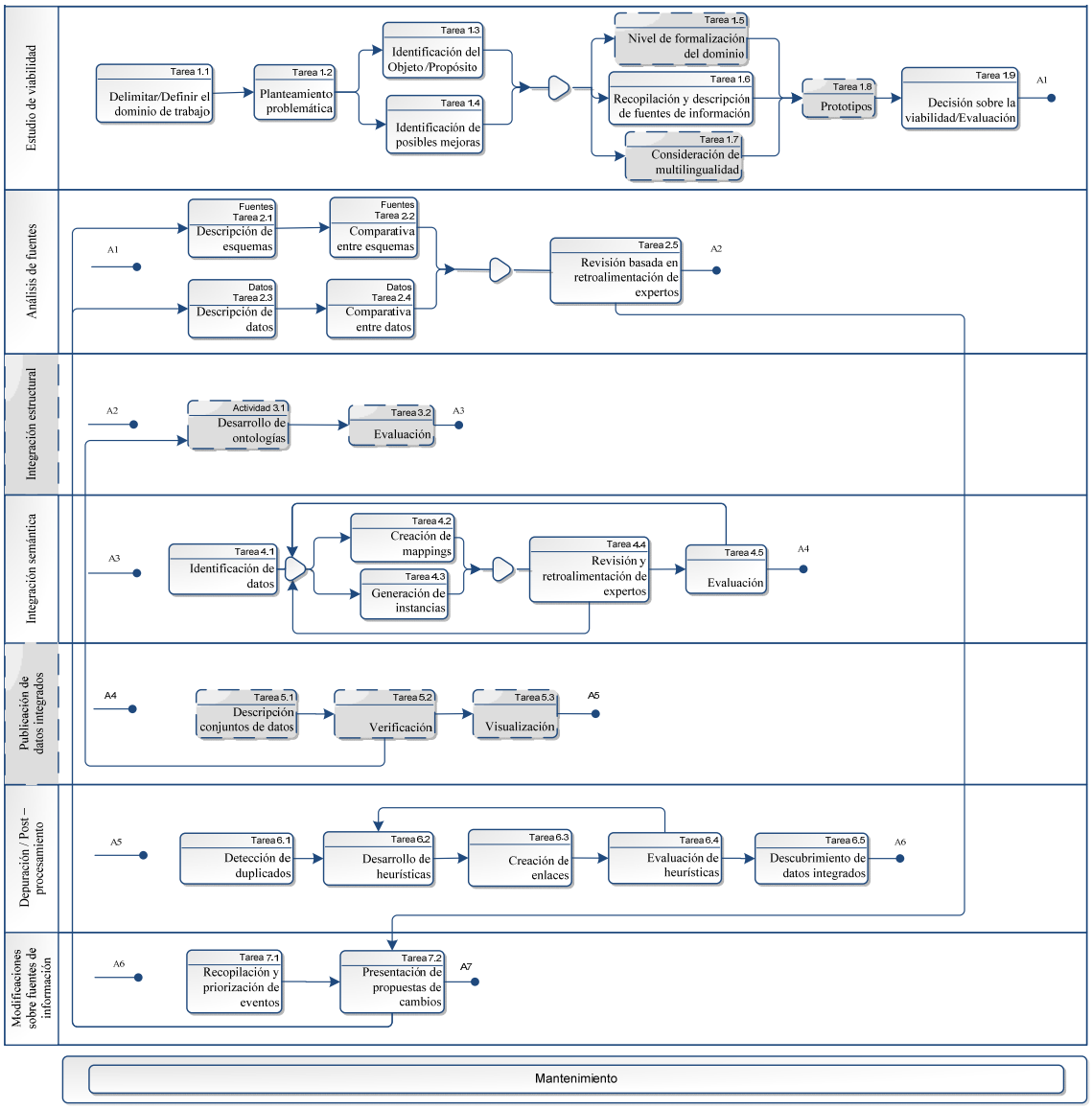


Figura 5.4. Visión general de la propuesta metodológica para la integración basada en ontologías de información de bases de datos heterogéneas en el dominio geográfico

5.4. Responsables implicados. Roles

Las diferentes tareas que componen el proceso de integración de información propuestas en esta metodología se llevan a cabo por diversos participantes según el tipo de rol que deba ser realizado en cada tarea. En esta sección se describen los participantes implicados en el desarrollo de esta propuesta metodológica.

- *Jefe de proyecto*. El jefe de proyecto es el encargado de realizar el control y seguimiento de la ejecución de cada una de las actividades y tareas implicadas en el desarrollo, así como de analizar los riesgos potenciales en el desarrollo y tomar medidas de contingencia para los mismos.
- *Ingeniero Software*. El ingeniero de software es el miembro (o miembros) encargado de aplicar los principios de la ingeniería del software al diseño, desarrollo, testeo y evaluación de cualquier componente software y/o de sistemas.
- *Ingeniero de conocimiento*. El ingeniero de conocimiento es el miembro experto (o miembros) en sistemas informáticos cualificado en el área de los sistemas basados en conocimiento e ingeniería ontológica. De forma habitual, reciben la información de los

expertos de dominio y este se encarga de interpretar y codificar la información haciéndola accesible al usuario final.

- *Experto del dominio*. Un experto de dominio es una persona con conocimiento especial o aptitudes en un área particular de interés. En este caso, el experto deberá tener conocimientos, de forma general, sobre el dominio de la información geográfica. Y, particularmente, derivado del área de aplicación de esta metodología, sobre el dominio hidrográfico.
- *Diseñador/Desarrollador web*. El diseñador/desarrollador web es el miembro responsable del diseño y la programación de la funcionalidad de un sitio web.

5.5. Proceso de desarrollo. Detalles

A continuación se describe en detalle el conjunto de actividades propuestas en esta metodología. Para estas descripciones se usará el caso de uso recogido en la Tabla 5.1.

5.5.1. Actividad 1. Estudio de viabilidad

En la ingeniería del software el objetivo del estudio de viabilidad es analizar la idea o las necesidades para un sistema de software, propuestas potenciales y todos sus beneficios y restricciones del ciclo de vida (IEEE Standard, 1996). Desde esta perspectiva, el estudio de viabilidad puede considerar técnicas para llevar a cabo modelados y prototipos. Asimismo, puede sugerir si es conveniente desarrollar o comprar el sistema, parcialmente o totalmente. Además, cada recomendación debe estar completamente justificada, documentada y formalmente aprobada por todas las organizaciones inmiscuidas (incluyendo a los usuarios y los desarrolladores).

Desde la óptica de la propuesta metodológica presentada en este trabajo el objetivo de esta actividad es analizar, desde una perspectiva genérica, la situación de partida (contexto) del dominio. Como resultado de esta actividad, se obtienen los principales requisitos a satisfacer por esta propuesta metodológica de integración de información. Asimismo, se identifican los principales beneficios y riesgos. Uno de los aspectos más importantes del estudio de viabilidad es el concerniente a las restricciones técnicas, que pueden determinar si es recomendable o no continuar con el proceso de integración de información de bases de datos heterogéneas.

Esta actividad y sus respectivas tareas están inspiradas por las tareas propuestas en MÉTRICA versión 3 (MÉTRICA v.3) y *Rational Unified Process* (RUP) (Kruchten, 2000) para el proceso de estudio de viabilidad de los sistemas software. Sin embargo, las técnicas (y documentos) utilizados en la ejecución de las tareas propuestas en esta actividad están inspiradas por diferentes propuestas de ingeniería del conocimiento, tales como (Gómez-Pérez *et al*., 1997) y (Schreiber *et al*., 1999).

En la Figura 5.5 se muestra una visión detalla de las diferentes tareas propuestas por esta actividad. De esta manera, las primeras tareas consisten en delimitar/definir el dominio de trabajo (tarea 1.1) y llevar a cabo un planteamiento de la problemática (tarea 1.2). El objeto de estas tareas es tener una panorámica de la situación de partida. Una vez llevado a cabo esta primera fase, el siguiente paso consiste en la identificación del objeto/propósito (tarea 1.3) y la identificación de posibles mejoras (tarea 1.4). Ambas tareas conllevan un proceso de análisis de las salidas de las tareas anteriores (1.1 y 1.2) para, de esta manera, definir el contexto de trabajo, los requisitos que con posterioridad caracterizaran el proceso de integración y, por supuesto, identificar posibles líneas de actuación y mejora. Las tareas 1.5, 1.6 y 1.7 contribuyen a profundizar en el proceso de análisis general iniciado en las tareas anteriores (1.3 y 1.4). Las salidas resultantes de este grupo de tareas condicionará en gran medida la decisión final sobre la viabilidad del proceso de integración propuesto en esta metodología. En la tarea 1.8, se realiza una primera aproximación para comprobar la viabilidad técnica del proyecto. Finalmente, en la tarea 1.9, se propone completar una serie de cuestiones, inspiradas en (Schreiber *et al*., 1999),

que permitirán tomar una decisión sobre la viabilidad del proceso de integración y proponer un conjunto de acciones y recomendaciones a seguir.

Tabla 5.1. Caso de uso de aplicación de la metodología

El contexto en el que se lleva a cabo el proceso de instanciación de esta metodología está compuesto de nueve bases de datos diferentes. El conjunto formado por la Base Topográfica Nacional (BTN25), Base Cartográfica Numérica (BCN200), Nomenclátor NOMGEO y Nomenclátor Conciso (NC) pertenecen al Instituto Geográfico Nacional. Por otro lado, un grupo conformado por bases de datos ligadas a proyectos europeos; tales como *EuroGlobalMap* (EGM), *EuroRegionalMap* (ERM) y *Water Information System for Europe*[7] (WISE); productores temáticos, como la Confederación Hidrográfica del Ebro (CHE); y productores locales, caso de la base de datos perteneciente al Instituto Cartográfico de Andalucía (ICA). De esta situación contextual se extraerán los diversos ejemplos que servirán como documentación y caracterización de las diferentes tareas que conforman esta metodología.

Desde una perspectiva de alto nivel, estas bases de datos contienen información a distintas escalas. Asimismo, de forma generalizada, la información que albergan estas fuentes contiene un importante componente multilingüe, con la excepción de las fuentes pertenecientes a WISE e ICA. Una descripción detallada de las fuentes de información que componen el proceso de integración se recoge en el Anexo de este trabajo.

El proceso de integración propuesto por esta metodología no conlleva la creación de un repositorio central de bases de datos, sistema federado o cualquier otro de los enfoques descritos en el Capítulo 3 (sección 3.2.7) de este trabajo. Este hecho deriva de la conformidad, por parte de esta metodología, con los principios INSPIRE. Esto supone que esta metodología de integración de información *garantiza el almacenamiento, disponibilidad y mantenimiento de datos espaciales al nivel más adecuado*; hace posible *combinar, de forma coherente, datos espaciales de diversas fuentes* en el contexto descrito y, al mismo tiempo, hace que (los datos espaciales) *puedan ser compartidos entre distintos usuarios y aplicaciones* (The European Parliament and Council of the EU, 2007).

Asimismo, en lo referente a los datos, las diferentes tareas relacionadas que componen esta metodología están condicionadas por tres tipos de relaciones entre las fuentes de información:

- **Intersección vacía**. Las bases de datos contienen información concerniente a diferentes áreas geográficas.
- **Intersección parcial**. En este tipo de relación entre las bases de datos se produce una intersección parcial en la información. Asimismo, la escala de las fuentes pueden no ser similares.
- **Intersección total**. Las bases de datos tratan con información coincidente, es decir, recogen información sobre un mismo espacio geográfico con escalas similares.

Estas relaciones resultan fundamentales en cualquier proceso de integración de información y, especialmente, para el caso de la IG. Por este motivo, las diferentes tareas de comparación entre fuentes consideradas en esta metodología tienen en cuenta dichas relaciones, centrándose especialmente en las relaciones parciales y totales.

[7] http://water.europa.eu/content/view/16/31/lang,en/

Tabla 5.2. Productos de tareas, técnicas y participantes de la actividad 1 (Estudio de viabilidad)

Nº Tarea	Nombre Tarea	Responsable	Producto	Técnicas aplicables
Tarea 1.1	Delimitar/Definir el dominio de trabajo	Jefe de proyecto / Ingeniero Software	Descripción de alto nivel del dominio y características de las fuentes de información	Sesiones de entrevistas y lista de comprobación
Tarea 1.2	Planteamiento problemática	Jefe de proyecto / Experto del dominio / Ingeniero Software / Ingeniero de conocimiento	Descripción de las diferentes problemáticas actuales	Sesiones de entrevistas
Tarea 1.3	Identificación del Objeto / Propósito	Jefe de proyecto / Experto del dominio / Ingeniero Software / Ingeniero de conocimiento	Descripción de alto nivel sobre el alcance	Sesiones de entrevistas y lista de comprobación
Tarea 1.4	Identificación de posibles mejoras	Experto del dominio/ Ingeniero Software / Ingeniero de conocimiento	Listado de las posibles mejoras identificadas	Sesiones de entrevistas y tormenta de ideas
Tarea 1.5	Nivel de formalización del dominio	Experto del dominio / Ingeniero de conocimiento	Listado recopilatorio sobre los recursos del dominio disponibles y su correspondiente formalización.	Catalogación
Tarea 1.6	Recopilación y descripción de fuentes de información	Experto del dominio/ Ingeniero Software	Catalogación y descripción (rellenado de los metadatos descriptores) de las fuentes de información disponibles	Búsqueda, recopilación y catalogación
Tarea 1.7	Consideración de multilingüismo	Experto de dominio / Ingeniero de conocimiento / Ingeniero Software	Descripción de las características generales del componente multilingüe	Sesiones de entrevistas y listas de comprobación
Tarea 1.8	Realización de prototipos	Experto del dominio, ingeniero software e ingeniero de conocimiento.	Prototipo 1, Prototipo 2, etc.	Técnicas provenientes del resto de tareas (aplicación somera)
Tarea 1.9	Decisión sobre la viabilidad / Evaluación	Jefe de proyecto / Experto del dominio/ Ingeniero Software / Ingeniero de conocimiento	Recomendaciones y propuestas de acción	Lista de comprobación

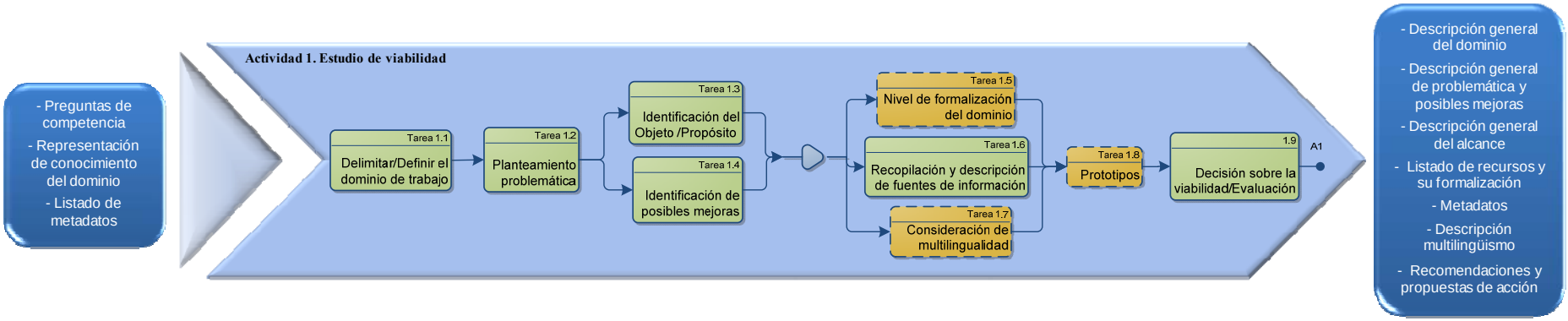


Figura 5.5. Descomposición de tareas de la actividad 1 (Estudio de viabilidad)

Las entradas de esta actividad consisten en diferentes listas de comprobación formadas por múltiples cuestiones, representaciones de conocimiento del dominio y una serie de metadatos a completar. Los resultados obtenidos del conjunto de tareas que conforman esta actividad están constituidos por la descripción general del dominio, una descripción general de sus problemáticas y posibles mejoras, la descripción del alcance del proyecto de integración, un listado de recursos y su formalización, una serie de metadatos, la descripción del componente multilingüe y un listado de recomendaciones y propuestas de acción.

De manera resumida, en la Tabla 5.2, se reflejan las entradas y salidas, técnicas y participantes asociados a cada una de las tareas. A continuación se procederá a describir en detalle cada una de las tareas recogidas en esta actividad.

Tarea 1.1 Delimitar/Definir el dominio de trabajo

Esta tarea trata de ofrecer una visión general de los diferentes aspectos que van a caracterizar el proceso de integración de las múltiples fuentes de información. Por tanto, se trata de recoger las características que describen el dominio de información. Con respecto a las técnicas se recomienda llevar a cabo entrevistas y listas de comprobación con los jefes de proyectos, ingenieros software y expertos de dominio de las partes implicadas en el proyecto. El objetivo es recoger los detalles generales relacionados con el dominio de trabajo.

Material de entrada y resultados de salida

Entradas: Conjunto de cuestiones para la definición del dominio de trabajo.

Salidas: Descripción de alto nivel del dominio y características de las fuentes de información.

Responsable

Ejecutor del proyecto: Jefe de proyecto, ingeniero software y experto de dominio.

Organización: Jefe de proyecto, ingeniero software y experto de dominio.

Técnicas

Sesiones de entrevistas y lista de comprobación (*checklist*). Se recomienda la realización de sesiones de entrevistas y la utilización en las mismas de la lista de comprobación que aparece en la Tabla 5.3 con objeto de poder conseguir una visión general del dominio. Las preguntas que aparecen en esta lista no son exhaustivas, ya que los miembros participantes en el proyecto pueden estar interesados en añadir otras preguntas que contribuyan a clarificar la situación del dominio.

Tabla 5.3. Cuestiones para definir el dominio de trabajo

	Cuestiones
Heterogeneidad	¿De cuántos productores es la información que se maneja / se desea integrar? **En un único productor:** 1. ¿Tiene diferentes fuentes de información (productos)? En caso afirmativo, proceda a enumerarlas. 2. ¿Existe heterogeneidad designativa (componentes léxicos-sintácticos) entre las fuentes de una misma institución/productor? 3. ¿Se han detectado problemas de calidad/consistencia (errores tipográficos, variaciones en medidas, etc.)? **En diferentes productores:** 1. ¿Cuáles son las fuentes de información del conjunto de productores? 2. Entre los diferentes productores, ¿aparecen productores especializados en el dominio hidrográfico? 3. ¿Tienen dichos productores el mismo grado de especialización en la información o pertenecen a diferentes dominios? 4. ¿Existe heterogeneidad designativa (componentes léxicos-sintácticos) entre

	las fuentes de una misma institución/productor? 5. ¿Se han detectado problemas de calidad/consistencia (errores tipográficos, variaciones en medidas, etc.)?
Estructuración de la información	**Formas de estructuración** 1. ¿Cuáles son las formas utilizadas para la estructuración de la información (catálogos, tesauros, diccionarios de datos, etc.)? 2. ¿Cuál es el tipo de formalización utilizada (XML, ficheros de texto, etc.) por las formas de estructuración? **Estructuración** 1. ¿Qué información contemplan las formas utilizadas para estructurar la información (tipos de fenómenos, atributos, características de representación gráfica, definiciones, etc.)? 2. ¿Presentan alguna división/clasificación la información del dominio hidrográfico? 3. ¿Existen dependencias relacionales/organizativas entre los conceptos asociados a los diferentes fenómenos hidrográficos? 4. ¿Existe información jerarquizada o jerarquizable asociada a los diferentes fenómenos hidrográficos? 5. ¿En caso de existir diferentes fuentes de información hidrográfica, tanto en un mismo proveedor como en diferentes, poseen una única clasificación o están clasificados bajo los puntos de vista de diferentes áreas? ¿En caso de diferenciación, cuáles son los criterios de clasificación? ¿Cuáles son diferencias organizativas entre formas de estructuración? **Estandarización** 1. ¿Existe conformidad con alguna norma, especificación, estándar (ISO, OGC, W3C, etc.)? 2. ¿Existe conformidad con algún vocabulario controlado ampliamente difundido/generalizado en el dominio hidrográfico o geoespacial? **Documentación** **En un único productor** 1. ¿Existen diccionarios que recopilen los tipos de fenómenos y sus definiciones en una organización? 2. ¿Aparecen otras características adicionales que contribuyan al enriquecimiento del diccionario? 3. ¿Existen diferentes diccionarios que recopilen los tipos de fenómenos, definiciones y otras características adicionales dentro de una misma organización? 4. ¿Tienen el mismo grado de especialización dichos diccionarios? 5. ¿Hay presencia de diversidad de definiciones sobre un mismo fenómeno hidrográfico (polisemia)? 6. ¿Se utilizan diversidad de fenómenos hidrográficos con un mismo significado (sinónimos) en las diferentes formas de estructuración? 7. ¿Pertenecen todos los fenómenos objeto de estudio al mismo dominio de trabajo o pueden estar considerados desde el punto de vista de otros dominios? **En diferentes productores** 1. ¿Existen diferentes diccionarios que recopilan variedad de definiciones sobre un mismo tipo de fenómeno de manera inter y/o intra institucional? 2. ¿Tienen el mismo grado de especialización dichos diccionarios? 3. ¿Hay presencia de diversidad de definiciones sobre un mismo fenómeno hidrográfico (polisemia)? 4. ¿Se utilizan diversidad de fenómenos hidrográficos con un mismo significado (sinónimos) en las diferentes formas de estructuración? 5. ¿Pertenecen todos los fenómenos objeto de estudio al mismo dominio de trabajo o pueden estar considerados desde el punto de vista de otros dominios?
Escala	¿Escalas/Granularidad asociadas a las fuentes de información? **Misma escala** 1. ¿Existen solapamientos de información (hidrónimos) entre diferentes fuentes a la misma escala? 2. ¿Existe superposición gráfica de un mismo fenómeno hidrográfico entre las diferentes fuentes que poseen la misma escala?

	3. ¿Se identifica conflación geométrica entre las diferentes fuentes? **Diferentes escalas** 1. ¿Existen solapamientos de información (hidrónimos) entre diferentes fuentes a diferentes escalas? 2. ¿Existe superposición gráfica de un mismo fenómeno hidrográfico entre las diferentes fuentes que poseen diferentes escalas? 3. ¿Podría producirse de forma sencilla un proceso de generalización y/o agregación entre los fenómenos de las distintas fuentes?
Bases de datos	1. ¿Cuántas bases de datos poseen? ¿Tienen información sobre el mismo espacio geográfico? ¿Cuáles están especializadas en el dominio hidrográfico? 2. ¿Se identifican metadatos asociados a las diferentes bases de datos para caracterizar a las mismas? 3. ¿Pertenece la información contenida en las fuentes de información a la misma fecha o, al menos, están próximas? 4. ¿Existen grandes diferencias entre las fuentes de información en las fechas de los datos? 5. ¿Cuál es la tasa de cambios relacionadas con las fuentes de información de diferentes localizaciones geográficas? 6. ¿Qué sistemas de información soportan las BD? 7. ¿Distintos tipos de sistemas utilizan una misma BD? 8. ¿Qué lenguajes de consulta se utilizan?
Bases de datos (modelado)	1. ¿Qué tipo de modelado tienen (relacional u orientado a objetos)? 2. ¿Existen relaciones entre tablas? ¿Están contenidas de forma implícita? ¿Se identifican de manera explícita relaciones entre tablas? 3. ¿Qué software (Sistemas Gestores de Bases de Datos) se utiliza para su almacenamiento y gestión de información? 4. ¿Existe documentación de buen nivel sobre el modelado e implementación de las bases de datos?
Bases de datos (datos)	1. ¿Cuáles son los procesos/tratamientos a los que se somete la información desde el proceso de captura hasta su publicación/difusión (linaje)? 2. ¿Sistemas de referencia utilizados? 3. ¿Existen recursos externos (bases de datos, diccionarios, etc.) que puedan ser asociados a estas fuentes de información? ¿Procedentes del mismo o diferente dominio? ¿Cuál es la formalización qué presentan? 4. ¿Se refleja la calidad de los datos? 5. ¿Se deja constancia en la base de datos de que los fenómenos incluidos corresponden o son conformes a alguna forma de estructuración? 6. ¿Aparecen instancias pertenecientes a otros dominios de conocimiento?
Dominio	1. ¿Qué tipo de fenómenos se cubren dentro del dominio hidrográfico? 2. ¿Existe alguna especialización sobre algún fenómeno hidrográfico concreto? 3. ¿Aparece información relativa a otros temas/dominios de conocimiento? 4. ¿La información contenida es generalizable a una visión global del dominio hidrográfico o está delimitada al contexto de trabajo? 5. ¿A qué cuencas hidrográficas pertenece la información recogida? ¿Existe información perteneciente a diversas cuencas hidrográficas dentro de una misma fuente? 6. ¿Cuál es la confederación hidrográfica encargada de la gestión de la cuenca?

Ejemplo práctico

En este ejemplo se completa la información de la lista de comprobación y los resultados obtenidos recogidos en la Tabla 5.4.

Tabla 5.4. Respuestas a las cuestiones para la definición del dominio de trabajo

Cuestiones	
Heterogeneidad	La información pertenece a cuatro productores distintos (Instituto Geográfico Nacional (IGN), Instituto Cartográfico de Andalucía (ICA), Confederación Hidrográfica del Ebro (CHE) y proyecto *Water Information System for Europe* (WISE). **En un único productor** 1. En el caso del IGN posee siete fuentes de información diferentes (BTN25, BCN200, Nomenclátor NOMGEO, Nomenclátor Conciso (NC), Atlas Nacional, *EuroGlobalMap* (EGM) y *EuroRegionalMap* (ERM)). 2. El IGN afirma que, en principio, no deberían existir heterogeneidad designativa entre sus diferentes productos, aunque no se descarta, consecuencia de que el proceso productivo y de gestión se lleva a cabo por diferentes secciones y responsables. 3. Los propietarios de las fuentes de información afirman no tener constancia del tipo de problemas citados en este punto. **En diferentes productores** 1. Al IGN pertenecen las siguientes fuentes BTN25, BCN200, Nomenclátor NOMGEO, Nomenclátor Conciso (NC), Atlas Nacional, *EuroGlobalMap* (EGM) y *EuroRegionalMap* (ERM)); al ICA la base de datos perteneciente a la capa de información hidrográfica; a la CHE la base de datos con la información sobre su cuenca hidrográfica y al proyecto WISE la base de datos *Waterbase*. 2. La información contenida en las diferentes fuentes pertenece al mismo dominio. 3. La diferencia entre los diferentes productores es el tipo de información que producen. Así, según Boletín Oficial del Estado (BOE) (1986) la información producida por el IGN e ICA es básica, mientras la Confederación Hidrográfica del Ebro y WISE producen información temática. 4. Los propietarios de las fuentes de información afirman no tener constancia del tipo de problemas citados en este punto. 5. Los propietarios de las fuentes de información afirman no tener constancia del tipo de problemas citados en este punto.
Estructuración de la información	**Formas de estructuración** 1. Catálogos de fenómenos (a modo de simple listado) y diccionario de datos. 2. Catálogos de fenómenos (ficheros de texto, *.pdf, hojas de cálculo [*.xls], base de datos [*.mdb]) y diccionarios de datos (*.pdf y *.doc). **Estructuración** 1. Entre otros dominios, todas disponen de información perteneciente al dominio hidrográfico. BTN25 y BCN200 contiene tipos de fenómenos, características de representación gráfica y clasificación jerárquica acorde al tema, grupo y subgrupo. NC y NOMGEO posee únicamente el tipo de fenómeno. EGM y ERM poseen tipos de fenómenos, definiciones, atributos y tipo de geometría. WISE, ICA y CHE, en su forma de estructuración de información en sentido tradicional, únicamente poseen el tipo de fenómeno. 2. Las diferentes fuentes de información poseen diversos criterios de clasificación en sus formas de estructuración, en función de las necesidades y criterios aplicados a cada producto. Considerando este hecho, las formas de estructuración se dividen entre aquellas que tienen una clasificación jerárquica implícita y aquellas que son simples listados. 3. Dadas las características de las formas de estructuración, este tipo de relaciones no se explicitan entre conceptos. Aunque, de las relaciones jerárquicas implícitas puede extraerse conocimiento sobre estas relaciones. 3. Sí, en los catálogos de fenómenos de BTN25 y BCN200. 4. Las distintas fuentes de información, aunque con algunas similitudes, poseen clasificaciones diferentes, incluso en aquellas que pertenecen a un mismo productor. **Estandarización** 1. El vocabulario contenido en cada forma de estructuración es propio, especializado y no conforme a ningún vocabulario controlado conocido. Aunque los diferentes productores comentan que este hecho no significa que la gran parte

	de la información contenida no pueda ser enriquecida con referencias a este tipo de vocabularios. 2. No, las diferentes formas de estructuración son propias de cada organización. **Documentación** **En un productor** 1. La existencia de diccionarios dentro de las organizaciones en un hecho excepcional. A pesar de este hecho, el IGN posee un diccionario de fenómenos y dentro de los proyectos *EuroGlobalMap* y *EuroRegionalMap* existen otro conjunto de definiciones asociadas a los fenómenos, aunque sin relación ni puesta en común entre ellos. 2. No se recoge información adicional en los diccionarios, limitándose estos a definir los fenómenos recogidos. 3. Sí, aunque no es muy habitual, se dan casos como el comentado en la respuesta 1. 4. No todos los diccionarios tienen el mismo grado de especialización, ya que la escala de la información recopilada varía de unas fuentes a otras. 5. Sí, aunque no es muy habitual, se dan casos como el comentado en la respuesta número 1. Aunque en este caso la polisemia se entiende como diversidad en el texto que define un mismo concepto (glosa) y no como multiplicidad de definiciones. 6. En principio la utilización de sinónimos parece estar controlado, aunque no se descarta, consecuencia de que el proceso productivo y de gestión se lleva a cabo por diferentes secciones y responsables. Diferentes responsables dentro del proyecto comentan que es posible que este problema aparezca al estudiar en profundidad las diferentes formas de estructuración tanto a nivel intra como inter organizacional. 7. Todos los fenómenos pertenecen al dominio hidrográfico y están clasificados bajo este punto de vista. **En diferentes productores** 1. Los diccionarios recopilados pertenecen a un único productor. Como se comentó con anterioridad, existen variaciones en las definiciones recopiladas en el ámbito intra-institucional. 2. – 3. – 4. – 5. –
Escala	BTN25 (1:25.000), BCN200 (1:200.000), NOMGEO (1:25.000) NC (1:1.000.000), Atlas Nacional (1:500.000), CHE (1:50.000), EGM (1:1.000.000), ERM (1:200.000), ICA (1:5.000), WISE (escala europea). **Misma escala** 1. Sí, existen solapamientos de información entre fuentes con la misma escala, por ejemplo, BTN25 y NOMGEO. 2. Existen superposiciones gráficas parciales de diversos fenómenos. 3. Los diferentes productos son desarrollados por diferentes secciones y responsables, por tanto, es previsible encontrar conflación geométrica incluso entre las fuentes con escala similar. **Diferente escala** 1. Sí, existen solapamientos de información entre fuentes con diferente escala, por ejemplo, fuentes con la misma escala BCN200 y ERM. Y en el caso de diferentes escalas BTN25 y CHE. 2. Existen superposiciones gráficas parciales de diversos fenómenos. 3. Dadas las características de las fuentes y las escalas de recogida de información el proceso de generalización y/o agregación entre fenómenos, aunque podría llevarse a cabo, resulta complicado en primera instancia.
	1. El proceso de integración consta de 10 bases de datos. Estas bases de datos contienen información comparable sobre el mismo espacio geográfico. Asimismo, todas estas fuentes, entre otros dominios, disponen de información perteneciente al dominio hidrográfico. 2. No se identifican metadatos descriptores de las fuentes. 3. Aunque no se poseen fechas concretas, los diferentes productores afirman que

Bases de datos	las fuentes de información consideradas son últimas versiones de sus productos y que la información contenida pertenece a fechas recientes. 4. – 5. De forma general, el proceso de adquisición de información (actualización) es más frecuente en los espacios urbanos y menor en las zonas rurales. 6. La información contenida en BTN25, BCN200, NOMGEO y NC abastece, entre otros productos, a la IDEE y SIGNA. ERM, EGM, WISE abastece a sistemas específicos de estos proyectos europeos. Finalmente, las bases de datos de ICA y CHE abastecen a IDEA e IDE-Ebro, respectivamente, y a diversos sistemas de información local. 7. Sí, como se deja manifiesto en la respuesta anterior. 8. SQL (*Structured Query Language*)
Bases de datos (modelado)	1. Las diferentes bases de datos con las que cuenta este proyecto están modeladas siguiendo el paradigma de bases de datos relacionales. 2. Sí existen relaciones entre las diferentes tablas. Estas relaciones están identificadas de forma explícita. 3. Aparece generalizada la utilización del software Oracle entre las diferentes bases de datos, aunque destaca la utilización de bases de datos *.mdb (Microsoft Access) en ERM y EGM. 4. No es frecuente la existencia de documentación de buen nivel sobre el modelado e implementación de las bases de datos en las distintas organizaciones productoras.
Bases de datos (datos)	1. De forma general, los procesos a los que se somete la información son: control topológico, geométrico y semántico. 2. El sistema de coordenadas utilizado es ED50- geográficas. 3. A pesar de la especialización del dominio, los expertos de dominio confirman la existencia de diferentes recursos externos, tales como diccionarios, tesauros, etc. En lo que respecta a la formalización, existen recursos tanto en papel como en digital. 4. No es frecuente recoger información sobre la calidad de los datos en las fuentes. 5. No hay una constancia explícita, aunque en el caso de BTN25 y BCN200 existe una tabla que recoge el catálogo de fenómenos que soporta la estructuración. 6. Las instancias de diferentes dominios, de forma generalizada aparecen en tablas distintas, aunque es frecuente la mezcla de instancias vinculadas a diversos tipos de fenómenos dentro de una misma tabla (por ejemplo, Río y Laguna).
Dominio	1. De forma general, en las fuentes aparecen fenómenos universalmente conocidos, por ejemplo, río, arroyo, ría, laguna, etc. 2. Las fuentes de información no presentan especialización sobre ningún fenómeno concreto. 3. Son diversas las fuentes que contienen información sobre otros dominios, por ejemplo, en BTN25 aparece información sobre vías de comunicación, relieve, etc. 4. La información contenida es generalizable a una visión global del dominio, aunque pueden existir fenómenos de tipo local. 5. Entre las fuentes de información aparece información del ámbito nacional, por tanto contienen información de las diferentes cuencas hidrográficas que conforman la red hidrográfica nacional, e información de una cuenca concreta, caso de la fuente referida a la cuenca hidrográfica del Ebro.

Tarea 1.2 Planteamiento de la problemática

El objetivo de esta tarea es identificar los problemas presentes en las diferentes fuentes de información que cubren el dominio mediante un proceso de abstracción sobre las diversas cuestiones identificadas en la tarea 1.1. De esta manera, esta tarea contribuye a identificar posibles focos de atención sobre determinados problemas que pueden requerir diferentes niveles de prioridad y/o actuación en el proceso de integración.

Para llevar a cabo esta tarea todas las personas responsables del proyecto están involucradas (jefe de proyecto, experto del dominio, ingeniero software e ingeniero de conocimiento). La técnica recomendada consiste en llevar a cabo sesiones de entrevistas con todos los miembros involucrados donde la intención es extraer la problemática que acontece en el dominio.

Material de entrada y resultados de salida

Entradas: Respuestas a las cuestiones

Salidas: Documento contenedor de las diferentes problemáticas actuales

Responsables

Ejecutor del proyecto: Jefe de proyecto, experto del dominio, ingeniero software e ingeniero de conocimiento.

Organización: Jefe de proyecto, experto del dominio, ingeniero software e ingeniero de conocimiento.

Técnicas

Sesiones de entrevistas.

Ejemplo práctico

Los problemas extraídos del listado de cuestiones, reflejado en la tarea 1.1, y de las sesiones de entrevistas con los responsables mencionados son agrupados bajo cada una de temáticas en las que se agrupan las diversas preguntas del mencionado listado. De forma general, en el contexto de integración objeto de este trabajo se identifican los siguientes problemas:

Heterogeneidad.

- La existencia de productores de información básica y temática conlleva diferente granularidad en el estudio de los fenómenos geográficos asociados.

Estructuración de la información.

- La diversidad (componentes, clasificación, desarrollo, gestión, etc.) es la principal característica de las diferentes formas de estructuración. Este hecho puede ser consecuencia directa de la ausencia de seguimiento de los diversos estándares internacionales.
- Los fenómenos geográficos contenidos en las formas de estructuración no son conformes a ningún vocabulario controlado.
- División y diversidad de los procesos de desarrollo y gestión provoca ausencia de control global y conocimiento profundo de las fuentes.
- Ausencia de relaciones entre diversas fuentes de una misma organización. Además, dentro de una misma forma de estructuración se producen ausencias de relaciones explícitas.

Escala.

- La presencia de diversas fuentes con heterogeneidad de escalas de recopilación de información hace difícil una comparación profunda entre las fuentes de información.

Bases de datos.

- La ausencia de metadatos descriptores dificulta la comparación en detalle de las fuentes de información.
- La heterogeneidad en el proceso de modelado, dominado por necesidades particulares, dificulta el proceso de integración.
- Existencia de información duplicada, consecuencia de los solapamientos de localizaciones geográficas y la consideración de escalas semejantes.

Tarea 1.3 Identificación del objeto / propósito

En esta tarea se va a determinar el alcance del proceso de integración de información de bases de datos heterogéneas con el objetivo de proporcionar una descripción general de las necesidades del proceso de integración en un determinado dominio de información. Esta descripción genérica resaltará, principalmente, los resultados esperados, el contexto y las problemáticas identificadas en el proceso de integración de fuentes de información.

La realización de esta tarea involucra a los diferentes responsables involucrados en el proyecto, es decir, implicará al jefe de proyecto, ingeniero de conocimiento, ingeniero software y experto del dominio. Las técnicas a utilizar en esta tarea serán las entrevistas, principalmente con el jefe de proyectos, sesiones de tormentas de ideas y diferentes preguntas con todos los participantes para identificar los requisitos de la integración propuesta.

Material de entrada y resultados de salida

Entradas: Abstracción de respuestas a las cuestiones y formularios.

Salidas: Descripción de alto nivel sobre el alcance.

Responsables

Ejecutor del proyecto: Jefe de proyecto, ingeniero de conocimiento, ingeniero software y experto del dominio.

Organización: Jefe de proyecto, ingeniero de conocimiento, ingeniero software y experto del dominio.

Tabla 5.5. Cuestiones para la identificación del propósito

	Cuestiones
Identificación del propósito	1. ¿Qué fuentes de información son elegidas para llevar a cabo el proceso de integración? 2. ¿Se selecciona alguna localización geográfica donde aparezcan múltiples fuentes? o ¿se lleva a cabo el proceso de integración de forma generalizada? 3. ¿Qué tema/dominio se seleccionan para llevar a cabo el proceso de integración? ¿Existen relaciones entre algunos de los temas/dominios? ¿Se identifica alguno con mayor prioridad o interés? 4. ¿Cuáles son las aplicaciones que soportan las fuentes de información implicadas en el proceso de integración? ¿Son las mismas aplicaciones a las que se destina la información integrada o se mantendrán independientes? 5. ¿Cuál es la finalidad del proceso de integración de información?

Técnicas

Sesiones de entrevistas y lista de comprobación. Se recomienda la realización de sesiones de entrevistas y la utilización en las mismas de la lista de comprobación que aparece en la Tabla 5.5. Las preguntas que aparecen en esta lista no son exhaustivas, ya que los usuarios pueden estar interesados en añadir otras preguntas que contribuyan a concretar el propósito del proyecto.

Ejemplo práctico

El listado de cuestiones se completa en este ejemplo y los resultados se muestran en la Tabla 5.6.

Tarea 1.4 Identificación de posibles mejoras

El objetivo de esta tarea es identificar las posibles mejoras a realizar en el proceso de integración de información de bases de datos heterogéneas. Esta tarea es dependiente de la tarea 1.2, es decir, las posibles vías de actuación van a depender de los problemas identificados en las sesiones de entrevistas realizadas con objeto de dicha tarea.

Tabla 5.6. Respuestas a las diversas cuestiones para la identificación del propósito

	Cuestiones
Identificación del propósito	1. Todas las fuentes de información mencionadas en el listado de comprobación de la tarea 1.1, aunque se prestará especial atención a BTN25, NOMGEO, CHE e ICA. 2. En este caso el proceso de integración constará de una integración generalizada, a través de fuentes como BTN25 y NOMGEO, y también se centrará en Aragón (CHE) y Andalucía (ICA). 3. El dominio seleccionado corresponde a la hidrografía. 4. Fundamentalmente las aplicaciones son Infraestructuras de Datos Espaciales (IDE), concretamente, IDEE, IDEA e IDE-Ebro. Asimismo, las aplicaciones de destino de este proceso son las mismas. 5. La finalidad de este proceso es ofrecer una visión integrada de la información hidrográfica en el contexto de la realidad geoespacial.

Esta tarea implica la participación del ingeniero de software, ingeniero de conocimiento y experto del dominio. Las técnicas empleadas en esta tarea se ciñen a entrevistas y sesiones de tormenta de ideas que contribuirán a la identificación de posibles mejoras.

Material de entrada y resultados de salida

Entradas: Problemática identificada en el documento obtenido de la tarea 1.2.

Salidas: Documento genérico con las posibles mejoras identificadas.

Responsables

Ejecutor del proyecto: Ingeniero software, ingeniero de conocimiento y experto del dominio.

Organización: Ingeniero software, ingeniero de conocimiento y experto del dominio.

Técnicas

Sesiones de entrevistas y tormentas de ideas.

Ejemplo práctico

En este ejemplo se identifican, de manera no exhaustiva, una serie de posibles mejoras a realizar en el proceso de integración de información llevado a cabo por esta propuesta metodológica.

Estructuración de la información.

- Realización/reutilización de una ontología para el dominio hidrográfico que sirva como:
  - Instrumento armonizador del vocabulario y sus componentes (atributos).
  - Metamodelo integrador de los diversos modelos relacionales de las fuentes.
  - Conformidad con estándares (ISO, W3C y/u OGC).
  - Conformidad o vinculación con vocabularios controlados extendidos.
  - Establecimiento de relaciones entre las diversas formas de estructuración y la forma global desarrollada.
  - Enriquecimiento semántico con las relaciones del dominio.

Escala.

- Integración de las diversas fuentes de información mediante el establecimiento de correspondencias para conformar una visión "completa" y conjunta de la realidad geoespacial. De esta manera se evitaran las islas de información y la creación múltiple e innecesaria de fuentes y se tenderá al enriquecimiento del proceso de integración.

Bases de datos.

- Descripción de alto nivel de los esquemas y datos de las diversas fuentes de información.
- Desarrollo de diferentes mecanismos de desambiguación de la información duplicada.

Tarea 1.5 Formalización del dominio

El objetivo de esta tarea consiste en la búsqueda de recursos disponibles sobre el dominio de información (catálogos, tesauros, ontologías, modelos UML, etc.), descripción de su formalización y análisis de las posibilidades de reutilización de los mismos. Esta tarea, dado su objetivo, puede ser opcional en caso de que la organización u organizaciones implicadas en el proceso de integración identifiquen, conforme al propósito (tarea 1.3), recursos propios como suficientes. En este caso, esta tarea se limitará a una mera recopilación de recursos existentes, por lo que podría llevarse a cabo en el contexto de la tarea 1.6.

En esta tarea tienen un papel fundamental tanto los expertos de dominio como los ingenieros de conocimiento implicados en el proyecto. Por un lado, los expertos de dominio pueden proporcionar una valiosa información sobre las fuentes de información utilizadas. Por otro lado, los ingenieros de conocimiento participantes en el proyecto pueden aportar recursos más formales, desde la perspectiva semántica, y el proceso de abstracción necesario para la viabilidad de la reutilización de los recursos disponibles dentro del dominio.

Tabla 5.7. Recopilación no exhaustiva de recursos disponibles dentro del dominio hidrográfico

Nombre del recurso	Tipo	Formalización
Ámbito nacional		
BTN25	Catálogo de fenómenos	*.pdf
BCN200	Catálogo de fenómenos	*.txt
ERM	Especificaciones	*.pdf
EGM	Especificaciones	*.pdf
Comunidades Autónomas	Catálogo de fenómenos	*.pdf, *xls, *.doc
Nom. Geográfico Nacional	Nomenclátor	*.doc
Nomenclátor Conciso	Nomenclátor	*.doc
Diccionarios del dominio	Diccionario	Papel / electrónico
Monografías	Libros	Papel
Ámbito internacional		
Directiva Marco del Agua	Directiva Europea	*.pdf
Proyecto SDIGER	Especificaciones	*.pdf, UML
UNESCO	Tesauro	HTML
Alexandria Digital Library	Tesauro	HTML
Getty	Tesauro	HTML
GEMET	Tesauro	HTML
GeoNames	Nomenclátor/Ontología	XML/OWL
Hydrology	Ontología	OWL
Topography	Ontología	OWL
Mereological	Ontología	OWL
Spatial relations	Ontología	OWL
ISO 19115	Ontología	OWL
GML	Ontología	OWL

Material de entrada y resultados de salida

Entradas: Representaciones existentes sobre el dominio.

Salidas: Listado recopilatorio sobre los recursos del dominio disponibles y su correspondiente formalización.

Responsables

Ejecutor del proyecto: Experto del dominio e ingeniero de conocimiento.

Organización: Experto de dominio.

Técnicas

Procesos de búsqueda y catalogación.

Ejemplo práctico

En este ejemplo se recoge un listado no exhaustivo de recursos disponibles dentro del dominio hidrográfico.

Tarea 1.6 Recopilación y descripción de fuentes de datos

El objetivo de esta tarea es recopilar todas las fuentes de información existentes que puedan ser reutilizadas, parcial o totalmente, en la organización u organizaciones sobre las que se va a llevar a cabo el proceso de integración de información de bases de datos heterogéneas. Junto al proceso de recopilación de las fuentes de datos, en esta tarea también se lleva a cabo una descripción genérica de las características de las fuentes mediante la recopilación de una serie de metadatos generales utilizados para describir las fuentes. Estos metadatos, reflejados en la Tabla 5.8, están formados por una combinación de metadatos geográficos, extraídos del Núcleo Español de Metadatos (NEM) (Subgrupo de Trabajo del Núcleo Español de Metadatos, 2005), y una serie de metadatos descriptores, basados en metadatos de bases de datos relacionales y almacenes de datos (*data warehouses*), de las características de las fuentes de información.

Esta tarea se centrará en el proceso de recopilación, catalogación y descripción (rellenado de metadatos descriptores) de las fuentes de datos. El objetivo de esta tarea implicará de forma directa al experto del dominio e ingeniero software participantes en el proyecto.

Material de entrada y resultados de salida

Entradas: Listado de metadatos a completar

Salidas: Catalogación y descripción de las fuentes de información disponibles.

Responsables

Ejecutor del proyecto: Experto del dominio e ingeniero software.

Organización: Experto del dominio e ingeniero software.

Técnicas

Recopilación, catalogación y descripción de fuentes de datos.

Ejemplo práctico

En este ejemplo se rellena el listado de metadatos propuesto en la Tabla 5.8 y se ofrecen los metadatos correspondientes a BTN25 en la

Tabla 5.10.

Tarea 1.7 Consideración del aspecto multilingüe

La consideración del aspecto multilingüe en las fuentes de información se presenta como una tarea opcional en esta metodología, ya que la presencia de este tipo de información no es generalizada entre las fuentes. A pesar de esto, el tratamiento de este aspecto supone un enriquecimiento tanto de los propios datos como del modelado de los mismos.

El objetivo de esta tarea consiste en obtener una visión de alto nivel de la componente multilingüe. De esta manera, se analiza el multilingüismo mediante una serie de preguntas que examinan la situación presente en los datos y modelos.

En esta tarea participan el experto de dominio, ingeniero software e ingeniero de conocimiento. Asimismo, las técnicas empleadas consistirán en entrevistas, llevadas a cabo con los miembros implicados, en los que se procederá a completar el formulario de preguntas que se presenta en la Tabla 5.11.

Tabla 5.8. Listado de metadatos sobre las fuentes de información

Metadatos geográficos	Metadatos de bases de datos
Recurso (base de datos) •Título •Resumen descriptivo •Versión •Escala equivalente •Uso •Formato de distribución •Consistencia conceptual •Consistencia lógica *Organización* •Autor •Nombre de la organización •Fecha de creación •Conservador •Idioma de los metadatos •Punto de contacto o Dirección física o Dirección electrónica	*Tablas de todas las tablas*: •Nombres de tablas •Tamaño •Número de filas en cada tabla. *Tablas de todas las columnas*: •Qué tablas se utilizan •Tipo de datos almacenados en cada tabla. •Sinónimos •Rango de valores •Índices *Esquemas lógicos de la fuente.* *Información descriptiva de la fuente*: •Frecuencia de actualizaciones •Métodos de acceso. *Gestión de las dimensiones de tabla*: •Definiciones de dimensiones •Asignación de claves de subrogación.

Material de entrada y resultados de salida

Entradas: Preguntas.

Salidas: Descripción de las características generales del componente multilingüe.

Responsables

Ejecutor del proyecto: Experto del dominio, ingeniero de software e ingeniero de conocimiento.

Organización: Experto del dominio e ingeniero de software.

Tabla 5.9. Cuestiones sobre el multilingüismo de las fuentes de información

	Cuestiones
multilingüismo	1. ¿Existe demanda de información multilingüe? 2. ¿Existe información en varios idiomas a nivel de fenómeno en las formas de estructuración? 3. ¿Qué idiomas contemplan las fuentes de información? ¿Cuáles son los idiomas que pretende cubrir? 4. ¿Se considera el multilingüismo en el modelado de las bases de datos? ¿Existen tratamiento/consideración a nivel de fenómenos? ¿y de instancias? 5. ¿Se identifica/asocia el idioma a cada instancia o fenómeno? 6. ¿Están controlados los sinónimos utilizados para un mismo fenómeno en el contexto multilingüe? ¿y las definiciones? 7. Las instancias multilingües presentes en las bases de datos ¿son consideradas como una instancia única? ¿Existe tratamiento como instancias diferentes (monolingües/multilingües)? 8. ¿Poseen los diferentes sistemas interfaces multilingües? 9. ¿Se dispone de recursos (humanos y/o material) para el tratamiento de la componente multilingüe en sistemas y/o información? 10. ¿Conoce el tratamiento multilingüe de la información en otras instituciones?

Tabla 5.10. Metadatos sobre las fuentes de información

Metadatos geográficos	Metadatos de bases de datos
Recurso (base de datos) •Base Topográfica Nacional (BTN25) •Base de datos correspondiente a la información de BTN 1:25.000 •Versión 3.0 •1:25.000 •Explotación para la producción cartográfica en entornos SIG •DGN (gráfico), Oracle (alfanumérico) •No se recoge información sobre la consistencia conceptual y lógica *Organización* •IGN - Subdirección General de Producción Cartográfica •Instituto Geográfico Nacional •2009 •Francisco Javier García García •Español •Punto de contacto o C/General Ibáñez de Íbero, 3 28003, Madrid. España o fjgarcia@fomento.es	*Tablas de todas las tablas*: •**Nombres tabla**: Esta base de datos contiene 411 tablas. Algunas de las tablas que aparecen son: BCN0104P_HITO, BCN0656L_PUERTO, BCN0101_LIM_ADM_CART, BCN0301L_RIO, BCN0304L_CAU_ART, etc. •**Tamaño:** BCN0104P_HITO (249), BCN0656L_PUERTO (581), BCN0101_LIM_ADM_CART (835), BCN0301L_RIO (985), BCN0304L_CAU_ART (781) , etc. •**Número filas**: BCN0104P_HITO (249 filas), BCN0656L_PUERTO (1.566), BCN0101_LIM_ADM_CART (41.835), BCN0301L_RIO (445.634), BCN0304L_CAU_ART (69.224), etc. *Tablas de todas las columnas*: •**Tablas utilizadas**:BCN0301L_RIO , BCN0304L_CAU_ART, BCN0307L_TUBERI, BCN0310L_RAMBLA, BCN0313L_RIA, BCN0316S_LAGUNA, BCN0319S_HUMEDA, BCN0334P_SURGEN, BCN0337P_CASCAD, BCN0355L_ISLA y BCN0358S_SALINA •**Tipos de datos**: De forma general, los tipos de datos utilizados para el conjunto de las tablas de BTN25 son: INTEGER, VARCHAR, FLOAT y SDO_GEOMETRY. •**Sinónimos**: No hay sinónimos definidos para estas tablas •**Rango de valores**. No hay información •**Índices**: Todas las tablas estudiadas contiene tres índices. El nombre de los índices está formado por el nombre de la tabla más la incorporación final de _ID_HOJA, _PK y _RTREE. A modo de ejemplo, se recoge el siguiente caso: BCN0301L_RIO (BCN0301L_RIO _ID_HOJA; BCN0301L_RIO _PK; BCN0301L_RIO _RTREE) *Esquemas lógicos de la fuente.* *Información descriptiva de la fuente*: •Nueva versión, puede tardar 1 o 2 años más en terminarse de forma completa. •Métodos de acceso. No hay información *Gestión de las dimensiones de tabla*: •Definiciones de dimensiones. No hay

	infor. •Claves de subrogación. PK: ID

Técnicas

Sesiones de entrevistas y lista de comprobación *(checklists)*. Se recomienda la realización de sesiones de entrevistas y la utilización en las mismas de la lista de comprobación que aparece en la Tabla 5.9. Las preguntas que aparecen en esta lista no son exhaustivas, ya que los usuarios pueden estar interesados en añadir otras preguntas con objeto de analizar ciertos aspectos dentro del componente multilingüe de la información.

Ejemplo práctico

En este ejemplo se completan las cuestiones anteriores y se ofrecen los resultados en la Tabla 5.11.

Tabla 5.11. Respuestas a las preguntas sobre la componente multilingüe de las fuentes de información

	Cuestiones
multilingüismo	1. Sí, consecuencia del contexto de la Sociedad de la Información y del hecho de que España es un país donde cohabitan 5 lenguas oficiales. 2. No, en las formas de estructuración únicamente aparecen fenómenos en un idioma. 3. En las fuentes de información aparece datos en vasco, gallego, catalán y castellano. Los idiomas a cubrir serían, principalmente, inglés y el de los países fronterizos a España, es decir, portugués y francés. 4. De forma genérica el multilingüismo no es considerada en el modelado de las bases de datos, aunque en el caso del Nomenclátor Conciso aparece un campo destinado a identificar el idioma en el que aparece el topónimo. 5. No, salvo en el caso mencionado con anterioridad. 6. No, los fenómenos o instancias multilingües son considerados un elemento más, ya que no existe relación entre estos elementos. Además, no existen definiciones de los tipos de fenómenos para el contexto multilingüe. 7. No existe tratamiento multilingüe de las instancias en las diversas bases de datos, por tanto, las instancias son consideradas elementos diferentes. 8. Sí, aunque no en todos los casos existe una traducción completa de todos los interfaces. 9. No se dispone de personal especializado en esta materia, aunque sería posible traducir los diferentes tipos de fenómenos a las lenguas oficiales de España así como contactar con ciertos responsables que pudieran encargarse de su traducción a otros idiomas.

	10. No, al menos, de forma directa, aunque dada la vinculación de diversos miembros de este proyecto en diferentes proyectos europeos e internacionales resultaría sencillo conseguir cierta información al respecto.

Tarea 1.8 Realización de prototipos

Esta tarea opcional trata de ofrecer un primer acercamiento, a modo de pruebas iniciales, para comprobar la viabilidad técnica del proyecto. Estas pruebas permiten la generación de prototipos de las diferentes actividades de esta metodología de integración de datos geográficos. Asimismo, estos prototipos iniciales pueden servir como punto de partida (entrada) para la ejecución de forma pormenorizada de las diversas tareas descritas. La aparición de dificultades para desarrollar prototipos sobre alguna de las tareas recogidas en esta metodología puede influir directamente sobre la decisión de la viabilidad del proyecto a ejecutar.

Material de entrada y resultados de salida

Entradas: Diversas bases de datos.

Salidas: Prototipo 1, Prototipo 2, etc.

Responsables

Ejecutor del proyecto: Experto del dominio, ingeniero software e ingeniero de conocimiento.

Organización: Experto del dominio, ingeniero software e ingeniero de conocimiento.

Técnicas

Las técnicas para la realización de estos prototipos aparecen detalladas en las diferentes tareas que componen esta metodología. La diferencia consiste, básicamente, en que las ejecuciones de las diferentes técnicas realizadas en esta tarea no se realizan de forma exhaustiva. Un ejemplo de la realización de un prototipo de la Actividad 4 se recoge en la Figura 5.6. En este ejemplo, se muestra la realización de unas pruebas iniciales para la generación de *mappings* entre una ontología y una base de datos seleccionada. Estos *mappings* se crean mediante la utilización del plugin ODEMapster, presente en la herramienta NeOn Toolkit.

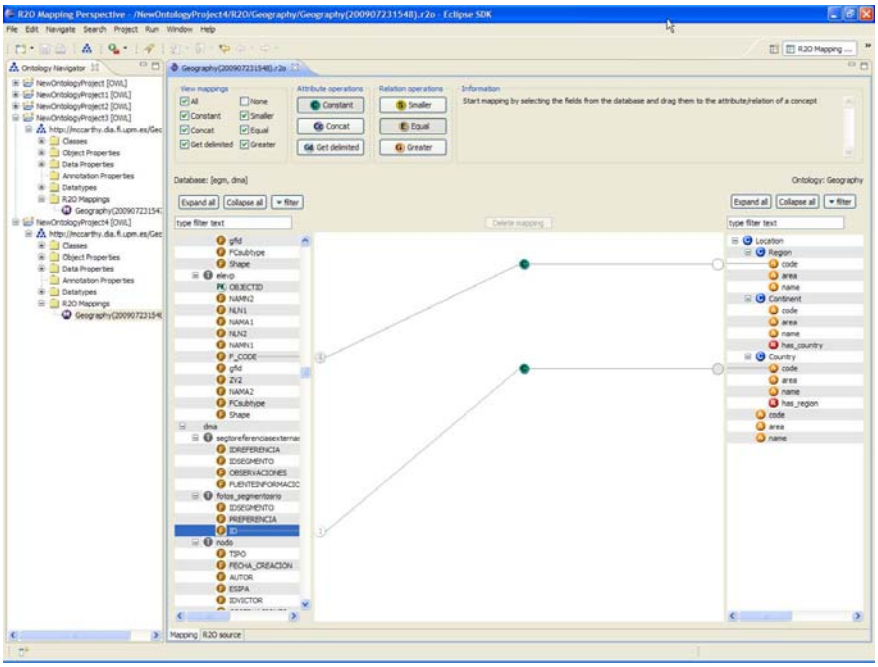


Figura 5.6. Utilización del plugin ODEMapter para la generación de *mappings*

Tarea 1.9 Decisión sobre la viabilidad / Evaluación

El objetivo principal de esta tarea es la toma de decisión sobre la viabilidad del proceso de integración de información de bases de datos heterogéneas, así como recomendar cuales son las acciones específicas que deberían ser consideradas para solventar los problemas que pueden aparecer durante el proceso de integración.

En el proceso de evaluación de la viabilidad del proyecto van a ser clave todas las salidas obtenidas en las tareas anteriores, ya que van a ser utilizadas para identificar los puntos fuertes y débiles del proceso. A esto se le une la utilización de una lista de comprobación (*checklist*) inspirada en la metodología *CommonKADS* (Schreiber *et al*., 1999), adaptada al contexto específico del proceso de integración de información.

Si el proyecto se considera viable, las tareas previas ayudarán a preparar el inicio del proceso de integración de información basado en ontologías y servirán como entrada de las subsiguientes actividades. Asimismo, se identifican los principales costes, beneficios y riesgos del mencionado proceso.

Material de entrada y resultados de salida

Entradas: Documentos pertenecientes a todos los productos finales (salidas) de las tareas anteriores y listado de preguntas, inspiradas en la ingeniería del conocimiento, sobre la viabilidad del proyecto en diferentes aspectos.

Salidas: Recomendaciones y propuestas de acción.

Responsables

Ejecutor del proyecto: Jefe de proyectos, experto del dominio, ingeniero software e ingeniero de conocimiento.

Organización: Jefe de proyectos, experto del dominio, ingeniero software e ingeniero de conocimiento.

Tabla 5.12. Lista de comprobación para la decisión sobre la viabilidad del proyecto (inspirada en los formularios de CommonKADS OM-5 (Schreiber *et al*., 1999))

	Preguntas para decisión sobre viabilidad
Viabilidad de negocio	1. ¿Cuáles son los beneficios esperados del proceso de integración de información para la organización u organizaciones? 2. ¿Cuál es el valor añadido esperado? 3. ¿Cuáles son los resultados de la comparación con las posibles soluciones alternativas? 4. ¿Se requieren cambios organizacionales?
Viabilidad del proyecto	1. ¿Existe suficiente compromiso por parte de los actores y las partes interesadas? 2. ¿Pueden estar disponibles los recursos necesarios en términos de tiempo, presupuesto, equipo y dotación de personal? 3. ¿Están disponibles los conocimientos y otras competencias? 4. ¿Son realistas las expectativas en relación al proyecto y sus resultados? 5. ¿Es adecuada la organización del proyecto y su comunicación interna y externa? 6. ¿Existen más riesgos e incertidumbres dentro del proyecto?
Viabilidad técnica	1. ¿Los participantes en el proyecto tienen un conocimiento detallado de las fuentes de información (incluyendo componentes del modelo, información contenida, etc.)? 2. ¿Las fuentes de información tiene un modelo claro y bien definido? 3. ¿Los participantes del proyecto tienen construido y/o utilizan ontologías? 4. ¿Están modeladas las diversas fuentes del proceso de integración conforme al mismo paradigma (entidad-relación / orientación a objetos)? 5. ¿Es oportuno continuar con el proceso de integración si las fuentes de información pertenecen a diferentes áreas geográficas que no se superponen? y/o ¿cuándo las fuentes

	pertenecen a distintas escalas? y/o ¿cuándo existen diversas fuentes que pertenecen a dominios distintos? 6. ¿Están definidas las medidas satisfactorias y cómo comprobar la validez y calidad de los resultados? 7. ¿Qué nivel de complejidad supone la interacción con otros sistemas de información u otros posibles recursos (interoperabilidad, integración de otras fuentes)? ¿Son adecuados y suficientes los métodos y técnicas presentes en el estado de la cuestión? 8. ¿Existen aspectos críticos implicados relativos al tiempo, calidad, recursos necesarios? 9. ¿Hay más riesgos tecnológicos e incertidumbres?
Acciones propuestas	1. Enfoque: ¿Cuál es el enfoque recomendado en las áreas problema-oportunidad? 2. ¿Cuáles son los resultados esperados y beneficios? 3. ¿Qué acciones del proyecto son obligadas de alcanzarse? 4. ¿Cuántas versiones de desarrollo se llevarán a cabo y que componentes serán transformados en cada versión? 5. Riesgos: Si las circunstancias, dentro o fuera de la organización cambian, ¿en virtud de qué condiciones resultaría conveniente reconsiderar las decisiones propuestas?

Técnicas

Lista de comprobación (*ckecklists*). Esta metodología recomienda la utilización de una lista de comprobación basada en los formularios para la toma de decisión de la viabilidad de CommonKADS (Schreiber *et al*., 1999). Esta lista de comprobación es presentada en la Tabla 5.12. El listado de preguntas contenidas no es exhaustivo, por tanto, los responsables participantes podrían incluir alguna pregunta o respuesta que contribuya al proceso de toma de decisión.

Ejemplo práctico

En este ejemplo se responde a las preguntas del listado de comprobación anterior, recogiéndose los resultados en la Tabla 5.13.

Tabla 5.13. Lista de comprobación para la decisión sobre la viabilidad del proceso de integración del conjunto de las bases de datos

Preguntas para decisión sobre viabilidad	
Viabilidad de negocio	1. Este proyecto promueve la integración de diferentes fuentes de información, hasta el momento gestionadas de forma individual, permitiendo aunar el conocimiento de las distintas organizaciones implicadas en torno a la semántica recogida en una ontología global de dominio. Asimismo, esta integración conlleva la eliminación de información duplicada, permitiendo una visión unificada de la realidad geoespacial. 2. El valor añadido generado de la ejecución de este proyecto proviene de la integración de conocimientos (antes aislados) a través del proceso de integración estructural y semántica recogido en la metodología. 3. La ausencia de metodologías de integración de información basada en ontologías y la conformidad con los principios INSPIRE (en lo que respecta al almacenamiento, disponibilidad y mantenimiento de las fuentes de información al nivel más adecuado) hace que, hasta el momento, no haya más solución para el proceso de integración que el recogido en la metodología propuesta. 4. No se requieren cambios organizacionales, debido a que esta metodología es consecuente con los principios INSPIRE mencionados con anterioridad.
Viabilidad del proyecto	1. El compromiso es real por parte de las partes interesadas, especialmente, por parte del IGN. Este compromiso es consecuencia de la necesidad existente de ofrecer al usuario final una visión integrada de la realidad representada a través de IG. 2. Los recursos requeridos están disponibles para la ejecución del proyecto. 3. Los conocimientos para la realización del proyecto están disponibles, ya que se cuenta con los miembros requeridos (ver sección 5.4) y con las fuentes de información que van a formar parte del proceso de integración. 4. Derivado del análisis realizado en el estudio de viabilidad (Actividad 1) y de la realización de diversas pruebas técnicas, a modo de prototipos, se puede afirmar que la

	propuesta de integración de este proyecto así como sus resultados esperados obedecen a un planteamiento realista. 5. La organización del proyecto es adecuada dado el personal implicado en su desarrollo. Además, hay diversos responsables pertenecientes a las distintas organizaciones implicados en el proceso. 6. Del análisis realizado (Actividad 1) no se detecta ningún riesgo que ponga en peligro la ejecución de este proyecto de integración.
Viabilidad técnica	1. No todos los participantes implicados en el proyecto tienen un conocimiento detallado de las fuentes de información, pero esta debilidad queda solventada con la realización de la Actividad 2 propuesta en la metodología. 2. Todas las fuentes de información (bases de datos) tiene un modelo definido. Las diferencias y/o carencias entre modelos queda detallada con la realización de la Actividad 2. 3. El cliente/las diferentes organizaciones implicadas no tiene construido ni utiliza ninguna ontología. 4. Todas las fuentes están modeladas conforme al paradigma entidad-relación. 5. Aunque en el proceso de integración aparecen fuentes vinculadas a diversas áreas geográficas y con diversidad de escalas, también se recogen fuentes con información del conjunto del territorio nacional español. Éstas últimas hacen que el proceso de integración sea factible. Asimismo, aunque existen fuentes con escales similares, se considera que la diversidad de escalas de entre las fuentes recogidas contribuye a enriquecer el proceso de integración. Finalmente, a pesar de que existen fuentes que contienen diversos dominios, como se mencionó con anterioridad, todas las fuentes contienen información sobre el dominio hidrográfico. 6. La metodología establecida define cuales son las técnicas para comprobar la validez y calidad del proceso de integración. Un ejemplo son las medidas de *precision* y *recall* para la evaluación de la Actividad 4. 7. La integración de fuentes adicionales de información en este proyecto estaría supeditada al análisis previo previsto en la Actividad 1, con objeto de comprobar su viabilidad. 8. La obtención de las fuentes de información en el menor tiempo posible y contar con el personal implicado (sección 5.4) son los aspectos más críticos del proyecto. 9. Los posibles riesgos tecnológicos queda minimizados con la realización de los diversos prototipos iniciales y con el listado de herramientas proporcionado por la metodología para la realización de diferentes tareas.
Acciones propuestas	1. Nuestra recomendación es realizar una integración basada en ontologías de las fuentes de información (bases de datos heterogéneas) recogidas. Esta propuesta de integración se realizará en base al proceso de desarrollo recogido en esta metodología. 2. Se prevén grandes beneficios de la ejecución de este proyecto de integración, ya que permitirá la generación de una visión integrada del conocimiento hidrográfico, con semántica explícita y sin información duplicada. 3. El objetivo final es la integración de los diferentes conjuntos de información RDF generados. Para ello, se realizarán todas las actividades/tareas indicadas en la metodología, aunque pueden obviarse (si no se consideran necesarias) aquellas que se establecen como optativas. 4. Esta metodología sigue un ciclo de vida iterativo, por tanto, se realizarán tantas versiones como el cliente considere necesarias, o dicho con otras palabras, se trabajará en las distintas actividades/tareas hasta conseguir resultados que satisfagan al cliente. 5. El principal riesgo proviene de que alguna de las organizaciones pueda realizar importantes cambios en el modelado de sus fuentes de información. Estos, en principio, no repercutirían sobre el estudio de viabilidad realizado (Actividad 1), pero sí podrían repercutir en una nueva realización de la Actividad 2.

5.5.2. Actividad 2. Análisis de fuentes

En esta actividad se profundiza en el análisis de las características de los componentes de las diferentes fuentes de información que forman parte del proceso de integración con el objetivo de obtener una descripción y comparación entre los esquemas y datos. La información obtenida en este proceso de análisis manual ayudará a comprender mejor cada una de las fuentes

de información y contribuirá a identificar similitudes y diferencias que favorecerán el proceso de integración propuesto en esta metodología.

Como ocurre con la Actividad 1, esta actividad y sus respectivas tareas (ver Figura 5.2), están inspiradas por las tareas propuestas en la ingeniería del software, concretamente MÉTRICA versión 3 (MÉTRICA v.3) y *Rational Unified Process* (RUP) (Kruchten, 2000), y por diferentes propuestas de ingeniería del conocimiento, tales como (Gómez-Pérez *et al*., 1997) y (Schreiber *et al*., 1999).

Para llevar a cabo las tareas de descripción se tienen en cuenta tanto aspectos relacionados con el dominio geográfico como otros puramente ligados con las fuentes de información. Así, en esta actividad se puede encontrar información relativa a los tipos de atributos, relaciones, cardinalidad, etc. (información sobre el modelado de las fuentes de información) y metadatos que describen el recurso, su calidad, organización a la que pertenece, etc. (información sobre los datos contenidos en las fuentes de información).

Las tareas de comparación están influidas por la información contenida en las diferentes fuentes de información y por un elemento fundamental en la representación cartográfica como es la escala. El proceso de comparación se lleva a cabo bajo la condición previa de la existencia de solapamientos de información entre fuentes. Una vez seleccionadas las fuentes que cumplen este criterio, el factor escala va a ser el elemento principal del proceso de comparación, ya que, en primer lugar, se compararan las fuentes que posean idéntica escala. Posteriormente se lleva a cabo un proceso de enriquecimiento de la información obtenida mediante un proceso de comparación entre fuentes de diferentes escalas.

Finalmente, en esta actividad resulta fundamental la participación y, consecuentemente, el proceso de retroalimentación de los expertos, ya que éstos contribuirán a identificar y abstraer los problemas detectados en las fuentes de información. Además, su participación ayudará a identificar y/o refinar las posibles mejoras propuestas en la tarea 1.4. Una visión detallada de las diferentes tareas propuestas por esta actividad se muestra en la Figura 5.7.

Las entradas de esta actividad están compuestas por las diferentes fuentes de información, sus metadatos y documentación sobre el dominio. Los resultados obtenidos del conjunto de tareas que conforman esta actividad están constituidos por una descripción de alto nivel de las fuentes de información y de los datos componentes, información resultante del proceso de comparación entre diferencias y semejanzas en base a fuentes y datos y un listado de recomendaciones y propuestas de acción. De manera resumida, en la Tabla 5.14, se reflejan las entradas y salidas, técnicas y participantes asociados a cada una de las tareas. A continuación se procederá a describir en detalle cada una de las tareas recogidas en esta actividad.

Tarea 2.1 Descripción de esquemas

El objetivo de esta tarea consiste en profundizar en el análisis de los esquemas de las diferentes fuentes de datos. Para esto, se lleva a cabo el análisis de cuatro aspectos fundamentales del modelado de los esquemas de las bases de datos desde la óptica del proceso de integración:

- *Análisis de las tablas* que conforman las diversas fuentes de información (número de tablas, aparecen nombres descriptivos de los fenómenos contenidos, contienen un único tipo de fenómeno o aparece una miscelánea dentro de cada tabla, etc.).
- *Análisis de los componentes especificados* que caracterizan a los diferentes tipos de fenómenos dentro del dominio de información en cada tabla (número de campos (columnas/atributos) cada tabla, identificación de claves primarias y/o extranjeras, tipos de datos asociados, etc.).
- *Análisis de las relaciones* y sus características especificadas (tipos de relaciones y cardinalidad) entre las distintas tablas de las fuentes de información.

- Análisis, si procede, del tratamiento o consideración del *aspecto multilingüe* en el modelado de las bases de datos.

La técnica utilizada en esta tarea consiste en la descripción de componentes de los esquemas de las diferentes fuentes de información. Las características técnicas que conlleva esta tarea provoca la implicación directa del ingeniero software del lado ejecutor del proyecto de integración. Mientras, por parte de la organización, el responsable también será el ingeniero software encargado de la administración de las bases de datos, aunque en este caso la participación se centrará en el apoyo al responsable del lado ejecutor ante cualquier evento. Asimismo, algunos datos pueden proceder de otros tipos de fuentes de información, es decir, no necesariamente de bases de datos, por ejemplo: hojas de cálculo, XML, etc. En esta tarea, la utilización de herramientas como TOAD[1] o SQuirreL SQL[2] pueden facilitar el trabajo llevado a cabo.

Material de entrada y resultados de salida

Entradas: Fuentes de información y metadatos identificados.
Salidas: Descripción de alto nivel de las fuentes de información.

[1] *TOAD.* http://www.toadsoft.com/
[2] *SQuirreL*. http://www.squirrelsql.org/

Tabla 5.14. Productos de tareas, técnicas y participantes de la actividad 2 (Análisis de fuentes)

N° Tarea	Nombre Tarea	Responsable	Producto	Técnicas aplicables
Fuentes Tarea 2.1	Descripción de esquemas	Ingeniero Software	Descripción de alto nivel de las fuentes de información	Descripción de componentes de los esquemas de las fuentes de información
Fuentes Tarea 2.2	Comparativa entre esquemas	Ingeniero Software	Comparación de las diferencias y semejanzas de las fuentes de información	Catalogación y comparación
Datos Tarea 2.3	Descripción de datos	Experto de dominio / Ingeniero Software	Descripción de alto nivel de los datos componentes de las fuentes de información	Descripción de los datos componentes de la bases de datos
Datos Tarea 2.4	Comparativa entre datos	Experto de dominio / Ingeniero Software	Comparación de las diferencias y semejanzas de los datos de las fuentes de información	Catalogación y comparación
Tarea 2.5	Revisión basada en retroalimentación de expertos	Experto del dominio/ Ingeniero Software / Ingeniero de conocimiento	Propuestas de cambios/mejoras con respecto a fuentes y datos para envío a los respectivos administradores	Entrevistas y Evaluación

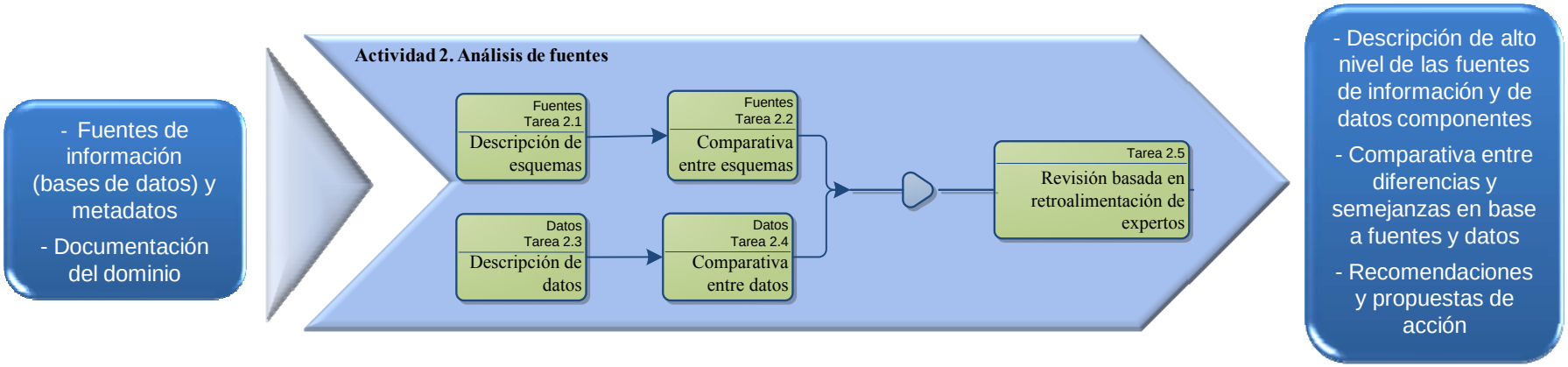


Figura 5.7. Descomposición de tareas de la actividad 2 (Análisis de fuentes)

Responsables

Ejecutor del proyecto: Ingeniero software.
Organización: Ingeniero software.

Técnicas

Descripción de componentes de los esquemas de las fuentes de información.

Ejemplo práctico

Un ejemplo detallado de la descripción de los esquemas de las bases de datos correspondientes a BTN25, BCN200, Atlas Nacional, Nomenclátor Conciso y Nomenclátor NOMGEO aparece en el Anexo.

Tarea 2.2 Comparativa entre esquemas

El punto de partida de la comparativa propuesta en esta tarea se toma de la recopilación y descripción de las fuentes de información (tarea 1.6). Este hecho es consecuencia de que el proceso de comparación se realiza entre fuentes con solapamientos de información parcial o total. Una vez seleccionadas las fuentes de información que contemplen este requisito el proceso de comparación se lleva a cabo en dos fases distintas. En primer lugar, se comparan esquemas de fuentes que tengan idéntica escala de recogida de información. Finalizado este proceso, se procede en una segunda fase, con el objetivo comprobar y enriquecer las diferencias y semejanzas entre diferentes esquemas, a un proceso de comparación del tratamiento de los esquemas desde la óptica enriquecedora de la visión transescalar del modelado de la información geográfica.

Consideradas las pautas iniciales del proceso de comparación de esquemas pertenecientes a las diferentes fuentes de información que constituyen el proceso de integración, los aspectos objeto del proceso de comparación son los analizados en la tarea anterior (tarea 2.1), es decir, tablas, componentes especificados, relaciones y la consideración de la multiligualidad. Consecuencia de que esta tarea es una continuación de la anterior (tarea 2.1) resulta evidente la implicación del ingeniero software por ambas partes implicadas en el proceso de integración. Las técnicas empleadas por estos responsables consisten en la catalogación y comparación de los diferentes esquemas de las diversas fuentes con el fin de obtener semejanzas y diferencias entre las mismas.

Material de entrada y resultados de salida

Entradas: Descripción de alto nivel de las fuentes de información.
Salidas: Comparación de las diferencias y semejanzas de las fuentes de información.

Responsables

Ejecutor del proyecto: Ingeniero software.
Organización: Ingeniero software.

Técnicas

Catalogación y comparación de los diferentes esquemas de las fuentes que componen el proceso de integración.

Ejemplo práctico

En el Anexo se presenta un ejemplo detallado de la comparación entre los esquemas de dos bases de datos con escalas distintas (Nomenclátor NOMGEO y Confederación Hidrográfica del Ebro).

Tarea 2.3. Descripción de datos

Esta tarea tiene por objetivo la descripción de los datos contenidos en las diferentes fuentes de información. Para alcanzar este objetivo se considera el análisis y descripción de ciertos parámetros cartográficos, características de los datos y la consideración del multilingüismo asociado a los datos de las diversas fuentes de información.

- *Parámetros cartográficos*. La descripción de los parámetros cartográficos de los datos se lleva a cabo mediante un listado de metadatos, reflejados en la Tabla 5.15, conformes al perfil del Núcleo Español de Metadatos (NEM) (Subgrupo de Trabajo del Núcleo Español de Metadatos, 2005). Esto supone que los metadatos contemplados para la descripción global del conjunto de los datos están formados por la combinación de *Dublin Core Metadata Initiative*, ISO 19115 (ISO, 2003) y otros elementos propuestos por este perfil de metadatos.
- *Características de los datos*. Se utilizan para enriquecer la descripción de los datos de las diversas fuentes de información. De esta manera, se recogerá información sobre las unidades de medida utilizadas, implementación de relaciones topológicas y sus repercusiones y la identificación, caracterización y justificación de ausencias u omisiones y/o excesos o comisiones de información.
- La *consideración del multilingüismo*, aspecto opcional que depende de las características de la información contenida en las fuentes, pretende realizar un análisis del tratamiento de esta información a nivel de los datos.

La técnica empleada en esta tarea está centrada en la descripción de los datos de las bases de datos con el fin de obtener una descripción de alto nivel de dichos datos. De esta manera, los responsables implicados, por ambas parte del proyecto, es el experto de dominio que son los encargados del análisis y descripción de los parámetros cartográficos y del multilingüismo asociado a los datos. Asimismo, por parte del ejecutor del proyecto también está implicado el ingeniero de software. Su vinculación con esta tarea es puramente testimonial, aunque resulta conveniente su implicación en la misma dada su participación en las tareas anteriores (2.1 y 2.2) y su intervención en la siguiente (tarea 2.4). Al igual que en la tarea 2.1, la utilización de herramientas como TOAD o SQuirreL SQL pueden facilitar el trabajo a realizar.

Material de entrada y resultados de salida

Entradas: Fuentes de información (bases de datos) y documentación sobre el dominio.
Salidas: Descripción de alto nivel de los datos componentes de las fuentes de información.

Responsables

Ejecutor del proyecto: Experto de dominio e ingeniero de software.
Organización: Experto de dominio.

Técnicas

Descripción de los datos componentes de las bases de datos.

Ejemplo práctico

Un ejemplo con la descripción de los datos propuesta en esta tarea (Parámetros cartográficos, características de los datos y consideración del multilingüismo) se recoge en el Anexo de este trabajo.

Tarea 2.4 Comparativa entre datos

Al igual que ocurre en la tarea destinada a la comparativa entre esquemas (tarea 2.2) el punto de partida de la comparativa propuesta en esta tarea se toma de la recopilación y descripción de las fuentes de información (tarea 1.6), consecuencia de que el proceso de comparación se realiza entre fuentes con solapamientos de información. Asimismo, al igual que

se describe con anterioridad, el proceso de comparación se lleva a cabo, en primer lugar, entre fuentes que tienen idéntica escala de recogida de información y que, además, presenten idénticas características respecto a ciertos parámetros cartográficos identificados en la colección de metadatos, tales como: sistema de referencia, información del datum vertical, resolución y exactitud posicional. En segundo lugar, con el objetivo de enriquecer el proceso desde la óptica que ofrece la visión de la información geográfica a diferentes escalas, se procede a un proceso de comparación entre las fuentes de información de diversas escalas caracterizadas por poseer idénticos sistemas de referencia, información del datum vertical y exactitud posicional.

Junto a los parámetros cartográficos (metadatos) descriptores del conjunto de los datos, recogidos en la Tabla 5.15, esta tarea considera un proceso de análisis y comparación de aspectos léxico-sintácticos, unidades de medida, ausencias generalizadas y/o lagunas puntuales de información (omisiones), exceso de información (comisiones) y presencia-tratamiento de información multilingüe. La finalidad de la consideración de este amplio abanico de aspectos es llevar a cabo una comparación detallada de las diferencias y semejanzas de los datos que conforman las diferentes fuentes de información.

La miscelánea entre aspectos geográficos y otros aspectos más técnicos vinculados a los datos provoca que a esta tarea estén vinculados tanto expertos de dominio como ingenieros de software por las partes implicadas en el proceso de integración. Las técnicas empleadas por estos responsables se ciñen a la catalogación y comparación de los datos componentes de las diversas fuentes de información para obtener las diferencias y semejanzas entre los mencionados datos.

Material de entrada y resultados de salida

Entradas: Descripción de alto nivel de los datos componentes de las fuentes de información.

Salidas: Comparación de las diferencias y semejanzas de los datos de las fuentes de información.

Responsables

Ejecutor del proyecto: Experto de dominio e ingeniero de software.

Organización: Experto de dominio e ingeniero software.

Técnicas

Catalogación y comparación de los datos componentes de las diversas fuentes de información que conforman el proceso de integración.

Ejemplo práctico

Una referencia al trabajo de comparación entre datos propuesta en esta tarea aparece en el Anexo. Asimismo, en este anexo aparece una recopilación detallada de la comparación entre datos y un listado exhaustivo de los problemas resultantes de dicha comparación aparece en la sección 8.2 de este trabajo.

Tarea 2.5. Revisión basada en retroalimentación de expertos

En esta tarea se exponen ante los expertos las conclusiones de los resultados obtenidos en las tareas concernientes a los procesos de comparación entre esquemas (tarea 2.2) y datos (tarea 2.4). La interacción con los expertos de la organización diversifica los objetivos para esta tarea. Así, por un lado, el objetivo es que los expertos contribuyan en el proceso de interpretación de resultados obtenidos, aportando una visión más enriquecida al proceso de análisis. Por otro lado, el objetivo es contribuir al proceso de abstracción necesario para obtener las propuestas de cambio necesarias en determinadas fuentes de información, procediendo al refinamiento de la tarea 1.4. Debido a que esta propuesta metodológica de integración de información de bases de datos se basa en el principio de distribución y mantenimiento de las fuentes de información por

parte de las propias organizaciones productoras, las propuestas de mejora serán sólo recomendaciones de actuación destinadas a la organización propietaria de la fuente. Finalmente, se lleva a cabo la plasmación en un informe de las lecciones aprendidas obtenidas del proceso de análisis de las fuentes de información y de la retroalimentación de los expertos.

Tabla 5.15. Metadatos sobre los datos contenidos en las fuentes de información

Metadatos geográficos
Recurso *(datos)*
• Título
• Nombre de la organización (responsable)
• Resumen descriptivo del conjunto de los datos
• Localización geográfica del conjunto de los datos
• Extensión geográfica *(Boundary Box)*
• Propósito
• Fuente
• Escala equivalente
• Descripción del nivel (Descripción detallada sobre el nivel de los datos del ámbito)
• Tipo de representación espacial
• Fecha de creación (de los metadatos)
• Fecha (de referencia para el recurso mencionado)
• Revisión
• Formato de distribución
• Fecha de referencia para el conjunto de los datos
• Idioma del conjunto de los datos
• Categoría del tema del conjunto de los datos
• Sistema de referencia
• Identificador de sistema de referencia
• Información del Datum vertical
• DQ_Completeness (Presencia y ausencia de entidades, de sus atributos y sus relaciones)
Calidad
• Resolución
• Resolución espacial del conjunto de los datos
• Exactitud Posicional Externa Absoluta
• Descripción de la medida
• Resultado de la medida
• Calidad de los datos
• Exactitud posicional
• Consistencia topológica
• Identificador geográfico
• Información de agregación
• Exactitud de los atributos no cuantitativos
• Exactitud de los atributos cuantitativos
• Resultado cuantitativo
• Paso de proceso
• Información sobre la calidad de los datos
Organización
• Autor
• Conservador
• Parte responsable del conjunto de los datos

La técnica manejada en esta tarea es la realización de reuniones a las que asistirán, por parte de la organización, el experto de dominio e ingeniero de software. Mientras, por parte del ejecutor del proyecto, asistirán idénticos responsables y, además, se unirá la participación del ingeniero de conocimiento con vistas a comprobar la situación general de las fuentes de información, una vez finalizado el proceso de análisis, y comenzar la planificación de las tareas correspondientes a la actividad de integración semántica.

Material de entrada y resultados de salida

Entradas: Comparativa de las diferencias y semejanzas de las fuentes de información y datos pertenecientes a las mismas.

Salidas: Propuestas de cambios/mejoras con respecto a fuentes y datos para envío a los respectivos administradores.

Responsables

Ejecutor del proyecto: Experto del dominio, ingeniero software e ingeniero de conocimiento.

Organización: Experto de dominio e ingeniero de software.

Técnicas

Entrevistas (reuniones) y evaluación.

Ejemplo práctico

Los procesos de análisis y comparación llevados a cabo en esta actividad finalizan con una revisión por parte de los expertos cuyo objetivo es obtener retroalimentación. Esto conlleva, previamente, un proceso de información a estos expertos con los resultados obtenidos en los procesos anteriores. En este ejemplo (ver Tabla 5.16) se muestra un listado de errores léxicos obtenidos de diversas fuentes del IGN.

Tras las sesiones de entrevistas y evaluación por parte de los expertos se aportan las siguientes sugerencias de cambio (Ver Tabla 5.17), con el compromiso de la organización de realizar estos cambios identificados.

Tabla 5.16. Errores léxicos

Error
Ca+ada
Ca+o
Rac3/4
Rio

Tabla 5.17. Sugerencias de cambio

Error	Sugerencia
Ca+ada	Cañada
Ca+o	Caño
Rac3/4	-
Rio	Río

Una descripción detallada de los diferentes problemas detectados se lista en la sección 8.2 de este trabajo.

5.5.3. Actividad 3. Integración estructural

El proceso de realización de esta actividad viene marcado por la existencia de recursos previos sobre el dominio y los niveles de formalización de los mismos (tarea 1.5). Este hecho provoca que dentro del proceso de desarrollo de esta propuesta metodológica el conjunto de esta actividad sea opcional. Por tanto, la opcionalidad de esta tarea va a ser manifiesta en caso de que existan recursos ontológicos accesibles, tanto en la organización como fuera de la misma. Asimismo, estos recursos deben ajustarse a las fuentes de información y al propósito establecido para el proceso de integración. En caso de no existir recursos ontológicos o que estos no

Tabla 5.18. Productos de tareas, técnicas y participantes de la actividad 3 (Integración estructural)

Nº Tarea/ Actividad	Nombre Tarea	Responsable	Producto	Técnicas aplicables
Actividad 3.1	Desarrollo de ontologías	Experto del dominio / Ingeniero de conocimiento	Ontología individual, ontologías individuales interconectadas o redes de ontologías.	Dependientes de la metodología seleccionada
Tarea 3.2	Evaluación	Experto de dominio / Ingeniero de conocimiento	Ontología individual, ontologías individuales interconectadas o red de ontologías de dominio formalizadas en lenguaje de ontologías verificadas formalmente respecto al lenguaje utilizado y evaluadas en relación a su consistencia lógica, requisitos y cuestiones de competencia de la ontología.	Verificación formal del lenguaje, evaluación de la consistencia lógica, requisitos y cuestiones de competencia de la ontología y prueba en la aplicación de destino.

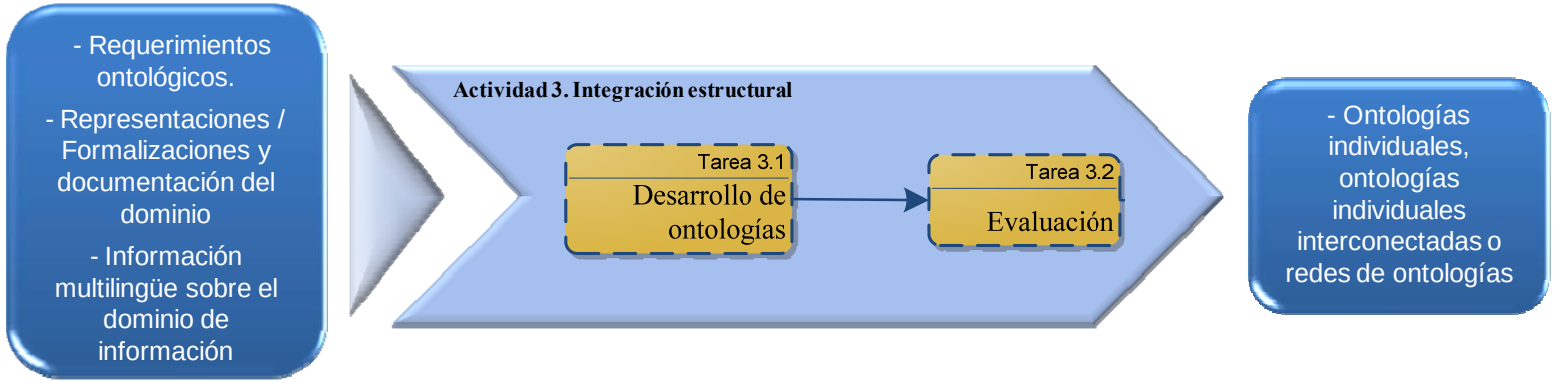


Figura 5.8. Descomposición de tareas de la actividad 3 (Integración estructural)

resultaran reutilizables, consecuencia del propósito fijado en el proceso de integración, se procederá al desarrollo de los diferentes pasos marcados en esta actividad.

El objetivo de esta actividad es el desarrollo y evaluación de ontologías que sirvan de instrumento integrador de los diferentes esquemas propuestos en las diversas fuentes de información que conforman el proceso de integración. Esta actividad aporta una gran flexibilidad a la propuesta metodológica realizada en este trabajo, ya que en función de las necesidades de la organización (identificadas en la tarea 1.3) la propuesta de integración estructural de las diferentes fuentes de información construirá un tipo diferente de ontologías y, por tanto, se adoptará una metodología de construcción de ontologías u otra distinta. De manera obvia, esta actividad y sus tareas, como puede apreciarse en la Figura 5.2, están inspiradas por las tareas propuestas en la ingeniería ontológica, concretamente, están basadas en diferentes propuestas metodológicas para la construcción de ontologías, tales como (Staab *et al*., 2001; Gómez-Pérez *et al*., 2003; Pinto *et al*., 2004; Suárez-Figueroa *et al*., 2008).

En esta actividad la evaluación del recurso ontológico está inspirada en las propuestas de (Gangemi *et al*., 2006; Gómez-Pérez, 2004). Así, en esta tarea de evaluación se considera la verificación formal del lenguaje, evaluación de la consistencia lógica, preguntas de competencia y requisitos.

Dadas las características de esta actividad, las entradas de la misma van a depender del propósito del proyecto de integración de información y, por tanto, de la propuesta metodológica seleccionada para el desarrollo de ontologías. Con respecto a los resultados, al igual que las entradas irán en función de la metodología seleccionada, aunque, de manera general, los productos finales de esta actividad consistirán en ontologías individuales, ontologías individuales interconectadas o redes de ontologías de dominio formalizadas en algún lenguaje de ontologías y evaluadas en relación a su consistencia lógica, requisitos y cuestiones de competencia de la ontología. De manera resumida, en la Tabla 5.18, se reflejan las entradas y salidas, técnicas y participantes asociados a cada una de las tareas. Además, en la Figura 5.8 se muestra una visión de las diferentes tareas propuestas por esta actividad. A continuación se describen en detalle cada una de las tareas recogidas en esta actividad.

Actividad 3.1 Desarrollo de ontologías

El objetivo de esta tarea es el desarrollo de ontologías que sirvan de instrumento integrador para las diferentes fuentes de información que conforman el proceso de integración. La decisión sobre la tipología de ontologías y su proceso de desarrollo conforman una serie de pasos directamente influidos por las necesidades y el propósito que originan la propuesta de integración de información. Este hecho motiva que esta tarea considere un amplio abanico de propuestas metodológicas, de cuya elección se determinará tanto la tipología de ontología como el proceso de desarrollo.

Los distintos escenarios que pueden presentarse ante las necesidades y propósitos del proceso de integración de información provocan que la clasificación propuesta en (Suárez-Figueroa *et al*., 2008) se ajuste mejor al abanico de posibilidades que pueden darse en el proceso de desarrollo de ontologías. Asimismo, vinculado con el objetivo de esta tarea, es decir, la integración estructural de las fuentes de información, esta clasificación posee cierto paralelismo con las diferentes vías de integración de información basadas en ontologías propuestas en (Wache *et al.*, 2001). Así, el tipo de *ontologías individuales* tiene su paralelismo en el *enfoque basado en una única ontología*, ya que ambas propuestas únicamente desarrollan/utilizan una ontología. Las *ontologías individuales interconectadas* se asemejan al *enfoque basado en múltiples ontologías*. En este caso, las ontologías están constituidas de una combinación y/o relación con otras ontologías. Por último, las *redes de ontologías* están vinculadas con los *enfoques híbridos*, debido a que estos conforman una colección de múltiples ontologías relacionadas entre sí mediante diferentes tipos de relaciones.

Para cada una de los tipos de ontologías, mencionados con anterioridad, existen diferentes propuestas metodológicas para su construcción. Una descripción exhaustiva de diferentes métodos y metodologías existentes puede ser encontrada en (Gómez-Pérez *et al*., 2003). Para esta tarea se seleccionan las metodologías más reconocidas, completas y extendidas (*METHONTOLOGY* (Gómez-Pérez *et al*., 2003), *On-To-Knowledge* (Staab *et al*., 2001), *DILIGENT* (Pinto *et al*., 2004) y *NeOn methodology* (Suárez-Figueroa, 2010)).

Las técnicas utilizadas en esta tarea son dependientes del tipo de metodología elegida, ya que cada metodología propone unas técnicas determinadas a seguir. Con respecto a los responsables que participan en esta tarea, por parte del ejecutor del proyecto habrá una implicación directa del experto de dominio y del ingeniero de conocimiento. Mientras, por el lado de la organización en la que se lleva a cabo el proceso de integración el responsable directamente implicado será el experto de dominio. Asimismo, resultaría conveniente la participación, aunque de una manera más puntual, de un ingeniero de conocimiento por parte de la organización, con el fin de tener una implicación directa en ciertas tareas del proceso de desarrollo de ontologías. En esta tarea pueden utilizarse diferentes herramientas en el proceso de modelado del dominio (por ejemplo; Microsoft VISIO, Rational Rose, MINDMAP, FreeMind, etc.) y en el desarrollo de ontologías (por ejemplo; Protégé, NeOn Toolkit, Top Braid Composer, etc.)

Material de entrada y resultados de salida

Entradas: Dependientes de la propuesta metodológica seleccionada y/o prototipos generados (tarea 1.8).

Salidas: Ontología individual, ontologías individuales interconectadas o redes de ontologías.

Responsables

Ejecutor del proyecto: Experto de dominio e ingeniero de conocimiento.

Organización: Experto de dominio e (ingeniero de conocimiento).

Técnicas

En esta tarea las técnicas empleadas son dependientes de la propuesta metodológica seleccionada para el proceso de desarrollo de ontologías (Fernández-López *et al.,* 1999; Staab *et al*., 2001; Pinto *et al*., 2004; Suárez-Figueroa *et al*., 2008), ya que cada metodología utiliza sus técnicas correspondientes.

Ejemplo práctico

Un ejemplo detallado sobre el proceso de construcción de ontologías aparece en el Capítulo 6 de este trabajo.

Tarea 3.2 Evaluación

El objetivo de esta tarea es llevar a cabo un proceso de verificación y evaluación de las ontologías desarrolladas. Esta tarea viene a cerrar los procesos de desarrollo de las diferentes propuestas metodológicas, consecuencia de que muchas de estas, aún teniendo entre sus tareas procesos de evaluación, no contemplan una evaluación completa de la ontología, ya que de forma generalizada se ciñen a una evaluación centrada en los requisitos y/o cuestiones de competencia de la ontología. De esta manera, la evaluación propuesta en esta tarea contempla la verificación formal del lenguaje (por ejemplo, a través de herramientas como *OWL validator*[1]), evaluación de la consistencia lógica (a través de la utilización de razonadores como, por ejemplo, Pellet[2], FaCT++[3] o Racer[4]), evaluación conforme a los requisitos y cuestiones de competencia.

[1] http://www.mygrid.org.uk/OWL/Validator
[2] http://www.mindswap.org/2003/pellet/
[3] http://owl.man.ac.uk/factplusplus/

Las características de esta tarea suponen la implicación directa por parte del ejecutor del proyecto, del ingeniero de conocimiento en todos los elementos de esta tarea, aunque especialmente en los aspectos más formales de la misma, y del experto de dominio, cuya función estará ligada al proceso de evaluación de la ontología conforme a los requisitos, cuestiones de competencia y a la realización de la prueba en la aplicación de destino. Por parte de la organización, la implicación en esta tarea será más secundaria en lo que a la participación del ingeniero de conocimiento se refiere, mientras que el experto de dominio tendrá un papel fundamental en tareas de apoyo y supervisión del trabajo del experto de dominio por parte del ejecutor del proyecto.

Material de entrada y resultados de salida

Entradas: Ontologías individuales, ontologías individuales interconectadas o redes de ontologías.

Salidas: Ontología individual, ontologías individuales interconectadas o red de ontologías de dominio formalizadas en lenguaje de ontologías verificadas respecto al lenguaje utilizado y evaluadas en relación a su consistencia lógica y requisitos de la ontología.

Responsables

Ejecutor del proyecto: Experto de dominio e ingeniero de conocimiento.

Organización: Experto de dominio e (ingeniero de conocimiento).

Técnicas

Verificación formal del lenguaje, evaluación de la consistencia lógica, requisitos y cuestiones de competencia de la ontología individual, ontologías individuales interconectadas o red de ontologías.

Ejemplo práctico

En la Figura 5.9 se muestra una parte del proceso de evaluación de los recursos ontológicos desarrollados en el contexto de este trabajo. En este caso, mediante la utilización de la herramienta *OWL validator* se realiza la verificación formal del lenguaje de *hydrOntology* (descrita en detalle en el Capítulo 6). Esta herramienta informa de la expresividad utilizada en la formalización de la ontología (OWL Lite, DL o Full) (McGuinness *et al.,* 2004) y muestra un resumen de los distintos elementos que la componen. En el caso de esta ontología (*hydrOntology*), esta herramienta muestra que su expresividad es OWL Full.

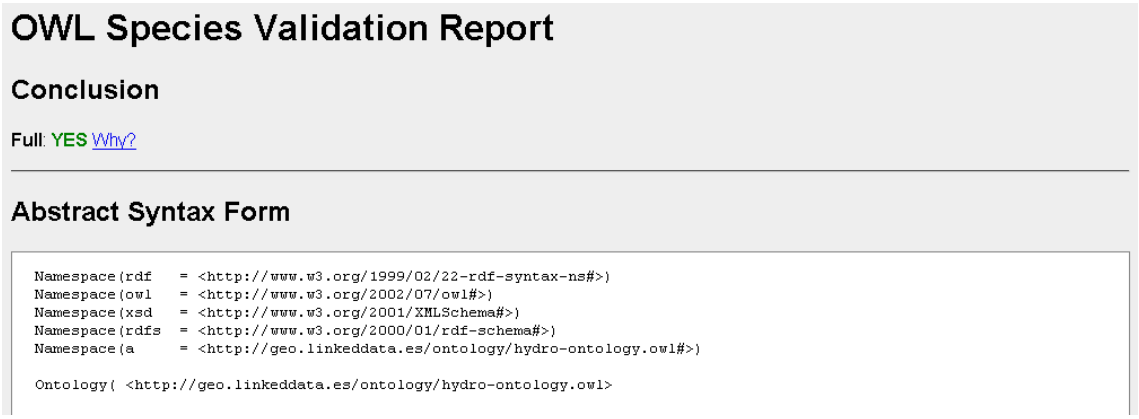


Figura 5.9. Verificación de *hydrOntology*

5.5.4. Actividad 4. Integración semántica

El objetivo de esta actividad es la transformación de las fuentes de información a conjuntos de datos RDF mediante el establecimiento de correspondencias (*mappings*) entre la ontología

[4] http://www.racer-systems.com/

individual, ontologías individuales interconectadas o red de ontologías desarrolladas procedentes de la actividad 3 (integración estructural) o de los recursos existentes localizados en la tarea 1.5 (nivel de formalización del dominio) y las diversas bases de datos. Con objeto de realizar un refinamiento o completitud de las correspondencias previamente establecidas (tarea 4.2) se requiere la participación del experto de dominio para realizar un proceso de depuración y/o mejora. Posteriormente, se procede a la fase de explotación de *mappings* (tarea 4.3) mediante la utilización de una aplicación de explotación de *mappings* y la ejecución de pruebas en la aplicación de destino. Finalmente, se llevará a cabo un proceso de evaluación de los *mappings* establecidos a través de los elementos de medida *precision* y *recall*. Esta actividad y sus respectivas tareas, como se refleja en la Figura 5.2, están inspiradas en las diversas actividades propuestas en los procesos de *ontology matching* (Euzenat *et al.*, 2007).

Las entradas a esta actividad están constituidas por las fuentes de información y la ontología individual, ontologías individuales interconectadas o redes de ontologías de dominio verificadas formalmente respecto al lenguaje utilizado y evaluadas en relación a su consistencia lógica, requisitos y cuestiones de competencia de la ontología. Los resultados de esta actividad consistirán en un/os conjunto/s de *mappings* (entre la ontología individual, ontologías individuales interconectadas o red de ontologías y las fuentes de información) y la verificación de los *mappings* establecidos y/o propuestas de cambio surgidas de la interrelación con los expertos. En la Tabla 5.19, de manera resumida, se reflejan las entradas, salidas, técnicas y participantes asociados a cada una de las tareas. Asimismo, en la Figura 5.10 se muestra una visión general de las diferentes tareas propuestas por esta actividad. A continuación se describen en detalle cada una de las tareas recogidas en dicha actividad.

Tarea 4.1 Identificación de datos

Una de las principales decisiones, previas al proceso de transformación de las fuentes de información a RDF (descrito en la tarea 4.2), es el formato o patrón en que los identificadores de las instancias (URI) van a ser generados. Las URI son extremadamente relevantes en este proceso, ya que éstas contribuirán de manera clave en el alineamiento de instancias provenientes de diferentes fuentes de información. Por ello, el objetivo de esta tarea es la generación de un patrón de URI para el conjunto de datos estudiados. La realización de esta tarea da conformidad con el principio de *Linked Data* (Berners-Lee, 2006) que propone la utilización de URI para identificar cosas.

En el desarrollo de dicha tarea se considera información de referencia las recomendaciones y buenas prácticas recogidas en (Ayers *et al*., 2008; Davidson, 2009). En esta tarea están implicados todos los participantes en el proyecto, que sobre la base de las recomendaciones y prácticas mencionadas deben aportar sus ideas para la generación del patrón de URI en una sesión de tormenta de ideas. De esta sesión deberá salir el patrón y las respectivas decisiones de diseño adoptadas.

Material de entrada y resultados de salida

Entradas: Documentación sobre diseño de URI.

Salidas: Diseño del patrón de URI para las fuentes de información y documento descriptivo de las decisiones adoptadas.

Responsables

Ejecutor del proyecto: Jefe de proyecto, experto del dominio, ingeniero software e ingeniero de conocimiento.

Organización: Jefe de proyecto, experto del dominio, ingeniero software e ingeniero de conocimiento.

Tabla 5.19. Productos de tareas, técnicas y participantes de la actividad 4 (Integración semántica)

Nº Tarea	Nombre Tarea	Responsable	Producto	Técnicas aplicables
Tarea 4.1	Identificación de datos	Jefe de proyecto / Experto del dominio/ Ingeniero Software / Ing. de conocimiento	Diseño del patrón de URI para las fuentes de información y documento descriptivo de las decisiones adoptadas	Análisis de documentación, tormenta de ideas y descripción de las decisiones adoptadas
Tarea 4.2	Creación de *mappings*	Ingeniero de conocimiento	Conjunto de *mappings* entre la ontología individual, ontologías individuales interconectadas o red de ontologías y las fuentes de información (bases de datos)	Formalización de *mappings* entre la ontología individual, ontologías individuales interconectadas o red de ontologías y las diferentes bases de datos que componen el proceso de integración de información.
Tarea 4.3	Generación de instancias	Ingeniero software / Experto de dominio	Conjuntos de datos RDF de las distintas fuentes	ODEMapster, D2R server, librería GEOMETRY2RDF, etc. para la generación de instancias y su información asociada.
Tarea 4.4	Revisión y retroalimentación de expertos	Ingeniero de conocimiento / Experto de dominio	Verificación de *mappings* establecidos y/o propuesta de cambios	Consulta de *mappings* establecidos y sesiones de entrevistas
Tarea 4.5	Evaluación	Experto de dominio e ingeniero software	Estudio de la calidad de los resultados (RDF) generados	*Precision & recall*

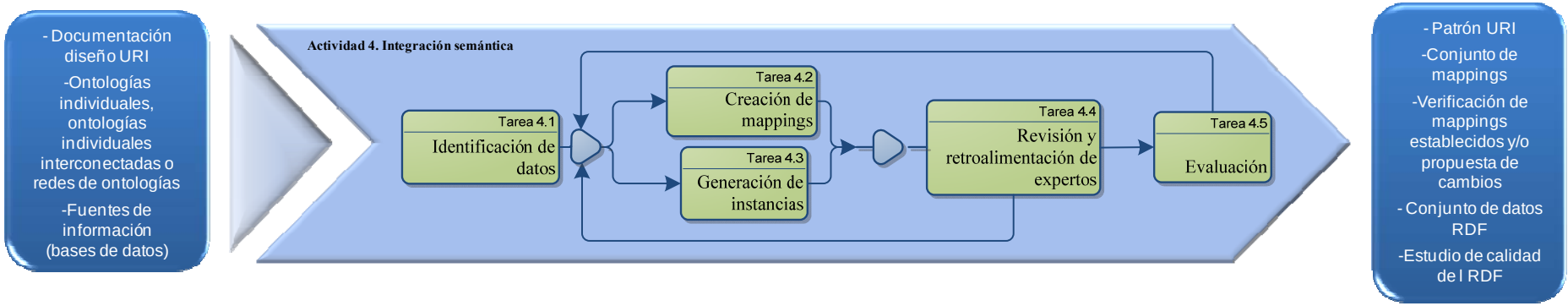


Figura 5.10. Descomposición de tareas de la actividad 4 (Integración semántica)

Técnicas

Análisis de documentación de referencia, tormenta de ideas, descripción del patrón de URI, así como las decisiones de diseño adoptadas en el contexto del proyecto.

Ejemplo práctico

La descripción del patrón de URI adoptado para la identificación de los datos geoespaciales considerados en este trabajo se recoge en el Capítulo 7 (sección 7.1) de este trabajo.

Tarea 4.2 Creación de correspondencias (mappings)

En esta tarea el objetivo es llevar a cabo el establecimiento de correspondencias entre los diferentes modelos relacionales provenientes de las bases de datos con el recubrimiento conceptual presente en las ontologías derivadas tanto de la actividad 3 como de la recopilación extraída en la tarea 1.5. El establecimiento de dichas correspondencias se realiza mediante la utilización de lenguajes como R_2O (Barrasa, 2007), D2R (Bizer, 2003), etc.

En esta tarea, la utilización de herramientas como ODEMapster, D2R Server o Triplify pueden facilitar el trabajo a realizar.

Los responsables implicados en esta tarea del proyecto son los ingenieros de conocimiento, ya que esta tarea pertenece a su dominio de acción. Asimismo, las técnicas utilizadas irán destinadas a la formalización de los *mappings*, mediante la utilización de algunos de los lenguajes mencionados, entre la ontología individual, ontologías individuales interconectadas o red de ontologías y las diferentes bases de datos que componen el proceso de integración de información.

Material de entrada y resultados de salida

Entradas: Ontologías individuales, ontologías individuales interconectadas o red de ontologías de dominio formalizadas en algún lenguaje de ontologías. Asimismo, las fuentes de información (bases de datos) son otra entrada fundamental para esta tarea.

Salidas: Conjunto de *mappings* entre la ontología individual, ontologías individuales interconectadas o redes de ontologías y las fuentes de información.

Responsables

Ejecutor del proyecto: Ingeniero de conocimiento.

Organización: Ingeniero de conocimiento.

Técnicas

Formalización de *mappings* entre la ontología individual, ontologías individuales interconectadas o red de ontologías y las diferentes bases de datos que componen el proceso de integración de información. Para una descripción detallada de las técnicas existentes en el contexto del *ontology matching* ver (Euzenat *et al.*, 2007).

Ejemplo práctico

En este ejemplo (ver Tabla 5.20) se muestra un extracto de los *mappings*, formalizados en lenguaje R_2O, establecidos entre la base de datos de EuroGlobalMap e *hydrOntology*. Concretamente, en este ejemplo se muestra el *mapping* entre el concepto "Acequia" de la ontología y el campo `NAMN1` de la tabla `WATRCRSL` perteneciente a la mencionada base de datos. Además, el código refleja el establecimiento de *mappings* entre atributos, concretamente, el referido a "nombre alternativo".

En el Capítulo 7 (sección 7.2.2) se recoge una descripción detallada de los diversos *mappings* generados entre *hydrOntology* y múltiples bases de datos.

Tarea 4.3 Generación de instancias

El objetivo de esta tarea es la generación de instancias de la información geográfica almacenada en diversas bases de datos geoespaciales. Para la obtención de dichas instancias y sus datos geográficos se pueden utilizar diferentes herramientas, tales como: ODEMaspter, D2R server, librería GEOMETRY2RDF[135], etc.

Los responsables implicados en esta tarea, por ambos lados, son los ingenieros de software, consecuencia del trabajo de implementación y ejecución de pruebas que se lleva a cabo en esta tarea. Asimismo, los expertos de dominio pueden tener una participación puntual en esta tarea, ya que su conocimiento puede ser necesario durante el desarrollo de esta tarea.

Tabla 5.20. Ejemplo de codificación de *mappings* R_2O entre *hydrOntology* y EuroGlobalMap

```
<?xml version="1.0" encoding="UTF-8"?>
<r2o>
 <conceptmap-def name="http://www.oeg-upm.net/ontology/hydrOntology.owl#
Acequia">
        <uri-as>
            <operation oper-id="concat">
                <arg-restriction on-param="string1">
                                    <has-value>http://www.oeg-
                    upm.net/ontology/hydrOntolo gy.owl#</has-value>
                </arg-restriction>
                <arg-restriction on-param="string2">
                    <has-column>egm.watrcrsl.NAMN1</has-column>
                </arg-restriction>
            </operation>
        </uri-as>
        <described-by>
            <attributemap-def name="www.oeg-upm.net/ontology/hydrOntology.ow
            l#nombre_alternativo">
                <selector>
                    <aftertransform>
                        <operation oper-id="constant">
                            <arg-restriction on-param="const-val">
                                <has-column>egm.watrcrsl.NAMN2</has-column>
                            </arg-restriction>
                        </operation>
                    </aftertransform>
                </selector>
            </attributemap-def>
        </described-by>
 </conceptmap-def>
</r2o>
```

Material de entrada y resultados de salida

Entradas: Conjunto de *mappings* verificados entre la ontología individual, ontologías individuales interconectadas o red de ontologías y las fuentes de información (bases de datos). Además, la creación de consultas iniciales a modo de prototipos (tarea 1.8) se podría considerar como recurso de entrada inicial, si se desarrollaron prototipos para esta tarea.

Salidas: Conjuntos de instancias en RDF de las distintas fuentes.

Responsables

Ejecutor del proyecto: Ingeniero software y (experto de dominio).

Organización: (Ingeniero software y experto de dominio).

[135] http://mayor2.dia.fi.upm.es/oeg-upm/index.php/es/downloads/151-geometry2rdf

Técnicas

ODEMapster, D2R server, librería GEOMETRY2RDF, etc. para la generación de instancias y su información asociada en RDF.

Ejemplo práctico

En el proceso de explotación de *mappings*, mostrado en este ejemplo, se utiliza la aplicación ODEMapster++ (Ver Figura 5.11). Un ejemplo detallado del trabajo realizado en el contexto de esta tarea se muestra en el Capítulo 7. Asimismo, se hace referencia a la ubicación de los diversos conjuntos de instancias en RDF generados.

Tarea 4.4 Revisión y retroalimentación de expertos

El objetivo de esta tarea consiste en llevar a cabo una revisión del trabajo realizado en la tarea destinada a la creación de *mappings* (tarea 4.2) y generación de instancias (tarea 4.3), con el fin de realizar un proceso de depuración. En esta tarea, se produce la participación de los expertos de dominio por ambas partes implicadas en el proyecto. Asimismo, la participación del ingeniero de conocimiento, del lado del ejecutor del proyecto, está destinada a mostrar los *mappings* realizados en la tarea anterior y a tomar cuenta de los comentarios realizados por parte de los expertos de dominio implicados.

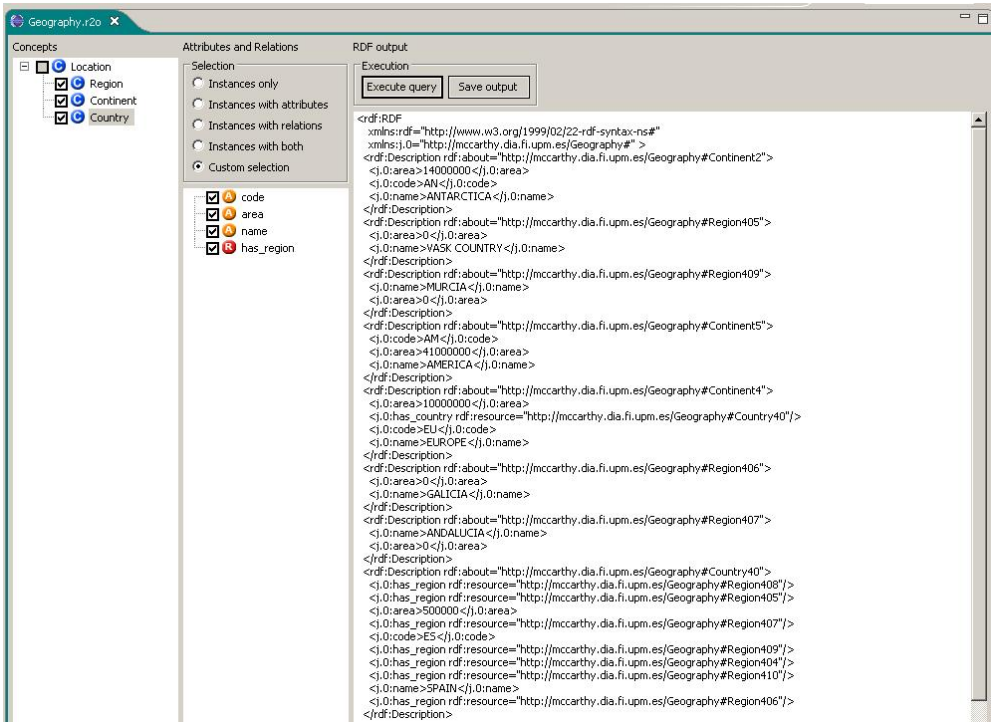


Figura 5.11. Resultados de la explotación de *mappings*

Las técnicas que implica esta tarea se concretan en procesos de consulta de *mappings* (para lo que resulta fundamental la utilización de lenguajes como ODEMQL (Barrasa, 2007), RDQL (Seaborne, 2004) o SPARQL (Prud'hommeaux *et al.*, 2008)) y diferentes sesiones de entrevistas con los expertos de dominio.

Material de entrada y resultados de salida

Entradas: Conjunto de *mappings* entre la ontología individual, ontologías individuales interconectadas o red de ontologías y las fuentes de información.

Salidas: Verificación de *mappings* establecidos y/o propuesta de cambios.

Responsables

Ejecutor del proyecto: Ingeniero de conocimiento y experto de dominio.
Organización: Experto de dominio e (ingeniero de conocimiento).

Técnicas

Consulta de *mappings* establecidos y sesiones de entrevistas.

Ejemplo práctico

Un ejemplo descriptivo del lenguaje de consulta ODEMQL se refleja en el Capítulo 7 de este trabajo. En este ejemplo, junto a un extracto de la codificación, se listan los componentes presentes en una consulta.

Del proceso de consulta y las sesiones de entrevistas con los expertos se obtuvieron importantes resultados que contribuyeron a mejorar los *mappings* establecidos. De esta manera se solventaron diversos errores asociados a duplicidades, inclusiones de otros tipos de fenómenos en nombres de instancia (por ejemplo, "*Barranco* Bernedo, *Arroyo* de"), existencia de instancias con y sin conectores dentro del nombre (por ejemplo, "Laguna de Carravalseca" y "Laguna Larga"), etc. Una descripción de los diversos problemas detectados y sus respectivas soluciones aparecen detalladas en el Capítulo 7 (sección 7.4) de este trabajo.

Tarea 4.5 Evaluación

El objetivo de esta tarea es la evaluación de la conformidad de todo el trabajo realizado a lo largo de esta actividad. Para ello, consecuencia de la aprobación de un primer filtro procedente de la retroalimentación de los expertos (tarea 4.4), esta tarea evaluará el funcionamiento global y del proceso de explotación de *mappings*. Esto hace que la participación de los expertos de dominio, de ambos lados del proyecto, resulte fundamental, debido a que estos tomaran un doble rol (experto de dominio y usuarios de aplicación). De esta manera, la participación del ingeniero de software, por parte del ejecutor del proyecto, se utilizará para depurar los posibles fallos detectados por los expertos/usuarios. Mientras, el ingeniero software tendrá una participación testimonial, ya que su participación será empleada, en caso de necesidad, como apoyo técnico.

En lo que respecta a las técnicas empleadas en este proceso de evaluación se utilizan las métricas *precision* y *recall*, con objeto de medir la exactitud o fidelidad y la completitud de los resultados, respectivamente.

Material de entrada y resultados de salida

Entradas: Conjuntos de datos RDF generados y fuentes originales (bases de datos)
Salidas: Estudio de la calidad de los resultados generados.

Responsables

Ejecutor del proyecto: Experto de dominio e ingeniero software.
Organización: Experto de dominio e (ingeniero software).

Técnicas

Medidas de *precision* y *recall.* Una descripción de estas y otras medidas de evaluación aparecen descritas en detalle en (Euzenat *et al.*, 2007).

Ejemplo práctico

La evaluación propuesta en esta tarea se refleja en el Capítulo 9 de este trabajo.

Tabla 5.21. Productos de tareas, técnicas y participantes de la actividad 5 (Publicación datos integrados)

Nº Tarea	Nombre Tarea	Responsable	Producto	Técnicas aplicables
Tarea 5.1	Descripción conjuntos de datos	Experto de dominio / Ingeniero de software	Conjuntos de datos RDF descritos con diversos elementos de metadatos	Descripción de los elementos de metadatos correspondiente a cada conjunto de datos RDF
Tarea 5.2	Verificación	Ingeniero software	Conjuntos de datos verificados	Generación de *scripts* para la validación automática de los datos publicados (URI) y/o modificación del código de VAPOUR
Tarea 5.3	Visualización	Ingeniero software / Diseñador/Desarrollador web / Experto de dominio	Generación de *FrontEnd*, visor geoespacial, sitio web (Internet o Intranet)	Implementación de un *FrontEnd* sobre los SPARQL Endpoint, implementación de visualizador geoespacial, diseño e implementación de un sitio web

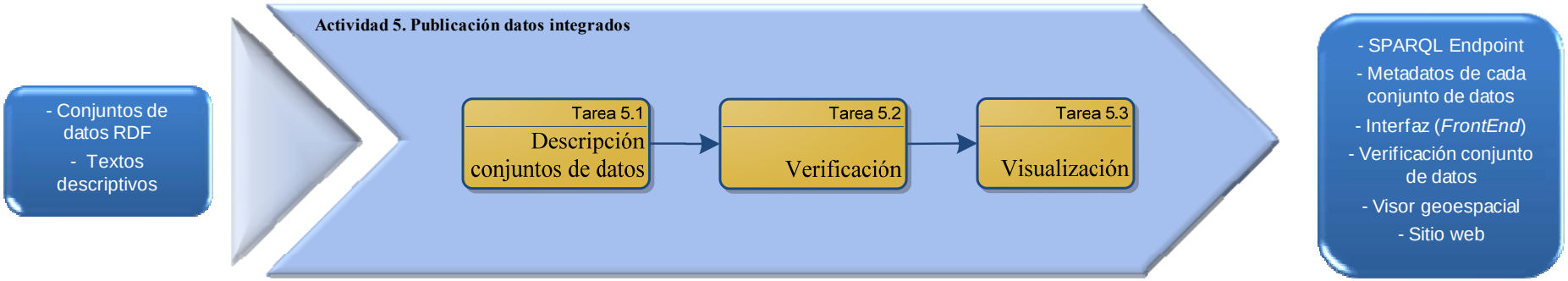


Figura 5.12. Descomposición de tareas de la actividad 5 (Publicación datos integrados)

5.5.5. Actividad 5. Publicación de datos integrados

Una vez que los datos procedentes de distintas fuentes de información han sido integrados, puede ser conveniente re-publicarlos para que puedan ser usados por terceros o enriquecidos con otras fuentes. Teniendo en cuenta el auge que ha tenido la publicación de datos como *Linked Data* (Bizer *et al.*, 2009) parece adecuado realizar esta publicación en el formato utilizado por la Web de los Datos como formato estándar. No obstante, en la realización de esta actividad se debe considerar las necesidades, requerimientos o cuestiones de licencia de los datos de la institución propietaria de la información, por tanto, dicha actividad en esta metodología aparece como opcional.

Las técnicas de publicación propuestas en el dominio del *Linked Data* pueden ser muy útiles, por proporcionar mecanismos sencillos de publicación y persistencia, así como por el hecho de poder ser relacionados con ontologías existentes. Para ello, el proceso de publicación de *Linked Data* geoespacial está basado en las recomendaciones y buenas prácticas recogidas en (Bizer, 2007).

Las entradas a esta actividad están compuestas por los conjuntos de datos RDF generados previamente (Actividad 4) y la ontología individual, ontologías individuales interconectadas o redes de ontologías que constituyen el modelo de conocimiento de los conjuntos de datos RDF. El resultado de esta actividad será la publicación de los conjuntos de datos conforme a los principios de *Linked Data* y la visualización de esta información.

En la Tabla 5.23 se muestran las entradas, salidas, técnicas y participantes asociados a cada tarea. Asimismo, en la Figura 5.12 se recoge una visión general de las diferentes tareas propuestas en esta actividad. A continuación se describen en detalle cada una de las tareas de esta actividad.

Tarea 5.1. Descripción de los conjuntos de datos

El objetivo de esta tarea es la descripción de los diversos conjuntos de datos publicados a través de un SPARQL Endpoint[1]. Para ello, se reutilizan parte de los metadatos recopilados en la tarea 2.3 (*Descripción de datos*) y se añaden otros elementos para la descripción del tipo de licencia de los datos, información sobre la procedencia de la información, cobertura espacial, etc. Por tanto, a los elementos procedentes del perfil del Núcleo Español de Metadatos, donde destaca la presencia de elementos de Dublin Core Metadata Initiative e ISO 19115, se añaden otros elementos procedentes de FOAF[2] y voID[3].

La técnica empleada en esta tarea está centrada en la descripción de los datos de los diversos conjuntos de datos con la finalidad de asociar esta descripción a la publicación de dichos datos. De esta manera, los responsables implicados, por ambas parte del proyecto, son los expertos de dominio. Estos son los encargados de la descripción de los diferentes elementos de metadatos para cada conjunto de datos. Asimismo, por parte del ejecutor del proyecto también está implicado el ingeniero de software, encargado de establecer la asociación entre la descripción (metadatos) y el conjunto de datos RDF correspondiente.

Material de entrada y resultados de salida

Entradas: Conjuntos de datos RDF

Salidas: Conjuntos de datos RDF descritos con diversos elementos de metadatos.

[1] *SPARQL Endpoint*. http://semanticweb.org/wiki/SPARQL_endpoint

[2] *FOAF* (*Friend Of A Friend*) http://www.foaf-project.org/

[3] *voID*. http://vocab.deri.ie/void/guide#sec_1_6_SPARQL_endpoint_and_Examp

Responsables

Ejecutor del proyecto: Experto de dominio e ingeniero de software.
Organización: Experto de dominio.

Técnicas

Descripción de los elementos de metadatos correspondiente a cada conjunto de datos RDF.

Ejemplo práctico

En este ejemplo se muestran los elementos de metadatos utilizados para describir cada recurso de información (instancia). A continuación se listan los metadatos y sus descripciones asociadas para el fenómeno "Río Lara", identificado por la URI `<http://geo.linkeddata.es/data /R%C3%ADo/Lara %2C%20R%C3%ADo>` (Ver Tabla 5.22).

Tarea 5.2. Verificación

El objetivo de esta tarea es comprobar que los conjuntos de datos se han publicado conforme a los principios de *Linked Data*. Para ello se puede utilizar un servicio de verificación como el ofrecido por VAPOUR[4] (Berrueta *et al.*, 2008). Este servicio comprueba que los datos publicados sean conformes con los principios de *Linked Data*, las recetas de las mejores prácticas para publicar vocabularios RDF[5] y las recomendaciones para conseguir *cool* URI (Ayers *et al.*, 2008). Este validador está disponible como un servicio web[6] y como un software de código abierto bajo los términos de licencia W3C[7].

Tabla 5.22. Elementos de metadatos y su descripción asociada

Metadatos	Descripción
rdf:type	<http://www.w3.org/2004/03/trix/rdfg-1/Graph>
foaf:primaryTopic	<http://geo.linkeddata.es/resource/Pa%C3%ADs/Espa%C3%B1a>
dcterms:title	Nomenclátor Geográfico Conciso de España (NGCE)
dcterms:description	El Nomenclátor Geográfico Conciso recoge información a escala 1:1.000.000. Esta fuente constituye un corpus básico de topónimos estandarizados creado por la Comisión Española de Nombres Geográficos. Esta base de datos tiene una única tabla donde aparece una miscelánea de fenómenos geográficos. En la versión controlada de la mencionada base de datos el número de entradas asciende a 3.667 registros pertenecientes a 22 fenómenos.
dcterms:creator	<http://geo.linkeddata.es/resource/Organizaci%C3%B3n/InstitutoGeogr%C3%A1ficoNacionalDeEspa%C3%B1a>
dcterms:publisher	<http://geo.linkeddata.es>
dc:rights	La información geográfica digital comprendida en el Equipamiento Geográfico de Referencia Nacional (artículo 1.1 de la Orden FOM/956/2008) así como los metadatos de los datos geográficos y servicios del IGN-CNIG, no requieren la aceptación de licencia y su uso será, en cualquier caso, libre y gratuito, siempre que se mencione al Instituto Geográfico Nacional como propietario de los datos.
dcterms:spatial	<http://geo.linkeddata.es/resource/Pa%C3%ADs/Espa%C3%B1a>

Los responsables implicados en esta tarea, por ambos lados, son los ingenieros de software, consecuencia del trabajo de implementación/utilización del servicio y ejecución de pruebas que se lleva a cabo en esta tarea.

[4] http://vapour.sourceforge.net/
[5] http://www.w3.org/TR/swbp-vocab-pub/
[6] http://validator.linkeddata.org/vapour
[7] http://www.w3.org/Consortium/Legal/copyright-software

Material de entrada y resultados de salida

Entradas: Conjuntos de datos publicados (URI de los recursos publicados y, opcionalmente, se pueden añadir como entradas las URI de las clases, propiedades e instancias de los vocabularios/ontologías utilizados).

Salidas: Conjuntos de datos verificados.

Responsables

Ejecutor del proyecto: Ingeniero software.

Organización: Ingeniero software.

Técnicas

Generación de *scripts* para la verificación automática del conjunto de datos publicados (URI) y/o modificación del código de VAPOUR para adaptarlo a las necesidades determinadas por el proyecto de integración.

Ejemplo práctico

Este ejemplo permite mostrar los resultados ofrecidos por VAPOUR en el proceso de verificación de los recursos publicados. Los recursos publicados como *Linked Data* pertenecen al tipo recursos sin información[8], es decir, corresponden a objetos pertenecientes al mundo real que existen fuera de la Web. A continuación, como muestra del proceso de verificación propuesto en esta tarea, se presentan los resultados del recurso <`http://geo.Linked Data.es/resource/Playa/Cristianos%2C%20Playa%20de%20Los`>.

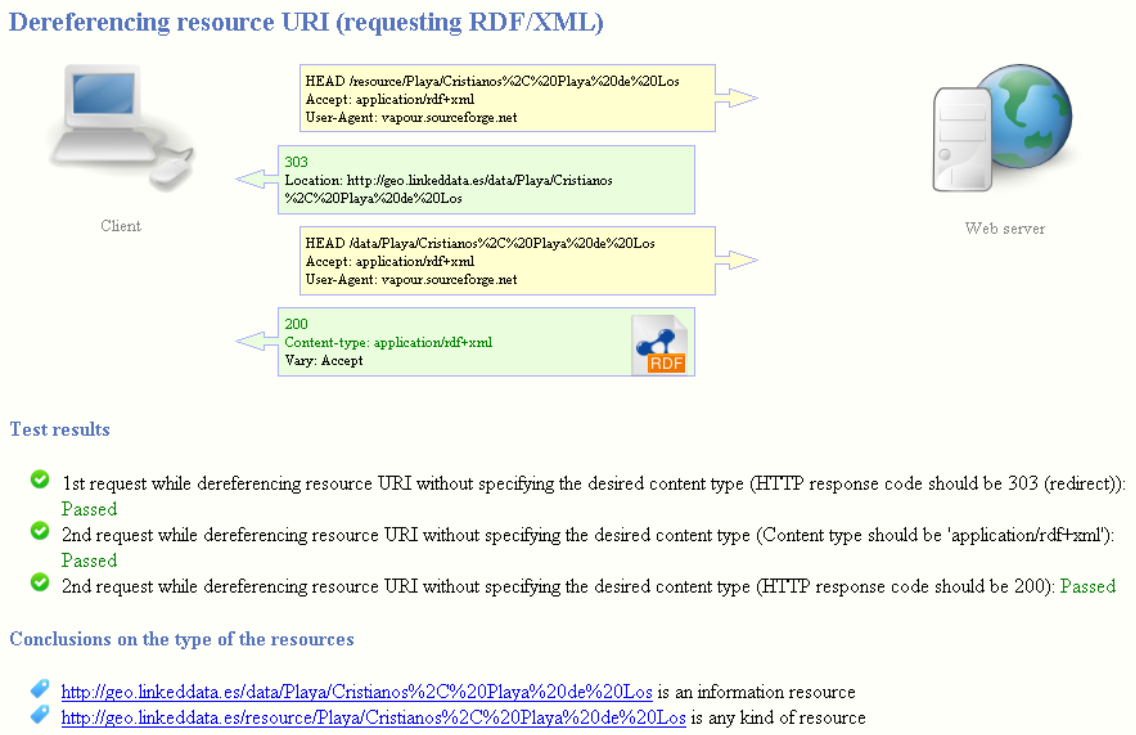


Figura 5.13. Desreferencia de un recurso (Conformidad con vocabulario asociado)

Este proceso de verificación se centra en la *desreferenciación* de una URI, es decir, en la búsqueda de una URI en la Web para obtener información del recurso referido. Aquí se valida la

[8] *Non-Information Resources.* Todos los recursos que aparecen en la Web tradicional, tales como: documentos, imágenes y otros archivos multimedia son recursos de información. Sin embargo, muchos de las cosas de las que se comparten datos no lo son (personas, lugares, conceptos científicos, etc.). Así, se establece como regla general que todos los "objetos del mundo real" que existan fuera de la Web son recursos sin información (http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/).

representación RDF/XML del recurso (es decir, el vocabulario/ontología utilizado) y la *desreferencia* del recurso URI.

- La validación de la **representación RDF/XML** de una URI de recurso consiste en (ver Figura 5.13):
  - **303 - Recursos sin información.** Debido a que los recursos sin información no pueden ser *desreferenciados* directamente, la arquitectura Web utiliza un *truco* para permitir que las URI de los recursos publicados (recursos sin información) sean *desreferenciadas*. Para esto, en lugar de enviar una representación del recurso, el servidor envía al cliente la URI de un recurso de información que describe a este tipo de recursos utilizando el código de respuesta HTTP 303. Esto se denomina una redirección 303. En un segundo paso, el cliente *desreferencia* esta nueva URI y obtiene una representación que describe el recurso sin información original.
  - ***Application*/RDF+XML**.
  - **200 - Recursos de información**. Cuando una URI que identifica a un recurso de información es *desreferenciada*, el servidor propietario de la URI, normalmente, genera una nueva representación, una instantánea del estado actual del recurso de información y lo devuelve al cliente utilizando el código de respuesta HTTP 200 OK.

En la Figura 5.13 se muestra que el recurso es correctamente verificado y ofrece como salida la descripción de las características del recurso, el listado de vocabularios utilizados y los diversos elementos de metadatos que lo describen.

- La validación de la ***desreferencia* de una URI de recurso** se centra en los códigos de respuesta HTTP 303 y 200, descritos con anterioridad. En este caso, VAPOUR realiza la verificación sin considerar la negociación de contenido[9]. Como se muestra en la Figura 5.14 el recurso es correctamente verificado y ofrece como salida la representación HTML del mismo (ver Figura 5.15).

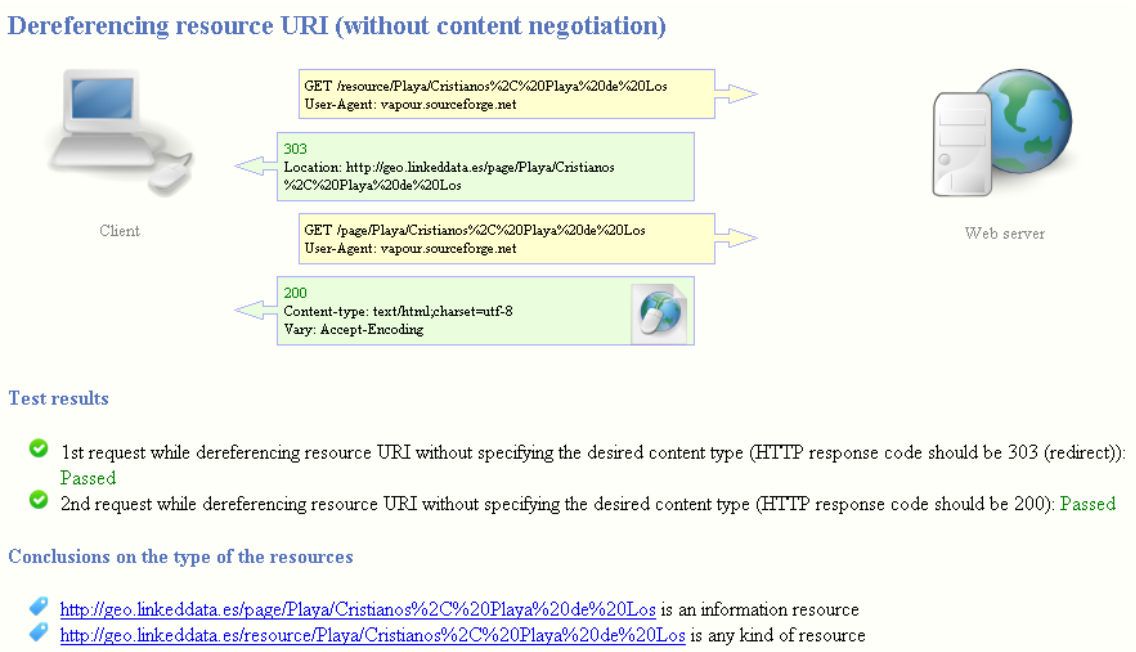


Figura 5.14. *Desreferencia* de un recurso

[9] *Content negotiation*. Es un mecanismo definido en la especificación HTTP que permite servir diferentes versiones de un mismo recurso en la misma URI.

Figura 5.15. Representación HTML del recurso

Asimismo, conforme a la recomendación propuesta en (Bizer *et al.*, 2007), en este proceso de verificación se pueden observar las tres diferentes versiones de URI ofrecidas para cada recurso en el conjunto de datos publicado. A continuación se muestran dichas URI utilizando el ejemplo anterior.

- `http://geo.linkeddata.es/resource/Playa/Cristianos%2C%20Playa%20de%20Los`
  La URI que identifica el recurso sin información "Playa de Los Cristianos"
- `http://geo.linkeddata.es/data/Playa/Cristianos%2C%20Playa%20de%20Los`
  El recurso de información con una representación RDF/XML que describe "Playa de Los Cristianos"
- `http://geo.linkeddata.es/page/Playa/Cristianos%2C%20Playa%20de%20Los`
  El recurso de información con una representación HTML que describe al recurso "Playa de Los Cristianos". En este trabajo este tipo de representación se ofrece a través de Pubby (ver Figura 5.15).

Tarea 5.3. Visualización

Esta tarea trata de facilitar la interacción con los conjuntos de datos publicados. Para ello, se propone el desarrollo de una interfaz (*FrontEnd*), un visor de información geoespacial y un sitio web.

La **generación de un *FrontEnd*** permite la transformación de los servicios SPARQL Endpoint a servidores de *Linked Data*. La realización de esta tarea da conformidad con los siguientes principios de *Linked Data* (Berners-Lee, 2006): a) Utilización del protocolo HTTP para nombrar y resolver la ubicación de los datos identificados mediante esas URI y b) Ofrecimiento de descripciones semánticas de los datos en RDF al resolver su URI.

Los *FrontEnd* desarrollados no son muy numerosos, aunque para la realización de esta tarea existen diversas aplicaciones, tales como: D2R Server, Talis Platform, Pubby o Paget.

Los responsables implicados en esta tarea, por ambos lados, son los ingenieros de software, consecuencia del trabajo de implementación y ejecución de pruebas que se lleva a cabo en esta tarea.

Para el desarrollo del **visor geoespacial** se pueden utilizar diversos componentes software y/o API[10], por ejemplo: Map builder, Google Maps API, OpenLayer, Geoserver o Deegree. Este visor permite visualizar de forma gráfica la componente espacial recogida en la información publicada.

[10] Una interfaz de programación de aplicaciones o API (*Application Programming Interface*) es el conjunto de funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece una librería para ser utilizada por otro software como una capa de abstracción.

Finalmente, como parte de la publicación de los datos, se propone la generación de un **sitio web** donde quedarían alojados los resultados de la tarea de visualización y toda la información resultante del proyecto de integración. No obstante, dependiendo de las características del proyecto a realizar la publicación de dicha información puede realizarse en Internet o Intranet.

Los responsables implicados en esta tarea son los ingenieros de software, consecuencia del trabajo de implementación y ejecución de pruebas a realizar en esta tarea, el diseñador/desarrollador web y los expertos de dominio, ya que su conocimiento puede ser necesario durante el desarrollo de esta tarea.

Material de entrada y resultados de salida

Entradas: SPARQL Endpoint (conjuntos de datos publicados) y textos descriptivos de la información disponible.

Salidas: Generación de *FrontEnd*, visor geoespacial y sitio web.

Responsables

Ejecutor del proyecto: Ingeniero software, diseñador/desarrollador web y expertos de dominio.

Organización: Ingeniero software y expertos de dominio.

Técnicas

Implementación de un *FrontEnd* sobre los SPARQL Endpoint generados, implementación de la aplicación web para la visualización de los conjuntos de datos RDF geoespaciales y diseño e implementación de un sitio web.

Ejemplo práctico

Un ejemplo detallado del trabajo realizado en el contexto de esta tarea se presenta en el Capítulo 7 (sección 7.5).

5.5.6. Actividad 6. Depuración / Post-procesamiento

El objetivo de esta actividad consiste en detectar los duplicados de instancia presentes en los conjuntos de datos geoespaciales publicados y utilizar dichos duplicados para interrelacionar el conocimiento presente en las fuentes del proceso de integración.

La propuesta de esta actividad deriva en dos posibles vías de actuación. Por un lado, una posibilidad sería exponer los datos integrados publicados en el conjunto de datos pertenecientes a Linking Open Data como entidades no relacionadas. Esta propuesta se recoge en la Actividad 5 de esta metodología. Por otro lado, otra posibilidad consiste en llevar a cabo un proceso de integración de información previo a la generación de *Linked Data*, asegurando de esta manera una mejor calidad en los datos generados. La elección de esta línea de trabajo conlleva centrarse en primer lugar en la detección de instancias duplicadas en conjuntos de datos heterogéneos dentro del dominio de la información geoespacial. De esta manera se pretende delimitar errores y definir relaciones (*owl:sameAs*) entre instancias idénticas presentes en los conjuntos de datos tratados. Esta propuesta de actuación se desarrolla en esta actividad. No obstante, la adopción del primer enfoque propuesto provoca que esta Actividad se considere opcional dentro de esta metodología.

Para alcanzar el objetivo propuesto, esta actividad propone realizar una detección de duplicados general (tarea 6.1), con objeto de completar la información obtenida en la tarea 2.4 (comparativa entre datos) y de efectuar una detección de problemas en diferentes capas (léxico, sintáctico, semántico y pragmático). Se acomete un proceso de abstracción del cual surge la formalización e implementación de diversos enfoques heurísticos para la detección exhaustiva de duplicados de instancia (tarea 6.2). Tras esta tarea, se aplican enfoques heurísticos en los diversos conjuntos de datos RDF con la finalidad de obtener, de forma exhaustiva, las duplicidades existentes y

formalizar las relaciones que interrelacionen a los datos publicados. Finalmente, se realiza un proceso de evaluación (tarea 6.4) de los diferentes enfoques heurísticos a través de una evaluación conforme a un *gold standard*. Asimismo, también se tomarán en cuenta los elementos de medida *precision* y *recall* para comprobar los resultados del proceso de detección de duplicados.

Esta actividad y sus respectivas tareas, como se muestra en la Figura 5.2, están inspiradas por las tareas propuestas en la ingeniería del software, concretamente MÉTRICA versión 3 (MÉTRICA v.3) y *Rational Unified Process* (RUP) (Kruchten, 2000), por diferentes propuestas de ingeniería del conocimiento, tales como (Gómez-Pérez *et al*., 1997) y (Schreiber *et al*., 1999), por el área de *ontology matching* (Euzenat *et al*., 2007), y por diversas técnicas y métodos de detección de duplicados (Winkler, 2006; Elmagarmid *et al*., 2007).

Las entradas a esta actividad están constituidas por las fuentes de información y la comparación de las diferencias y semejanzas de los datos de las diversas fuentes. En cuanto a los resultados, estos están constituidos por una detección general y exhaustiva de los duplicados de instancia existentes, detección de problemas a diferentes niveles (léxico, sintáctico, semántico y pragmático), formalización y desarrollo de diversos enfoques heurísticos, obtención de interrelaciones entre conjuntos de datos y, por último, un informe de resultados de la evaluación realizada.

De manera resumida, en la Tabla 5.23, se reflejan las entradas, salidas, técnicas y participantes asociados a cada una de las tareas. Asimismo, en la Figura 5.16 se muestra una visión general de las diferentes tareas propuestas por esta actividad. Las diversas tareas que conforman esta actividad son descritas en detalle a partir del siguiente apartado.

Tarea 6.1. Detección de duplicados

En esta tarea se lleva a cabo un proceso de detección general de instancias duplicadas mediante un proceso de detección-comparación donde se requiere la intervención de los expertos de dominio. Asimismo, en base a la detección de duplicados realizada, esta tarea produce una abstracción de los problemas detectados.

Al igual que ocurre en la tarea destinada a la comparativa entre esquemas (tarea 2.2) y datos (tarea 2.4) el punto de partida de esta tarea es la recopilación y descripción de las fuentes de información (tarea 1.6), consecuencia de que el proceso de comparación se lleva a cabo entre fuentes con solapamientos de información. Asimismo, al igual que se describe en las mencionadas tareas del proceso de comparación, el proceso de detección de duplicados se realiza, en primer lugar, entre fuentes que tengan idéntica escala de recogida de información y que, además, presenten idénticas características respecto a ciertos parámetros cartográficos identificados en la colección de metadatos, tales como: sistema de referencia, información del datum vertical, resolución y exactitud posicional. En segundo lugar, con el objetivo de enriquecer el proceso desde la óptica que ofrece la visión de la información geográfica a diferentes escalas, se procede a un proceso de detección de duplicados entre las fuentes de información de diversas escalas caracterizadas por poseer idénticos sistemas de referencia, información del datum vertical y exactitud posicional.

Esta tarea se retroalimenta con la información proveniente del análisis y comparación de aspectos léxico-sintácticos, unidades de medida, ausencias generalizadas y/o lagunas puntuales de información (omisiones), exceso de información (comisiones) y presencia-tratamiento de información multilingüe, es decir, en este caso la retroalimentación se produce con las salidas procedentes de la tarea 2.4. Esta información sirve de ayuda para la clasificación de los problemas detectados en las diferentes capas propuestas.

Tabla 5.23. Productos de tareas, técnicas y participantes de la actividad 6 (Depuración / Post-procesamiento)

Nº Tarea	Nombre Tarea	Responsable	Producto	Técnicas aplicables
Tarea 6.1	Detección de duplicados	Ingeniero software / Experto de dominio	Descripción general de duplicados (*gold standard*), detección de problemas a diferentes niveles (léxico, sintáctico, semántico, pragmático) e informe de resultados	Detección-comparación manual de duplicados, catalogación de problemas identificados y realización de informes
Tarea 6.2	Desarrollo de heurísticas	Ingeniero software / Experto de dominio	Formalización e implementación de enfoques heurísticos	Implementación de diversos enfoques heurísticos
Tarea 6.3	Creación de enlaces entre conjuntos de datos	Ingeniero software	Resultados de la interrelación entre conjuntos de datos (formalización *owl:sameAs*).	Aplicación de las heurísticas desarrolladas sobre los conjuntos de datos RDF y/o sobre dichos conjuntos y otros conjuntos de interés
Tarea 6.4	Evaluación de heurísticas	Ingeniero software / Experto de dominio	Evaluación de la calidad del proceso de detección de instancias duplicadas (informe de resultados).	Gold standard, precision, recall y Medida E
Tarea 6.5	Descubrimiento de datos integrados	Ingeniero software	Fichero *sitemap* (XML)	Desarrollo de scripts para la generación de los ficheros XML (*sitemap*) asociado a los conjuntos de datos RDF publicados

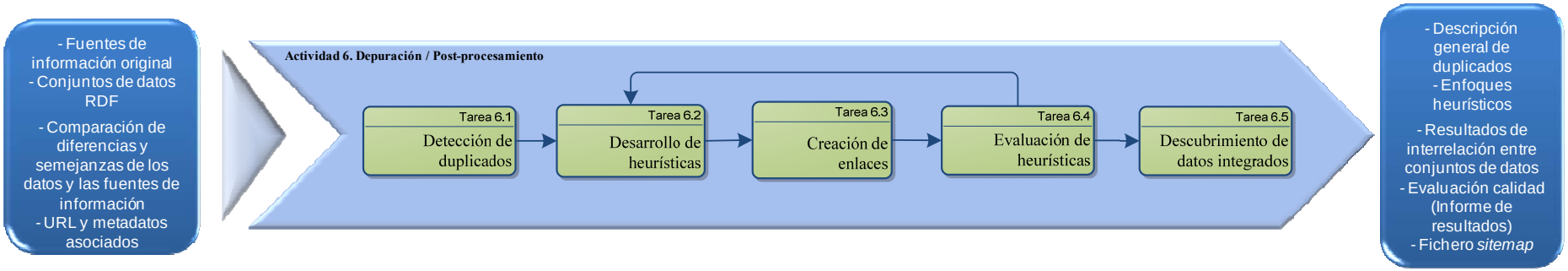


Figura 5.16. Descomposición de tareas de la actividad 6 (Depuración / Post-procesamiento)

La mezcla entre aspectos puramente de dominio y otros más técnicos hacen que esta tarea deba ser realizada tanto por expertos de dominio como por ingenieros de software (con una implicación menos fuerte por parte de los responsables de la organización). Las técnicas empleadas se ciñen, en una primera fase, a la detección-comparación de duplicados de instancia entre las diversas fuentes de información componentes, en gran parte, a través de un proceso manual. La información obtenida sirve para generar una parte importante del *gold standard* utilizado en la tarea 6.4. En una segunda fase, se lleva a cabo un proceso de abstracción y catalogación de los problemas identificados, para posteriormente realizar tablas e informes de los problemas observados.

Material de entrada y resultados de salida

Entradas: Fuentes de información en formato original (bases de datos) y conjuntos de datos RDF. Además, se utiliza como entrada la comparación de las diferencias y semejanzas de los datos de las fuentes de información y los posibles prototipos (tarea 1.8) desarrollados.

Salidas: Descripción general de duplicados, detección de problemas a diferentes niveles (léxico, sintáctico, semántico, pragmático) e informe de resultados.

Responsables

Ejecutor del proyecto: Ingeniero software y experto de dominio.

Organización: (Experto de dominio e ingeniero software).

Técnicas

Detección-comparación de duplicados (generación de un primer acercamiento para la obtención del *gold standard*), catalogación de problemas identificados y realización de tablas e informes.

Ejemplo práctico

Un ejemplo de esta tarea se recoge en el Capítulo 8 (sección 8.2) de este trabajo.

Tarea 6.2. Desarrollo de heurísticas

El objetivo de esta tarea es desarrollar diferentes enfoques heurísticos que contribuyan a la detección exhaustiva de duplicados de instancia en grandes conjuntos de datos. Para ello, resulta fundamental la clasificación de los problemas detectados en la tarea anterior (tarea 6.1). Esta información se utiliza para proponer una serie de propuestas heurísticas conformes a las diferentes capas de clasificación (léxica, sintáctica, semántica y pragmática).

En esta tarea resulta fundamental la participación de expertos de dominio e ingenieros software. La implicación de estos miembros se debe a la importancia del conocimiento de dominio que posee el experto (contribuye a tener una visión más completa en el proceso de formalización de los diversos enfoques heurísticos) y de la necesidad del trabajo de formalización e implementación por parte del ingeniero software. Los responsables implicados en esta actividad tienen participación por las diferentes partes implicadas en este proceso, si bien, por parte de la organización, la participación del ingeniero de software supone un apoyo técnico más que una implicación directa.

Material de entrada y resultados de salida

Entradas: Descripción general de duplicados, detección de problemas a diferentes niveles (léxico, sintáctico, semántico, pragmático), informe de resultados y posible prototipos desarrollados (tarea 1.8).

Salidas: Formalización e implementación de enfoques heurísticos.

Responsables

Ejecutor del proyecto: Ingeniero software y experto de dominio.

Organización: Experto de dominio, (ingeniero software).

Técnicas

Implementación de diversos enfoques heurísticos propuestos a diferentes niveles (léxico, sintáctico, semántico y pragmático) para la detección de duplicados. Asimismo, diferentes técnicas y métodos de detección de duplicados se recogen en (Winkler, 2006; Elmagarmid *et al.*, 2007). Una recopilación detallada del estado de la cuestión sobre este tema aparece en el Capítulo 3 (sección 3.3) de este trabajo.

Ejemplo práctico

Un ejemplo detallado de los diferentes enfoques heurísticos propuestos en esta tarea se describe en el Capítulo 8 (sección 8.3) de este trabajo.

Tarea 6.3. Creación de enlaces entre conjuntos de datos

Esta tarea pretende conectar los diferentes conjuntos de datos RDF publicados a través de las instancias duplicadas detectadas mediante la aplicación de los enfoques heurísticos desarrollados. Esta conexión entre los diversos conjuntos se realiza mediante la formalización de la relación *owl:sameAs*. Esta relación (*owl:sameAs*) es la forma habitual de relacionar un conjunto de URI que se refieren al mismo recurso. La semántica de esta relación significa que todas las URI vinculadas (recursos) con este predicado tienen la misma identidad (Bechhofer *et al.*, 2004), es decir, el sujeto y el objeto debe ser el mismo recurso.

Para la realización de esta tarea también puede utilizarse el *framework* propuesto por *Silk* (Volz *et al.*, 2009) o aplicar algunos de los enfoques para la detección de duplicados descritos en el Capítulo 3 (sección 3.3) de este trabajo.

Los responsables implicados en esta tarea, por ambos lados, son los ingenieros de software, consecuencia del trabajo de ejecución de pruebas que se realiza en esta tarea.

Material de entrada y resultados de salida

Entradas: Conjuntos de datos RDF publicados.

Salidas: Resultados de la interrelación entre conjuntos de datos (formalización de la relación *owl:sameAs* entre instancias duplicadas).

Responsables

Ejecutor del proyecto: Ingeniero de software.

Organización: Ingeniero de software.

Técnicas

Conjunto de heurísticas desarrolladas aplicadas sobre los conjuntos de datos RDF y/o sobre dichos conjuntos y otros conjuntos de datos de interés presentes en la Web de los Datos.

Ejemplo práctico

Un ejemplo del trabajo realizado en el contexto de esta tarea se presenta en el Capítulo 8 de este trabajo.

Tarea 6.4. Evaluación de heurísticas

El objetivo de esta tarea es la evaluación de los diferentes enfoques heurísticos propuestos en la tarea 6.2, a través de los resultados obtenidos en la tarea 6.3. Este proceso de evaluación se realiza mediante la comparación de los resultados obtenidos del proceso de detección manual de duplicados realizado por un experto de dominio (*gold standard*) (tarea 6.1) y los resultados de la ejecución de los diversos enfoques heurísticos propuestos (tarea 6.3). La comparación de resultados conduce a un proceso de retroalimentación directa que contribuye a la mejora y refinamiento de las diferentes propuestas heurísticas en las distintas capas.

Dadas las características de este proceso de evaluación, la participación de los expertos de dominio, de ambos lados del proyecto, resulta fundamental, consecuencia de la importancia de su conocimiento y de su implicación directa en el proceso de evaluación. Asimismo, la participación del ingeniero de software, por parte del ejecutor del proyecto, es necesaria para solventar posibles fallos detectados por los expertos/usuarios y para realizar posibles refinamientos de los enfoques heurísticos.

Finalmente, junto a la utilización del *gold standard* en el proceso de evaluación, también se utilizan las métricas de *precision*, *recall* y Medida E con objeto de evaluar la calidad de los resultados del proceso de detección de duplicados, así como, desde una perspectiva más abstracta, la calidad del proceso llevado a cabo en esta actividad.

Material de entrada y resultados de salida

Entradas: Conjuntos de datos RDF, resultados de la detección manual de duplicados (tarea 6.1), resultados de la aplicación de los enfoques heurísticos (tarea 6.3).

Salidas: Evaluación de la calidad del proceso de detección de instancias duplicadas (informe de resultados).

Responsables

Ejecutor del proyecto: Ingeniero software y experto de dominio.

Organización: Experto de dominio.

Técnicas

Gold standard, medidas de *precision*, *recall* y Medida E.

Ejemplo práctico

En el Capítulo 9 de este trabajo se recoge la evaluación realizada sobre las diferentes heurísticas desarrolldas.

Tarea 6.5. Descubrimiento de datos integrados

El objetivo de esta tarea es indexar los conjuntos de datos RDF en los distintos navegadores de *Linked Data* existentes, por ejemplo: Sindice. Para ello, resulta necesario generar un fichero XML (*sitemap*) que recoja las diferentes URL de un sitio web y sus metadatos asociados (cuándo fue la última actualización, frecuencia de modificación, importancia en relación con otras URL del sitio web). Su finalidad es facilitar el rastreo del sitio web por parte de los motores de búsqueda existentes.

Los responsables implicados en esta tarea, por ambos lados, son los ingenieros de software, cuyo trabajo en esta tarea está centrado en el desarrollo de *scripts* para la generación de los ficheros XML (*sitemap*).

Material de entrada y resultados de salida

Entradas: Diferentes URL del sitio web y sus metadatos asociados.

Salidas: Fichero *sitemap* (XML).

Responsables

Ejecutor del proyecto: Ingeniero software.

Organización: Ingeniero software.

Técnicas

Desarrollo de *scripts* para la generación de los ficheros XML (*sitemap*) asociado a los conjuntos de datos RDF publicados.

Ejemplo práctico

A continuación se recoge un extracto del fichero sitemap generado para la publicación de los diversos conjuntos de datos publicados en este trabajo (Ver Tabla 5.24).

Tabla 5.24. Extracto del sitemap de los datos integrados publicados

```
<?xml version="1.0" encoding="UTF-8"?>
<urlSet xmlns="http://www.sitemaps.org/schemas/sitemap/0.9" xmlns:sc= "http:
//sw.deri.org/2007/07/sitemapextension/scschema.xsd">
  <url>
    <loc>http://geo.linkeddata.es/resource/Acequia/Murciano%2C%20Acequia%20del</loc>
    <lastmod>2005-01-01</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.8</priority>
  </url>
  <url>

<loc>http://geo.linkeddata.es/resource/Barranco/Val%20de%20Murcia%2C%20Barranco%20de</loc>
    <lastmod>2005-01-01</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.8</priority>
  </url>
<sc:dataset>
    <sc:datasetLabel>Este es el dataset de geolinkeddata</sc:datasetLabel>
    <sc:linkedDataPrefix slicing="subject-object">http://geo.linkeddata.es/ </sc:linkedDataPrefix>
    <sc:dataDumpLocation>http://geo.linkeddata.es/geo.rdf</sc:dataDumpLocation>
  </sc:dataset>
</urlSet>
```

5.5.7. Actividad 7. Modificaciones sobre fuentes de datos

Esta actividad se lleva a cabo con objeto de fusionar la información obtenida en parte de la actividad 2, concretamente las tareas 2.3 y 2.4, y en la actividad 6. Con esta fusión se pretende combinar los resultados obtenidos de la aplicación de procesos manuales y semi-automáticos, respectivamente, con el objetivo de obtener y proponer modificaciones sobre las fuentes de información que contribuyan a la depuración y mejora de las mismas. A pesar de esto, los resultados de esta actividad no serán vinculantes para las fuentes de información implicadas en el proceso de integración, debido a que esta metodología se sostiene en la base de que dichas fuentes permanecen distribuidas y gestionadas por sus productores de origen, es decir, conforme a los principios INSPIRE. Esta actividad y sus respectivas tareas, como puede apreciarse en la Figura 5.2, están inspiradas por las tareas propuestas en la ingeniería de datos.

En esta actividad las entradas son los resultados de la comparación de las diferencias y semejanzas de los datos de las fuentes de información (tarea 2.4), propuestas de cambios/mejoras con respecto a fuentes y datos (tarea 2.5), descripción exhaustiva de duplicados, detección de problemas a diferentes niveles (léxico, sintáctico, semántico, pragmático) e informe de resultados (tarea 6.1) y fuentes de información heterogéneas integradas e informe de resultados (tarea 6.4). Asimismo, los resultados son un catálogo de propuestas de depuración de bases de datos y un informe de propuestas de depuración.

Las diversas entradas, salidas, técnicas y participantes asociados a cada una de las tareas se muestran, de manera resumida, en la Tabla 5.25. Asimismo, la Figura 5.17 contiene una visión general de las diferentes tareas propuestas por esta actividad. Estas tareas son descritas en detalle a partir del siguiente apartado.

Tabla 5.25. Productos de tareas, técnicas y participantes de la actividad 7 (Modificaciones sobre fuentes de datos)

Nº Tarea	Nombre Tarea	Responsable	Producto	Técnicas aplicables
Tarea 7.1	Recopilación y priorización de eventos	Ingeniero de conocimiento / Ingeniero software	Identificación de problemáticas Catálogo de propuestas de depuración de bases de datos Informe de propuestas de depuración	Catalogación
Tarea 7.2	Presentación de propuestas de cambios	Ingeniero software	Informe de propuestas de depuración.	Sesiones de entrevistas y reingeniería de fuentes

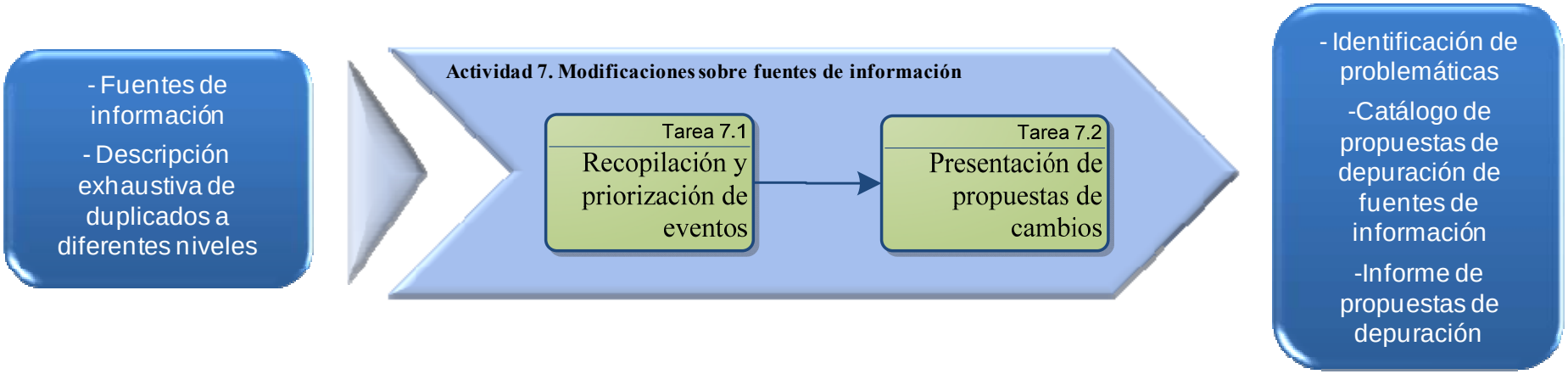


Figura 5.17. Descomposición de tareas de la actividad 7 (Modificaciones fuentes de datos)

Tarea 7.1. Recopilación y priorización de eventos

En esta tarea se lleva a cabo una recopilación de la información obtenida en la actividad 2, concretamente en las tareas 2.3, 2.4 y 2.5, y en la actividad 6 la información concerniente a las tareas 6.1 y 6.3. Esta recopilación de información pretende combinar los resultados obtenidos de la aplicación de procesos manuales (actividad 2) y semi-automáticos (actividad 6) en la detección y desambiguación de los datos. El objetivo pretendido en esta tarea consiste en la identificación de problemáticas y la realización de un catálogo de propuestas de depuración de las diferentes bases de datos que componen el proceso de integración, acompañado de un informe detallado sobre dichas propuestas.

Los responsables implicados en esta tarea, únicamente por parte del ejecutor del proyecto, son el ingeniero de conocimiento y el ingeniero de software. La técnica a aplicar por estos responsables será la catalogación de los diferentes problemas encontrados en las referidas fuentes de información.

Material de entrada y resultados de salida

Entradas: Descripción de alto nivel de los datos componentes de las fuentes de información (tarea 2.3), comparación de las diferencias y semejanzas de los datos de las fuentes de información (tarea 2.4), propuestas de cambios/mejoras con respecto a fuentes y datos para envío a los respectivos administradores (tarea 2.5), descripción exhaustiva de duplicados, detección de problemas a diferentes niveles (léxico, sintáctico, semántico, pragmático) e informe de resultados (tarea 6.1) y fuentes de información heterogéneas integradas e informe de resultados (tarea 6.4).

Salidas: Catálogo de propuestas de depuración de bases de datos.

Responsables

Ejecutor del proyecto: Ingeniero de conocimiento e ingeniero de software.

Organización: --

Técnicas

Catalogación

Ejemplo práctico

En este ejemplo se muestra, de forma breve, algunos problemas encontrados (errores léxicos (Tabla 5.26), identificación del multilingüismo (Tabla 5.27), relación entre fenómenos asociados (Tabla 5.28), presencia de fenómenos múltiples (Tabla 5.29) y diversidad de instancias en un mismo campo (Tabla 5.30) y sus pertinentes propuestas de cambio recomendadas. Una descripción más detallada de los problemas detectados en las diferentes bases de datos objeto de este trabajo se muestra en el Capítulo 8 (sección 8.2).

Tabla 5.26. Errores léxicos

Error	Sugerencia
Ca+ada	Cañada
Ca+o	Caño
Rac3/4	-

Tabla 5.27. Identificación de multilingüismo

Fenómeno	Idioma
Cavorco	Gallego
Erreka	Euskera
Barranqueira	Gallego
Barranquet	Catalán
Estany	Catalán

Tabla 5.28. Relación entre fenómenos asociados

Fenómeno	Relación
Cavorco	Barranco
Erreka	Arroyo
Barranqueira	Barranco
Barranquet	Barranco
Estany	Laguna

Tabla 5.29. Presencia de fenómenos múltiples

Nombre instancia	Fenómeno al que pertenece
Nuevo cauce del Río Guadiana	Río
Primer canal transversal de Albatera	Canal
Encharcamientos de Salburua - Balsa de Arcaute	Balsa

Tabla 5.30. Diversidad de instancias en un único campo

Nombre instancia	Sugerencia
Salada Grande o Laguna de Alcañiz	División en dos registros para distinción de nombre oficial y alternativo
ALMONACID O CÁMARAS	Inclusión tipo fenómeno y división en dos campos
BEGATILLO O BORDÓN	Inclusión tipo fenómeno y división en dos campos

Tarea 7.2. Presentación de propuestas de cambios

En esta tarea se lleva a cabo, en primer lugar, un informe de las diferentes propuestas de depuración extraídas del catálogo realizado con anterioridad (tarea 7.1). En segundo lugar, se produce un contacto con el administrador de cada una de las fuentes de información del proceso de integración. El objetivo es presentar a estos administradores las propuestas de cambios recogidas en el informe, comprobar la viabilidad y las posibles actividades de reingeniería asociadas a dichas propuestas. De esta manera, resulta fundamental la participación de los ingenieros de software de ambas partes implicadas en el proyecto.

Material de entrada y resultados de salida

Entradas: Catálogo de propuestas de depuración de bases de datos.
Salidas: Informe de propuesta de depuración.

Responsables

Ejecutor del proyecto: Ingeniero software.
Organización: Ingeniero software.

Técnicas

Sesiones de entrevistas y reingeniería de fuentes.

Ejemplo práctico

En este ejemplo se muestra, de forma muy breve, un fragmento de un informe en el que se consideran los problemas referidos a la presencia de fenómenos multilingües sin vinculación dentro de una fuente de información. Asimismo, se establecen diferentes vías de actuación con objeto de solucionar esta problemática y realizar un proceso de depuración o curación de las fuentes.

Informe sobre fenómenos multilingües

Los problemas presentados en la tabla referida a "*relación entre fenómenos asociados*" derivan de la consideración de este tipo de información como fenómenos independientes, a pesar de su manifiesta vinculación. A esto se le une que los sistemas de información que soportan las diferentes bases de datos con las que cuenta este proyecto realizan procesos de búsqueda a través de palabras clave. Estos factores repercuten en procesos de búsquedas lentos e ineficientes consecuencia del escaso o nulo tratamiento de esta información. Las propuestas de solución a esta problemática que este informe considera son:

1. Adición de una nueva columna a las tablas de las bases de datos para la introducción de la componente multilingüe asociada a cada tipo de fenómeno.
2. Creación de un listado de genéricos de los diferentes fenómenos geográficos en múltiples idiomas y establecer vínculos entre estos.
3. Adición de multilingüismo a la ontología individual, ontologías individuales interconectadas o redes de ontologías diseñadas (un ejemplo de esta propuesta se muestra en el Capítulo 6 de este trabajo) y gestión de la información multilingüe a través de estas información añadida y los *mappings* establecidos con la base de datos.

De estas tres propuestas, únicamente, las dos primeras suponen una intervención directa sobre la fuente (base de datos). Sin embargo, la última propuesta conlleva un proceso de depuración de las fuentes de información mediante procesos externos que no repercuten en cambios directos.

5.5.8. Actividad 8. Mantenimiento

El conocimiento no es algo estático sino que cambia con el transcurrir del tiempo. Desde la óptica de los sistemas de información, estos siempre tienden a cambiar y evolucionar, algunos son más frecuentes y otros menos. Esto hace que los procesos sean vistos como un desarrollo continúo.

El mantenimiento del proceso de integración de información no es esencialmente diferente de su proceso de desarrollo. La principal diferencia es que, por razones organizacionales, este es llevado a cabo por otras personas (Schreiber *et al.*, 1999). Este es el principal motivo por el que se pone especial atención en la documentación de cada una de las diferentes actividades y tareas propuestas por esta metodología.

Esta actividad horizontal, reflejada en la Figura 5.2, está inspirada por las tareas propuestas en la ingeniería del software, concretamente MÉTRICA versión 3 (MÉTRICA v.3) y *Rational*

Unified Process (RUP) (Kruchten, 2000), y por diferentes propuestas de ingeniería del conocimiento, tales como (Gómez-Pérez *et al*., 1997) y (Schreiber *et al*., 1999).

En esta metodología de integración de datos geográficos, si se llevan a cabo las diferentes actividades y sus respectivas tareas como se propone en esta metodología, se puede esperar que la mayoría de las actividades de mantenimiento estén ligadas a tres eventos concretos:

1. ***Actualización de información***. Los procesos de actualización de información geoespacial son habituales, derivados de los cambios asociados a la realidad geográfica. De esta manera, un proceso de estas características puede conllevar el añadido de nueva información con repercusión en el nivel conceptual, en los datos o en ambos niveles al mismo tiempo. En lo concerniente al proceso de mantenimiento, este evento puede implicar reproducir el proceso marcado por esta propuesta metodológica desde tres etapas distintas:

   a. Desde la actividad 3 (*integración estructural*) si se producen cambios y/o añadidos de nuevos conceptos asociados al dominio.
   b. Desde la actividad 2 (*análisis de fuentes*), concretamente desde la tarea 2.3 (*descripción de los datos*), si se llevan a cabo cambios profundos en los datos.
   c. Desde la actividad 4 (*integración semántica*) si se produce adición de nuevos datos.

2. ***Adición de una nueva fuente de datos***. El añadido de una nueva fuente de datos dentro de las diferentes fuentes de información integradas previamente conlleva un inicio de las tareas asociadas a esta propuesta metodológica.
3. ***Cambios en las bases de datos***. Asociado a la evolución de los conocimientos están las formas de modelar los dominios. Esta evolución puede conllevar cambios en los procesos de modelado u optimización de los modelos contenidos en las fuentes de información. Desde la perspectiva del proceso de mantenimiento, los cambios que se produzcan en las fuentes de información requieren recurrir a la actividad 2 (*análisis de fuentes*) de esta metodológica.

El material de entrada y los resultados de salida de cada uno de estos eventos dentro del proceso de mantenimiento viene marcado por cada una de las actividades y/o tareas que dichos eventos tomen como punto de partida. Una situación similar sucede con los responsables de realizar las diferentes tareas del proceso de integración que estos eventos llevan asociadas.

5.6. Conclusiones

En este capítulo se ha presentado nuestra propuesta metodológica para la integración basada en ontologías de información de bases de datos heterogéneas en el dominio geográfico. Esta metodología surge ante la ausencia de metodologías que puedan ser aplicadas en este área de conocimiento.

Desde una perspectiva general, esta metodología contribuye a alcanzar interoperabilidad/integración estructural y semántica. Además, en el plano sintáctico da un paso adelante al utilizar los lenguajes de la Web Semántica (RDF y OWL), que poseen mayor expresividad que XML, ampliamente utilizado en el contexto de la interoperabilidad (sintáctica) geoespacial. Esta interoperabilidad/integración permite ofrecer una visión integrada de la realidad geoespacial abstraída y almacenada en diferentes fuentes de información. Esta metodología consta de 7 actividades inspiradas e influenciadas por las buenas prácticas existentes en otros acercamientos metodológicos. Dichas actividades pueden dividirse en tres niveles, que son: análisis, integración y recomendación.

En el nivel de análisis se encuentran las actividades 1 y 2 de esta metodología. La actividad 1 está orientada a comprobar la **viabilidad del proyecto** de integración. Este proceso de análisis se apoya en 9 tareas, algunas de ellas opcionales, descritas en la sección 5.5.1. En una segunda fase (sección 5.5.2), se procede al **análisis** de las características de las **fuentes de información** que componen el proceso de integración. El objetivo de esta actividad, apoyada en 5 tareas, es obtener una descripción y comparación entre los distintos esquemas y datos.

Finalizado el análisis, se procede a la realización del proceso de integración de las fuentes de información. Este nivel está compuesto por las actividades 3, 4 y 5 (sección 5.5.3 – 5.5.5).

En la actividad 3 se realiza la **integración estructural** de la información a través del desarrollo de ontologías. Esta actividad puede resultar opcional, consecuencia de la posibilidad de utilizar recursos ontológicos existentes. Esta actividad, dependiendo de las necesidades, propone la utilización de diversas metodologías para la construcción de ontologías.

El proceso de **integración semántica** se lleva a cabo en la actividad 4. Esta actividad, apoyada en 4 tareas, se centra en la generación de *mappings* entre el recurso ontológico (proveniente de la actividad anterior) y las diversas fuentes de información implicadas. Una vez establecidos estos *mappings*, se realizan diversas tareas, previas a la explotación, destinadas a mejorar los resultados obtenidos (conjuntos de datos RDF).

La **publicación de datos integrados** se realiza en la actividad 5. Esta actividad puede resultar opcional en función de las características del proyecto de información. El objetivo principal de esta tarea es publicar conjuntos de datos geoespaciales conforme a los principios de *Linked Data*.

En la actividad 6 se realiza una **depuración o post-procesamiento** de la información RDF obtenida de la actividad anterior. El objetivo principal de esta actividad, compuesta por 3 tareas, es la detección y desambiguación de duplicados de instancias. La finalidad es ofrecer una visión integrada de los conjuntos de datos geoespacial almacenados en las diversas fuentes de información implicadas en el proyecto.

Tras el proceso de integración, se procede a realizar una serie de **recomendaciones** (sección 5.5.7) a la/s organización/es implicadas en el proceso de integración. El objetivo es transmitir y promover una serie de buenas prácticas y propuestas de depuración extraídas como conclusiones del proyecto de integración. Esta actividad es consecuente con los principios INSPIRE relacionados con el almacenamiento y mantenimiento de las fuentes de información al nivel más adecuado, de ahí que únicamente se realicen recomendaciones y no modificaciones directas.

En la actividad 8, consecuencia de la evolución del conocimiento, la metodología propone (sección 5.5.8) un **proceso de mantenimiento**, vinculando dicho proceso, principalmente, a 3 eventos concretos.

Finalmente, para demostrar la aplicabilidad de esta metodología se realiza un caso de uso para el dominio de la información hidrográfica. Algunos ejemplos vinculados con las distintas tareas de esta metodología se recogen a lo largo de este capítulo. Asimismo, gran parte de la información contenida en dichos ejemplos se amplía en los Capítulos 6, 7, 8 y 9 de este trabajo.

Capítulo 6

Desarrollo de recursos ontológicos de dominio en el contexto de la información hidrográfica

"Una realidad que vista desde cualquier punto resultase siempre idéntica es un concepto absurdo".

Ortega y Gasset (1939)

La diversidad de fuentes de conocimiento (catálogos, diccionarios de datos, tesauros, etc.) existentes en la comunidad geoespacial a diferentes escalas (internacional – nacional – regional – local) pone de manifiesto los problemas de estructuración y heterogeneidad presentes en las diferentes fuentes. Esta diversidad es especialmente relevante en el dominio hidrográfico, donde destaca la presencia, en diferentes fuentes, de fenómenos universalmente conocidos con definiciones dispares, junto a otras peculiaridades propias de la IG y de la realidad geoespacial. Una descripción detallada de los diversos problemas existentes vinculados con la heterogeneidad de la información geográfica se recoge en el Capítulo 2 (sección 2.2) de este trabajo.

Ante la necesidad de conseguir la armonización semántica de esta información, en este trabajo se pretende establecer un marco de referencia, común y compartido a través del desarrollo de recursos ontológicos para el dominio hidrográfico. Estos recursos pretenden solventar los problemas de estructuración y heterogeneidad existentes en las múltiples fuentes para, de esta manera, alcanzar la interoperabilidad semántica de la información en este área de conocimiento.

Este capítulo describe el proceso de desarrollo de una ontología, denominada *hydrOntology*, y su posterior evolución hacia una red de ontologías. Así, en la sección 6.1 se describe el proceso de desarrollo seguido para la construcción de los recursos ontológicos. En la sección 6.2 se presenta el contexto metodológico en el que se lleva a cabo el proceso de desarrollo. Una descripción detallada de *hydrOntology* como ontología individual se recoge en la sección 6.3. Asimismo, en la sección 6.4 se describen los resultados del proceso de evolución de la ontología individual a la red de ontologías en torno a *hydrOntology*. Finalmente, en la sección 6.5, se recogen las conclusiones del trabajo descrito en este capítulo.

6.1. Proceso de desarrollo

Esta sección describe el proceso de desarrollo y evolución de *hydrOntology*. Una visión general del conjunto de pasos llevados a cabo en este proceso se refleja en la Figura 6.1. A continuación se detallan algunas de las características de los mencionados pasos del proceso de de desarrollo.

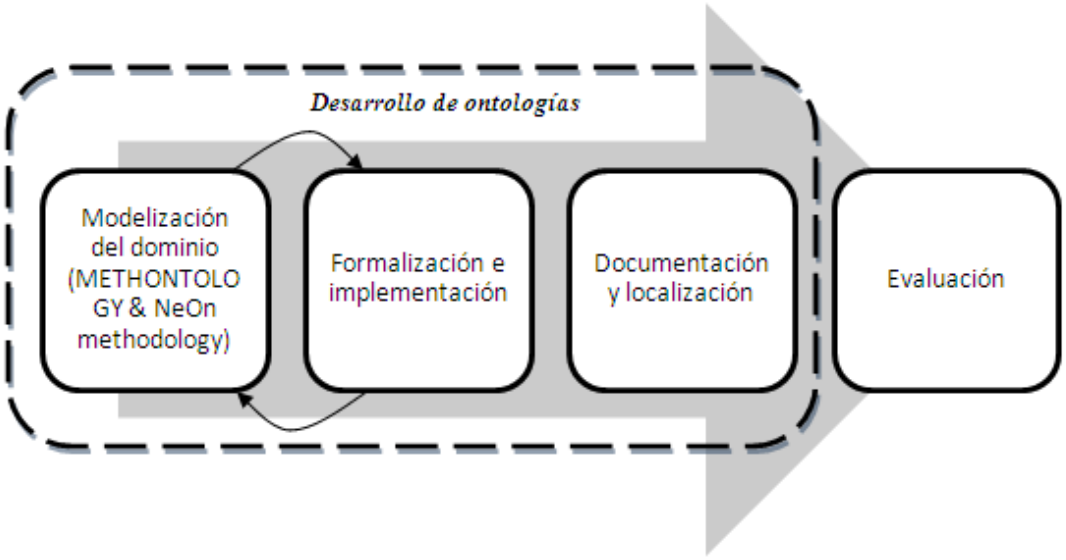


Figura 6.1. Visión general del proceso de desarrollo de *hydrOntology*.

Vinculadas a la tarea 3.1 (*desarrollo de ontologías*) de la metodología propuesta, aparecen las siguientes subtareas:

- Estudio y modelización del dominio hidrográfico conforme a las tareas marcadas por las metodologías de desarrollo de ontologías elegidas.
    - En un primer momento, se eligió METHONTOLOGY (Fernández-López *et al.*, 1999) como metodología de construcción de ontologías. Esta metodología se centra en el desarrollo de ontologías individuales, enfatizando la reutilización de ontologías de dominio y alto nivel existentes. Asimismo, propone la utilización de un conjunto de representaciones intermedias que pueden ser transformadas automáticamente, con posterioridad, en diferentes lenguajes formales.
    - En un segundo estadio, consecuencia de la evolución del propósito y alcance de la ontología, se adopta la metodología NeOn (Suárez-Figueroa, 2010). Este cambio metodológico supone un cambio de perspectiva en el proceso de desarrollo, ya que se pasa de la construcción de una ontología individual a una red de ontologías. Una descripción detallada de este proceso de cambio metodológico se muestra en la sección 6.4 de este capítulo.

- A la realización de las diferentes tareas propuestas en ambos acercamientos metodológicos prosigue la formalización de la ontología individual, en un primer momento, y, con posterioridad, de la red de ontologías en lógica de descripciones y sus correspondientes procesos de implementación en OWL mediante la utilización de la herramienta de creación y edición de ontologías Protégé[1].
- Tras el proceso de formalización e implementación, se realiza una documentación exhaustiva de la red de ontologías creada. Este proceso de documentación se caracteriza

[1] Protégé. http://protege.stanford.edu/

por el añadido de información correspondiente a: sinónimos, definiciones en lenguaje natural, fuente de procedencia de estas definiciones y definición de la procedencia de conceptos y sinónimos para cada una de las clases y propiedades que contiene dicha red de ontologías. Asimismo, en esta fase del proceso de desarrollo de la red de ontologías entorno a *hydrOntology* se procede a la inclusión de información multilingüe.

Finalmente, en relación con la tarea 3.2 de la mencionada metodología, se realiza la evaluación de la ontología desarrollada. Esta evaluación estará compuesta por la verificación formal del lenguaje, evaluación de la consistencia lógica y la evaluación conforme a los requisitos y cuestiones de competencia.

6.2. Contexto metodológico

Como se ha descrito anteriormente, el proceso de desarrollo de *hydrOntology* se divide en dos fases diferenciadas tanto en sus objetivos como en las metodologías seguidas.

Inicialmente, el desarrollo de *hydrOntology* se inició con el objetivo de crear una ontología local individual (Suárez-Figueroa *et al.*, 2008). Desde el punto de vista metodológico el desarrollo de *hydrOntology* se inicia conforme a las pautas de METHONTOLOGY (Gómez-Pérez *et al*., 2003). La Figura 6.2 muestra las tareas para la conceptualización de ontologías sugeridas en el marco de esta propuesta metodológica. Asimismo, esta figura enfatiza los componentes de la ontología (conceptos, atributos, relaciones, constantes, axiomas formales, reglas e instancias) construidos en cada tarea. Además, esta figura ilustra los pasos que esta metodología propone para la creación de tales componentes durante la actividad de conceptualización. El proceso reflejado en esta figura no es un proceso de modelado secuencial, aunque cierto orden debe ser seguido para asegurar la consistencia y completitud de la representación de conocimiento (Corcho *et al.*, 2005).

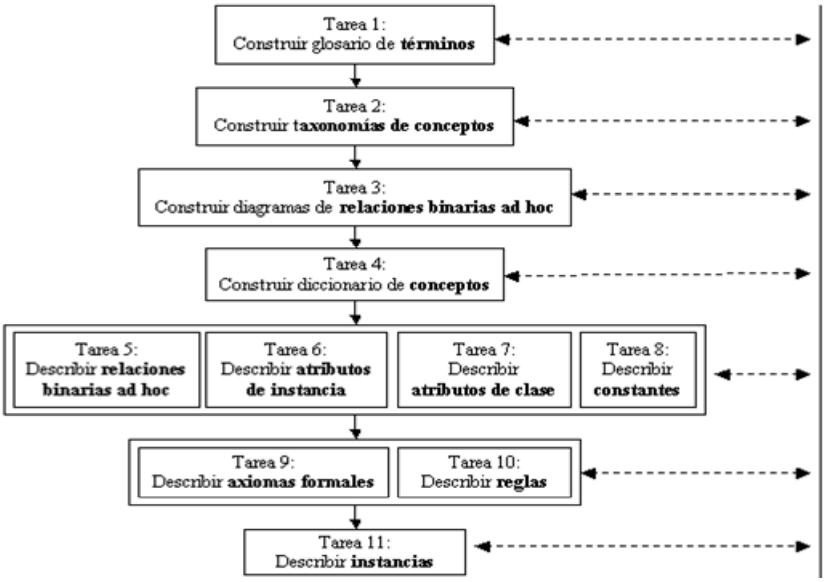


Figura 6.2. Tareas de la actividad de conceptualización conformes a METHONTOLOGY (Gómez-Pérez *et al*., 2003)

METHONTOLOGY propone un conjunto de tareas para capturar el conocimiento del dominio (Gómez-Pérez *et al*., 2003). Estas tareas pueden dividirse en tres grupos:

- El primer grupo está dirigido a delimitar y estructurar el dominio mediante el conjunto de tareas que van de la tarea 1 a la 4.

- El segundo grupo abarca de la tarea 5 a la 7. Este grupo de tareas ayudaría a documentar el conocimiento adquirido de las tareas previas.
- Finalmente, METHONTOLOGY propone otras tareas, de la tarea 8 a la 11, para completar el conocimiento. Este conjunto de tareas no fue llevado a cabo en el proceso de desarrollo de esta ontología global, consecuencia del cambio en la metodología seguida.

En el proceso de desarrollo de *hydrOntology* el conjunto de las tareas anteriores fueron llevadas a cabo de una manera pragmática. Este hecho supone un seguimiento no exhaustivo de las diferentes tareas mencionadas, especialmente, en lo que concerniente a la compleción de las tablas que proponen distintas tareas en esta metodología.

En un segundo estadio del proceso de desarrollo de *hydrOntology* se produce una evolución del propósito de esta ontología. Así, el objetivo pasa a ser la construcción de una ontología global de dominio cuya finalidad es cubrir la mayoría de los fenómenos asociados al dominio hidrográfico representables cartográficamente y servir como marco de armonización entre los diferentes productores de información geoespacial en el entorno nacional e internacional. Este cambio de visión conlleva un proceso de evolución metodológica y del tipo de ontología desarrollada, manifiesto en el paso de METHONTOLOGY a la metodología propuesta por NeOn (Suárez-Figueroa, 2010) y en el paso de la construcción de una ontología individual a una red de ontologías, respectivamente. La construcción de esta red de ontologías está vinculada con los tipos de integración de ontologías propuestos en (Fonseca *et al*., 2002). La relación con esta propuesta de integración se deriva de la integración de recursos ontológicos pertenecientes al mismo dominio (hidrografía) y a diferentes áreas concernientes o estrechamente vinculadas al dominio geoespacial (relaciones espaciales, metadatos, mereología, etc.).

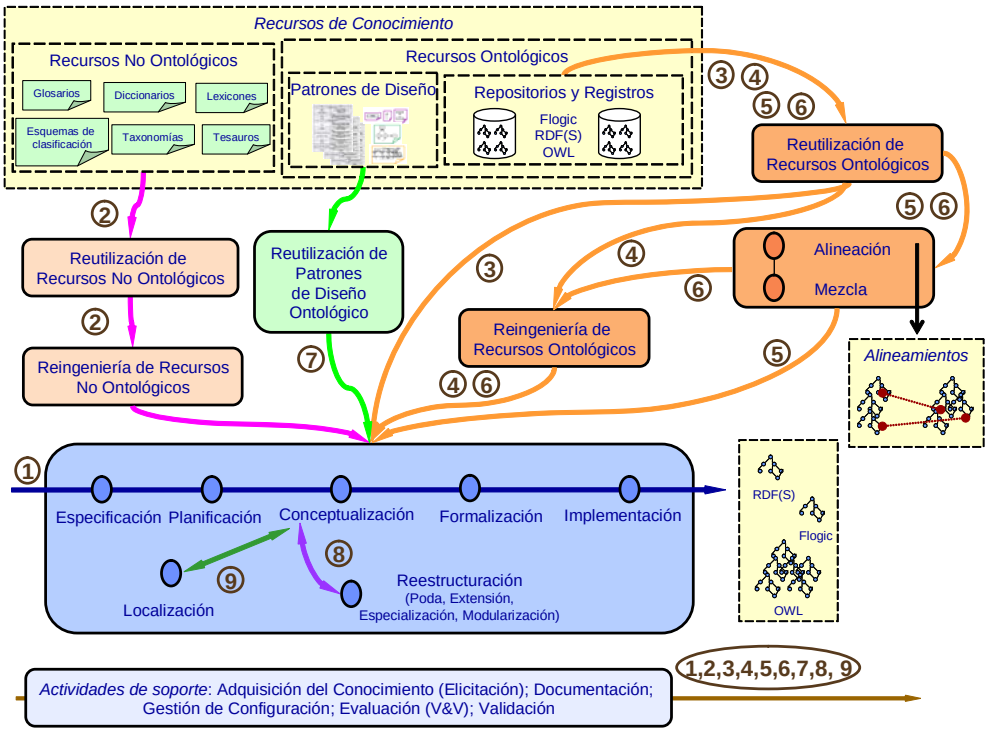


Figura 6.3. Escenarios para la construcción de ontología y redes de ontologías (Suárez-Figueroa, 2010)

El enfoque metodológico propuesto por la metodología NeOn requiere el seguimiento de una serie de escenarios. En la Figura 6.3 se muestra una visión general de los diferentes escenarios mencionados con anterioridad. Una descripción detallada de estos escenarios se encuentra en

(Suárez-Figueroa, 2010). En el contexto del desarrollo de la red de ontologías en torno a *hydrOntology* los escenarios adoptados se detallan en la sección 6.4.4 de este trabajo.
Descrito el marco metodológico, a continuación se recoge una descripción de las características que marcan el proceso de desarrollo de la ontología individual (*hydrOntology*) y de la red de ontologías.

6.3. *hydrOntology*

En esta sección se recoge una descripción detallada de las características del proceso de desarrollo de *hydrOntology* como ontología individual.

6.3.1. Especificación de requisitos

El objetivo de la labor de especificación es recoger, en lenguaje natural, la siguiente información:

Propósito: El desarrollo de esta ontología está destinado al establecimiento de correspondencias entre diferentes fuentes de datos del Instituto Geográfico Nacional (catálogos de fenómenos, nomenclátores, etc.) con objeto de conseguir una visión armonizada de los diferentes fenómenos hidrográficos, recogidos en las diversas fuentes, caracterizados por ser representables cartográficamente.

Nivel de formalización: Esta ontología debe ser implementada conforme a la especificación de OWL (Dean *et al.*, 2004), ya que este lenguaje es una recomendación técnica del W3C Consortium para la Web Semántica.

6.3.2. Preguntas de competencia

Las preguntas de competencia (ver Tabla 6.1) que esta ontología tiene que conseguir responder son las siguientes:

Tabla 6.1. Preguntas de competencia de *hydrOntology*

1.	¿Qué afluentes alimentan a los ríos de la cuenca “x”?
2.	¿Qué afluentes tiene el río “x”?
3.	¿Cuáles son los elementos de los que consta el río “x”?
4.	¿Embalses que forman parte del río “x”?
5.	¿Los ríos “x”, “y” y “z” de qué río son afluentes y a qué acuíferos alimentan?
6.	¿Qué ríos son conducidos por una tubería?
7.	¿Qué lagunas son originadas por un glaciar?
8.	¿Qué aguas alimentan a la albufera “x” y qué caños la comunican con el mar?
9.	¿Qué río alimenta a la marisma “x” y dime si origina algún lucio y cuáles son?
10.	¿Qué embalse almacena las aguas del río “x” y qué arroyos alimentan a dicho río?
11.	¿Qué esteros son inundados por una laguna / aguas costeras?
12.	¿Qué ríos tienen un sumidero que convierten su caudal en una corriente subterránea?

6.3.3. Fuentes de conocimiento

El proceso de adquisición de conocimiento para el desarrollo de *hydrOntology* considera múltiples fuentes de información. Estas fuentes están conformadas por diferentes recursos no ontológicos (catálogos de fenómenos, diccionarios de datos, tesauros, modelos UML (*Unified Modeling Language*), nomenclátores, etc.) relacionados, mayoritariamente, con el dominio de la información geoespacial. Asimismo, una característica común de estas fuentes es la presencia de información vinculada al dominio hidrográfico. Estos recursos se caracterizan por cubrir un amplio espectro en lo que a fuentes de información geográfica se refiere, así como por su procedencia diversa.

El proceso de selección de recursos no ontológicos que forman parte de *hydrOntology* se lleva a cabo considerando criterios dependientes e independientes del dominio de trabajo:

- **Criterios dependientes**
  - Existencia de relación estrecha con el dominio de conocimiento (dominio hidrográfico) y/o la vinculación con el área de la información geográfica.
  - El peso de las fuentes de información asociado a la procedencia, difusión y repercusión de las mismas en la comunidad geoespacial.
  - Consideración de la escala (granularidad) asociada a cada fuente de información, consecuencia del proceso de filtrado de la información en la representación cartográfica.
- **Criterios independientes**
  - Cantidad de conocimiento (Poveda, 2009) reutilizable.
  - Esfuerzo de transformación de un recurso no ontológico en una ontología (Poveda, 2009). Este criterio no se refiere sólo al proceso de automatizar la transformación de un recurso no ontológico en una ontología, sino también al esfuerzo requerido para evaluar el resultado de dicha transformación y añadir manualmente el conocimiento semántico que no se haya representado de manera automática.

A continuación se listan, de forma exhaustiva, las fuentes de información utilizadas y se describen las principales fuentes utilizadas como base en el proceso de desarrollo siendo diferenciadas por su nivel de procedencia. En la Tabla 6.2 se recoge una visión general de los recursos no ontológicos utilizados en el desarrollo de *hydrOntology*.

En el **marco internacional** las fuentes de información utilizadas en el proceso de desarrollo de *hydrOntology* se detallan a continuación:

- Dentro de estos recursos, una de las principales fuentes de información utilizadas son los **tesauros**. Uno de los tesauros más conocido es el *Getty Thesaurus of Geographic Names* (TGN) (Harpring, 1997; Harpring *et al.*, 2006). En TGN, cada registro muestra las coordenadas espaciales asociadas, la posición jerárquica, una breve explicación de la región y los nombres de lugares asociados. Además, los nombres para un lugar pueden incluir nombres en la lengua vernácula, inglés, otros idiomas, nombres históricos. Para más detalles, sobre éste y otros recursos, puede consultarse la recopilación de tesauros geográficos e históricos realizada en (Vilches-Blázquez *et al.,* 2007c).
- *Alexandria Digital Library Feature Type Thesaurus*[2] contiene un esquema jerárquico de términos descritos como entradas de nombres de lugares geográficos. Está previsto para un uso general para muchos tipos de nomenclátores y vocabularios compartidos con objeto de conseguir interoperabilidad entre nomenclátores. Este nomenclátor cubre términos de temáticas diversas, tales como: administrativas, hidrografía, fisiografía y lugares regionales. En su última versión (Julio, 2002) recoge 210 términos preferentes, 1046 términos no preferentes.
- Otra serie de tesauros utilizados en el proceso de construcción de *hydrOntology* son *UNESCO Thesaurus*[3], GEMET[4].

Otra de las fuentes de información que forman parte de *hydrOntology* son los **catálogos de fenómenos**. Dentro de esta tipología de fuentes destacan los siguientes recursos:

- El proyecto *EuroRegionalMap* (ERM) presenta un catálogo de fenómenos asociado a una base de datos topográfica vectorial paneuropea a escala 1:250.000. La información

[2] *ADL*. http://www.alexandria.ucsb.edu/~lhill/FeatureTypes/ver070302/index.htm
[3] *UNESCO thesaurus*. http://www2.ulcc.ac.uk/unesco/
[4] *GEMET thesaurus*. http://www.eionet.europa.eu/gemet

presente en este catálogo está basada en la armonización de las bases de datos geográficas nacionales de los diversos Organismos Cartográficos y Catastrales Nacionales de Europa existentes. *EuroGlobalMap* (EGM) posee un catálogo de datos topográficos que cubre toda Europa a escala 1:1.000.000, producido en cooperación con los Organismos Cartográficos Nacionales (*National Mapping Agencies* – NMAs) de Europa, usando bases de datos oficiales.

Tabla 6.2. Recursos no ontológicos utilizados en el desarrollo de *hydrOntology*.

Tipo de fuente	Ámbito Internacional	Ámbito Nacional	Ámbito Local
Catálogos de fenómenos y/o Diccionarios de datos	EuroGlobalMap (EGM) EuroRegionalMap (ERM) EuroGeonames (EGN) DIGEST Catálogo Gfco de Colombia GeoAlex	BCN25 BCN200 BTN25 Comisión de Normas Cartográficas (CNC)	Xunta de Galicia ICC - Cataluña Navarra La Rioja Junta de Extremadura Gobierno Vasco Castilla y León
Nomenclátores	GeoNames	Nomenclátor Gfco. Conciso Nomenclátor NOMGEO	-
Tesauros	*Alexandria Digital Library Feature Type Thesaurus* *UNESCO Thesaurus* *Getty Thesaurus of Geographic Names* GEMET	Tesauro CEDEX	-
Legislación & Estándares	INSPIRE *Directive* *Water Framework Directive Data Specifications – deliverable D2.3* *Swedish Standard – Surface water systems*	Real Decreto Legislativo 1/2001, de 20 de julio, por el que se aprueba el texto refundido de la Ley de Aguas.	-
Glosarios	Glosario Términos Gfcos - Chile Glosario Hidrológico Internacional - UNESCO	-	-

En esta tipología de fuentes de información se recogen fuentes como EuroGeonames (EGN)[5], The Digital Geographic Information Exchange Standard (DIGEST)[6], Catálogo de Objetos Geográficos Básicos de Colombia y el catálogo de fenómenos del proyecto GeoAlex.

- Con respecto a los **recursos legislativos** y/o **estándares** se consideran: Por un lado, dos directivas europeas; la Directiva Europea por la que se establece una infraestructura de información espacial en la Comunidad Europea (INSPIRE) (Commission of the European Communities, 2007) y la Directiva Europea por la que se establece un marco comunitario de actuación en el ámbito de la política de aguas (European Parliament and the Council of the European Union, 2000). Por otro lado, se recoge información procedente del *Data Specifications* de INSPIRE (Drafting Team "Data Specifications", 2007) y del *Swedish Standard* con información sobre "*surface water systems*".

[5]*EGN*. http://www.eurogeographics.org/eng/03_projects_EuroGeoNames.asp
[6]*DIGEST*. https://www.dgiwg.org/digest/

- Finalmente, en *hydrOntology* se recoge información procedente de diferentes **glosarios** de carácter internacional. Así, la información que aparece en esta ontología proviene del Glosario Hidrológico Internacional[7] (perteneciente a la UNESCO) y del Glosario de Términos Geográficos (Biblioteca del Congreso Nacional de Chile)[8].

En el **ámbito nacional** las fuentes que forman parte de la base de conocimiento de *hydrOntology* proceden del Instituto Geográfico Nacional (IGN), productor oficial de IG en este ámbito. El IGN dispone, entre otras, de las siguientes formas de estructuración de conocimiento: Nomenclátor Conciso (NC) (1:1.000.000), Nomenclátor NOMGEO (1:50.000), Base Cartográfica Numérica a escala 1:200.000 (BCN200) y la Base Topográfica Numérica a escala 1:25.000 (BTN25), anteriormente denominada BCN25. Estas fuentes se mantienen separadamente y presentan gran heterogeneidad en diferentes aspectos. A continuación se va a proceder a describir cada una de estas fuentes de manera más detallada.

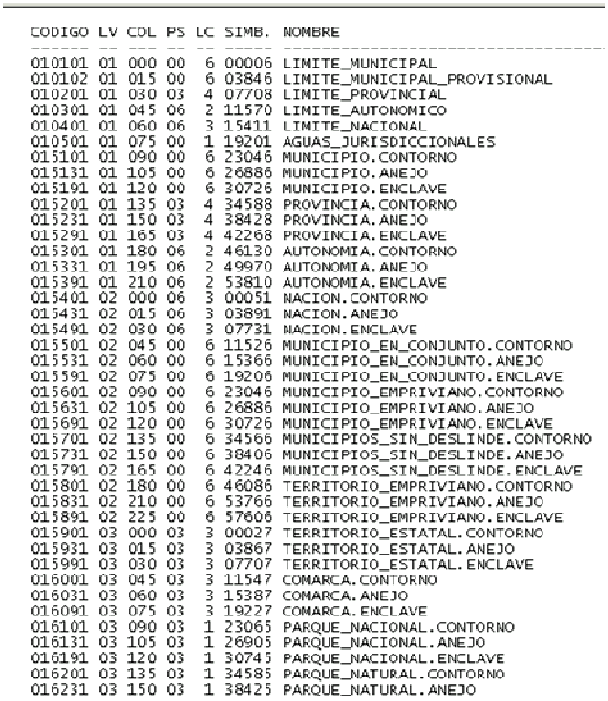
```
! Tipo_ dgn...NNSCCCGG                    Codigo_bcn...TTGGSS
!       NN  : Nivel elemento                      TT : Tema
!       S   : Estilo linea dgn                    GG : Grupo
!       CCC : Color  linea dgn                    SS : Subgrupo
!       GG  : Grosor linea dgn
!
! Entidad                                  Tipo_istram....???
!       104 : polilínea
!       203 : célula se convierte a símbolo
!        -1 : célula se explota en sus componentes
!       304 : rótulo
!
! Grupo
!         0 : sin determinar
!         1 : carreteras
!         2 : hidrografía
!         3 : conducciones
!         4 : administrativo
!
!         En textos el grupo corresponde a la fuente Microstation + Mayúsc
!
! Cerrado
!        en lineas                              en textos
!                1 : perimetral                        n : altura
!                0 : entidad lineal abierta
!               -1 : cultivo perimetral
!               -2 : cultivo línea abierta
! Trato
!   I: Intocable  A: Altimetría  N: No tratar  T: Textos Asociados
!   S: Textos Sueltos  C: Cultivo  F: Solo salida !: Tratar normalmente
!                                 TTGGSS
02000900   104     1   0   090101   1   !I   Marco de hoja
02300902   104     2   0   100200   0   !    Base Geodésica de Madridejos
06003900   104     3   0   025102   0   !    Acantilado
06006900   104     4   0   025302   0   !    Costa rocosa no acantilada
06009900   104     5   2   037402   1   !    Playa fluvial de guijarros. C
06012900   104     6   0   025501   1   !    Lavas. Contorno
06015900   104     7   0   058303   0   !I   Dique de hormigón >15 metros
06018900   104     8   0   058304   0   !I   Dique de hormigón < 15 metros
07013400   104     9   0   058302   0   !I   Dique de tierra
07016400   104    10   0   055401   1   !    Vertedero. Contorno
11003003   104    11   1   062202   0   !    Autopista. Enlace
11012000   104    12   0   056001   1   !I   Patio. Contorno
13003300   104    13   1   060101   0   !    Autopista. Eje
13303300   104    14   1   060131   0   !    Autopista en Construcción. Eje
14002401   104    15   1   066901   1   !I   Puesto de S.O.S.
14003301   104    16   1   067901   1   !I   Peaje
15003003   104    17   1   062204   0   !    Autovía. Enlace
15003004   104    18   1   060701   0   !    Autovía
```

Figura 6.4. Fenómenos en BCN25

- La Base Cartográfica Numérica[9] a escala 1:25.000, denominada BCN25, se diseñó como un producto derivado del Mapa Topográfico Nacional creado a la escala mencionada (MTN25). Fue construida para obtener información geográfica que cumpliera con los requisitos de datos requeridos para ser explotados dentro del entorno de los Sistemas de Información Geográfica (SIG). Por lo tanto, BCN25 contiene esencialmente la misma información que MTN25, con propiedades topológicas y geométricas adicionales, siguiendo un modelo específico orientado a base de datos y un catálogo de fenómenos (Rodríguez *et al.*, 2005). La Figura 6.4 muestra una pequeña parte del catálogo de fenómenos de BCN25.

[7] http://www.cig.ensmp.fr/~hubert/glu/HINDES.HTM
[8] http://www.bcn.cl/siit/glosario
[9] Una *Base Cartográfica Numérica* (BCN) es un conjunto de datos concebido y orientado para su carga en un SIG, que contiene toda la información relevante representada en la cartografía a la misma escala. En la concepción de las BCNs, se mantiene todo el caudal de conocimiento contenido en la cartografía clásica, en cuanto a selección y modelización del mundo real, y se adapta a las tecnologías más modernas, que ofrecen mecanismos de análisis de los datos más potentes que el ojo humano. En consecuencia la representación cartográfica clásica (simbolización) de las entidades geográficas se sustituye por una estructura de datos eficiente, manteniendo siempre la semántica del mapa. *Fuente*: Instituto Geográfico Nacional.

- Por otro lado, la primera versión de la Base Cartográfica Numérica a escala 1:200.000 (BCN200) se inició en 1985. Este trabajo se realizó mediante la digitalización y edición de la información contenida en mapas analógicos provinciales a la escala citada (Sevilla *et al.*, 2006). La primera versión de esta Base de Datos se completó en el año 1.992 y se actualiza anualmente. En la Figura 6.5 se recoge parte del catálogo de fenómenos de BCN200:



CODIGO	LV	CDL	PS	LC	SIMB.	NOMBRE
010101	01	000	00	6	00006	LIMITE_MUNICIPAL
010102	01	015	00	6	03846	LIMITE_MUNICIPAL_PROVISIONAL
010201	01	030	03	4	07708	LIMITE_PROVINCIAL
010301	01	045	06	2	11570	LIMITE_AUTONOMICO
010401	01	060	06	3	15411	LIMITE_NACIONAL
010501	01	075	00	1	19201	AGUAS_JURISDICCIONALES
015101	01	090	00	6	23046	MUNICIPIO.CONTORNO
015131	01	105	00	6	26886	MUNICIPIO.ANEJO
015191	01	120	00	6	30726	MUNICIPIO.ENCLAVE
015201	01	135	03	4	34588	PROVINCIA.CONTORNO
015231	01	150	03	4	38428	PROVINCIA.ANEJO
015291	01	165	03	4	42268	PROVINCIA.ENCLAVE
015301	01	180	06	2	46130	AUTONOMIA.CONTORNO
015331	01	195	06	2	49970	AUTONOMIA.ANEJO
015391	01	210	06	2	53810	AUTONOMIA.ENCLAVE
015401	02	000	06	3	00051	NACION.CONTORNO
015431	02	015	06	3	03891	NACION.ANEJO
015491	02	030	06	3	07731	NACION.ENCLAVE
015501	02	045	00	6	11526	MUNICIPIO_EN_CONJUNTO.CONTORNO
015531	02	060	00	6	15366	MUNICIPIO_EN_CONJUNTO.ANEJO
015591	02	075	00	6	19206	MUNICIPIO_EN_CONJUNTO.ENCLAVE
015601	02	090	00	6	23046	MUNICIPIO_EMPRIVIANO.CONTORNO
015631	02	105	00	6	26886	MUNICIPIO_EMPRIVIANO.ANEJO
015691	02	120	00	6	30726	MUNICIPIO_EMPRIVIANO.ENCLAVE
015701	02	135	00	6	34566	MUNICIPIOS_SIN_DESLINDE.CONTORNO
015731	02	150	00	6	38406	MUNICIPIOS_SIN_DESLINDE.ANEJO
015791	02	165	00	6	42246	MUNICIPIOS_SIN_DESLINDE.ENCLAVE
015801	02	180	00	6	46086	TERRITORIO_EMPRIVIANO.CONTORNO
015831	02	210	00	6	53766	TERRITORIO_EMPRIVIANO.ANEJO
015891	02	225	00	6	57606	TERRITORIO_EMPRIVIANO.ENCLAVE
015901	03	000	03	3	00027	TERRITORIO_ESTATAL.CONTORNO
015931	03	015	03	3	03867	TERRITORIO_ESTATAL.ANEJO
015991	03	030	03	3	07707	TERRITORIO_ESTATAL.ENCLAVE
016001	03	045	03	3	11547	COMARCA.CONTORNO
016031	03	060	03	3	15387	COMARCA.ANEJO
016091	03	075	03	3	19227	COMARCA.ENCLAVE
016101	03	090	03	1	23065	PARQUE_NACIONAL.CONTORNO
016131	03	105	03	1	26905	PARQUE_NACIONAL.ANEJO
016191	03	120	03	1	30745	PARQUE_NACIONAL.ENCLAVE
016201	03	135	03	1	34585	PARQUE_NATURAL.CONTORNO
016231	03	150	03	1	38425	PARQUE_NATURAL.ANEJO

Figura 6.5. Fenómenos en BCN 200

- La información contenida en ambas Bases Cartográficas Numéricas está estructurada en ocho temas diferentes (límites administrativos, relieve, hidrografía, vegetación, edificaciones, comunicaciones, líneas de conducción y toponimia). Cada fenómeno está codificado con 3 pares de dígitos: dos para el tema, dos para el grupo (parte de información homogénea estructurada en tópicos) y dos para el subgrupo (una extensión del fenómeno geográfico asociado al grupo). Estos números describen y clasifican diferentes fenómenos de acuerdo con su temática y atributos. El siguiente ejemplo muestra esta estructura:

  030101 Río permanente representado con una línea simple
  Tema 03: Hidrografía
  Grupo: **01** Curso de agua permanente
  Subgrupo: **01** Curso de agua simbolizado con una línea simple

  Junto a los catálogos de fenómenos descritos con anterioridad destacan por su implicación en el desarrollo de *hydrOntology* la información proveniente de la Base Topográfica Numérica (BTN25), así como la procedente del catálogo de la Comisión de Normas Cartográficas (CNC).

- Respecto a los **nomenclátores**[10], en el desarrollo de *hydrOntology* se consideran: Por un lado, el Nomenclátor NOMGEO, también llamado Base de Datos Georreferenciada

[10] *Nomenclátor*. La definición más extendida de este concepto proviene de la norma (ISO, 2003b). Este estándar define un nomenclátor como un directorio de instancias de una clase o clases de fenómenos que contienen alguna información de posición.

de Nombres Geográficos, tiene 44 fenómenos diferentes en español, gallego, catalán, vasco y aranés. En la Figura 6.6 puede verse la clasificación de los fenómenos de este nomenclátor. Por otro, el Nomenclátor Conciso (NC) es un corpus básico de topónimos estandarizados creado por la Comisión Española de Nombres Geográficos. Este nomenclátor sigue las recomendaciones de la Conferencia de Naciones Unidas sobre normalización de nombres geográficos y es conforme con el Modelo de Nomenclátor Español (Rodríguez Pascual *et al*., 2006). Este nomenclátor está constituido por 22 fenómenos, cuya tipología se refleja en la Figura 6.7. Para más detalles, se puede consultar (NGC, 2006).

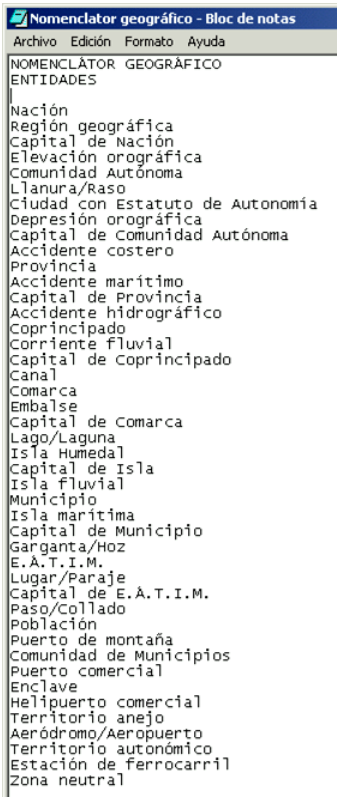


Figura 6.6. Fenómenos del Nomenclátor Geográfico Nacional

Figura 6.7. Fenómenos del Nomenclátor Conciso

- Finalmente, en cuanto a las fuentes de información correspondientes al nivel nacional en el proceso de adquisición de conocimientos de *hydrOntology* participan el tesauro CEDEX[11] y el Real Decreto Legislativo 1/2001, de 20 de julio, por el que se aprueba el texto refundido de la Ley de Aguas.

Derivado del objetivo de abarcar un amplio abanico de fuentes de información para conseguir un proceso exhaustivo de adquisición de conocimiento del dominio se recoge información vinculada al **contexto local** de producción de información. A este nivel la diversidad de fuentes de información también resulta considerable, consecuencia de que algunas Comunidades Autónomas (CCAA) tienen organismos o institutos propios destinados a la producción de IG y cada uno tiene catálogos de información propios. A continuación se describen las características principales de algunos de éstos recursos considerados:

- Castilla y León (Junta de Castilla y León, 2005): Los fenómenos están codificados jerárquicamente con 6 dígitos, agrupados en temas (11), grupos (60) y subgrupos.

[11] http://hispagua.cedex.es/documentacion/tesauro/Tes_hidro/tesauro.php?localizacion=Tesauro%20Ingen ier%EDa%20Civil

- Galicia (Xunta de Galicia, 2004): Los fenómenos están codificados jerárquicamente con 6 dígitos, agrupados en 17 bloques, que se dividen a su vez en sub-bloques. Los nombres de los fenómenos aparecen en gallego.
- Cataluña (Institut Cartogràfic de Catalunya, 2001): Los fenómenos están codificados jerárquicamente con 5 dígitos, el primero de ellos los divide en 7 bloques y los dos siguientes en 61 sub-bloques.
- País Vasco (Gobierno Vasco, 2005): Los fenómenos se llaman capas y se codifican jerárquicamente con 6 dígitos, de los cuales la primera división indica el tema al que pertenecen, habiendo 28 temas posibles.
- Extremadura (Junta de Extremadura, 2005): Los fenómenos están codificados con 5 dígitos, los dos primeros fijos y el tercero indica el tema (10).
- La Rioja (Gobierno de La Rioja, 2005): Los fenómenos están codificados jerárquicamente con 6 dígitos y organizados jerárquicamente en temas (13), grupos y subgrupos.
- Navarra (Gobierno de Navarra, 2000): Los fenómenos están codificados con 4 dígitos y divididos en 16 temas.

6.3.4. Modelado: Enfoque y criterios de estructuración

El proceso de modelado de *hydrOntology* sigue una estrategia *top-down* de desarrollo de ontologías (Uschold *et al.*, 1996), es decir, un proceso de desarrollo de arriba hacia abajo. Este enfoque se caracteriza por la identificación, en primer lugar, de los conceptos más abstractos y, en una fase posterior, se especializa los conceptos más específicos. La utilización de este enfoque conlleva como principal resultado un mejor control del nivel de detalle. Sin embargo, iniciar el proceso desde lo más general puede suponer tener que elegir y, más tarde, imponer arbitrariamente y posiblemente innecesariamente categorías de alto nivel. A continuación se detallan algunas características del proceso de modelado de la ontología relativas al enfoque y a los criterios de estructuración adoptados.

6.3.4.1. Enfoque

La conceptualización del dominio hidrográfico propuesta por *hydrOntology* se encuadra dentro del enfoque Global-as-View, descrito en detalle en el Capítulo 3 (sección 3.2.4), consecuencia de que el modelo que sostiene esta ontología está expresado conforme a los términos que aparecen en las diferentes fuentes de información tomadas como referencia.

La selección de dicho enfoque (Global-as-View) es resultante de la pretensión de este trabajo de originar una ontología de dominio que abarque, desde la óptica de la IG y su plasmación cartográfica, en gran parte, a todos aquellos fenómenos hidrográficos (tanto universalmente conocidos como aquellos dependientes de una determinada área geográfica dentro de nuestro territorio) que gestionan las fuentes que contienen este tipo de información. Con esto, se obtendrá una ontología novedosa e innovadora en cuanto a las ontologías relacionadas con la IG, por dos motivos fundamentales. Primero, por la ausencia de una ontología que recopile a todos los fenómenos hidrográficos representables cartográficamente, aunque son numerosas las referidas a la temática de hidrología y, segundo, porque entre las ontologías relacionadas con *hydrOntology* no hay ninguna que contemple tantas y variadas fuentes de información sobre el dominio.

Este enfoque es el óptimo para la labor de conceptualización de los diferentes dominios geográficos desde una perspectiva global, ya que aporta la visión general y, por tanto, la abstracción necesaria para la modelización de un dominio de conocimiento tan complejo como es el geográfico. Este hecho descarta el enfoque Local-as-View, debido a que el modelo general va a estar condicionado por el número de fuentes de trabajo. Por tanto, la consideración de diferentes modelos o esquemas dependientes de una determinada fuente de datos impide alcanzar la abstracción necesaria para la obtención de un modelo común, compartido y global de cualquier dominio.

6.3.4.2. Criterios de estructuración

En el proceso de desarrollo de *hydrOntology* se adoptaron una serie de criterios para la estructuración del conocimiento hidrográfico recogido en las múltiples fuentes de información controladas por esta ontología. Así, la organización de los fenómenos hidrográficos contenidos en esta ontología se rige principalmente por cuatro criterios (Vilches Blázquez *et al*., 2007b):

1. **Directiva Europea Marco del Agua**. Esta Directiva establece un marco comunitario de actuación en el ámbito de la política de aguas (European Parliament and the Council of the European Union, 2000). Concretamente, en su artículo 2 de esta Directiva, se recoge un listado de fenómenos hidrográficos y sus correspondientes definiciones en lenguaje natural. Este conjunto de fenómenos recoge una clasificación implícita. De esta manera, este listado contribuye a modelar el componente abstracto del modelo de *hydrOntology*. Las definiciones de los fenómenos hidrográficos recogidas en este documento son propuestas por el Parlamento Europeo y el Consejo de la Unión Europea. Este hecho implica la necesidad de que cualquier modelado sobre el dominio recoja este mínimo listado de fenómenos y sus definiciones.

2. El trabajo resultante del **proyecto SDIGER**[12] (Latre *et al.*, 2005) se toma como información de referencia consecuencia de la interrelación con el dominio de trabajo. Este proyecto fue seleccionado por Eurostat[13] como proyecto piloto de aplicación de INSPIRE (Commission of the European Communities, 2007). Asimismo, junto a los documentos de referencia del proyecto (Institute Géographique National France International *et al.*, 2005a) en el proceso de modelado de *hydrOntology* tiene una importancia estratégica los diversos modelos UML (Institute Géographique National France International *et al*., 2005b) resultantes del mencionado proyecto y del *Swedish Standard – Surface water systems* (Swedish Standard, 2006). El proceso de modelado de *hydrOntology* adopta en gran medida estos modelos y, además, incluye varios aportes a los modelos realizados en el proyecto SDIGER producto del consenso con el grupo de trabajo de la Universidad de Zaragoza que llevó a cabo el desarrollo del proyecto SDIGER. Estos cambios actualizaron los modelos propuestos en dicho proyecto.

3. La **semántica vinculada** a cada tipo de **fenómeno hidrográfico**, es decir, la definición en lenguaje natural asociada a cada fenómeno tiene un papel fundamental en proceso de modelado de los diferentes componentes de esta ontología. Este factor hace que el modelado de fenómenos hidrográficos recogidos en *hydrOntology* sea consecuente con el significado asociado a cada fenómeno.

4. **Análisis y asesoramiento** por parte del grupo de **expertos** pertenecientes a la sección de **toponimia** del IGN.

5. Finalmente, en el proceso de modelado de esta ontología se incluyen términos generales utilizados por los expertos del dominio. Estos términos se adoptan en la ontología para referirse a los conceptos de la realidad del dominio tratado (Smith *et al.*, 2010). De esta manera, **persiste** cierta **herencia** proveniente de las diferentes fuentes de información consideradas. Este hecho permite, por un lado, un posible establecimiento de *mappings*

[12] *Proyecto SDIGER*. Este proyecto desarrolla una Infraestructura de Datos Espaciales para dar acceso a los recursos de información geográfica referidos a la Directiva Marco del Agua (European Parliament and the Council of the European Union, 2000) en un escenario limítrofe e inter-administrativo que involucra a dos países (Francia y España) así como las dos principales cuencas hidrográficas a ambos lados de la frontera, la región de la cuenca del Adour-Garonne, gestionada por la Agencia Hidrográfica de la cuenca del río Adour-Garonne (L'Agence de l'Eau Adour-Garonne) y la región de la cuenca del río Ebro, gestionada por la Confederación Hidrográfica del Ebro.

[13] *Eurostat.* http://www.ec.europa.eu/eurostat/

entre la ontología y las diversas fuentes de conocimiento y, por otro, mantener la coherencia con el modelado del dominio realizado por diferentes expertos.

6.3.5. Características de hydrOntology

En esta sección se muestra una visión de alto nivel de *hydrOntology* con objeto de poder mostrar un esbozo de los diferentes niveles reflejados en esta ontología. De forma global, como puede apreciarse en la Figura 6.8, esta ontología está dividida en dos niveles: nivel superior y nivel inferior. El nivel superior contiene los fenómenos hidrográficos más abstractos de la ontología, mientras y el nivel inferior recoge un conjunto de fenómenos universalmente conocidos. Con respecto a los conceptos, el nivel superior contiene, por ejemplo, el concepto "Fenómeno hidrográfico", bajo el cual se ubican conceptos más especializados como "Aguas continentales" o "Aguas marinas". Entre estos conceptos existe un diferente grado de especialización, consecuencia de que esta ontología se centra en los fenómenos vinculados a las "Aguas continentales". Este concepto, conforme a la Directiva Marco del Agua (European Parliament and the Council of the European Union, 2000), es dividido en "Aguas superficiales" (cuyas clases hijas son: "Aguas de transición", "Aguas quietas", "Aguas corrientes", "Surgencias") y "Aguas subterráneas" (Vilches-Blázquez *et al.*, 2007a). Para cada una de estas clases se han identificado conceptos en el nivel inferior, donde se provee de un conjunto detallado de fenómenos hidrográficos.

En el amplio abanico de fenómenos considerados en *hydrOntology* se encuentran diversos conceptos que pueden ser utilizados para relacionar esta ontología con otros dominios de conocimiento, tales como el marco legal ("Aguas territoriales", "Zona contigua", "Alta mar", etc.), el dominio geológico o hidrogeológico ("Corriente Subterránea", "Acuífero", etc.) o la ingeniería civil urbana (tubería, canal, embalse, etc.) como queda patente en (Vilches Blázquez *et al.*, 2007b). Esta inclusión de conceptos relacionados con otros dominios enriquece la ontología desarrollada, consecuencia de la facilidad ante un posible crecimiento gradual en conocimiento y valor añadido.

Además, en el desarrollo de *hydrOntology* se han tenido en cuenta algunos conceptos relacionados con fenómenos que son pertenecientes de manera exclusiva a una determinada región o espacio geográfico. Entre todos estos fenómenos merecen destacarse "ibón", "lavajo", "chortal", "bodón" y "lucio". Estos conceptos son designados por su nombre local, aunque son sinónimos del concepto "Charca (Laguna pequeña y poco profunda)".

El modelado de esta ontología recoge los cuatro tipos de relaciones taxonómicas definidos en el Frame Ontology (Chaudhri *et al*., 1998) y OKBC Ontology (Ganter *et al.*, 1997): Subclase de (*Subclass-Of*), Descomposición disjunta (*Disjoint-Decomposition*), Descomposición exhaustiva (*Exhaustive-Decomposition*) y Partición (*Partition*).

Asimismo, esta ontología es desarrollada conforme a los principios de diseño de ontologías propuestos en (Gruber, 1995; Arpírez *et al.*, 1998). Algunas de las características más importantes corresponden a que los nombres de los conceptos (clases) son suficientemente explicativos y están correctamente escritos. Además, cada clase intenta agrupar sólo un único concepto y, por tanto, las clases entre paréntesis y/o con enlaces ("y", "o") son evitados. En cuanto a la convención de nombres, cada clase es escrita con letras mayúsculas al comienzo de cada concepto, mientras relaciones y atributos son escritos con letras minúsculas (Vilches-Blázquez *et al.*, 2009).

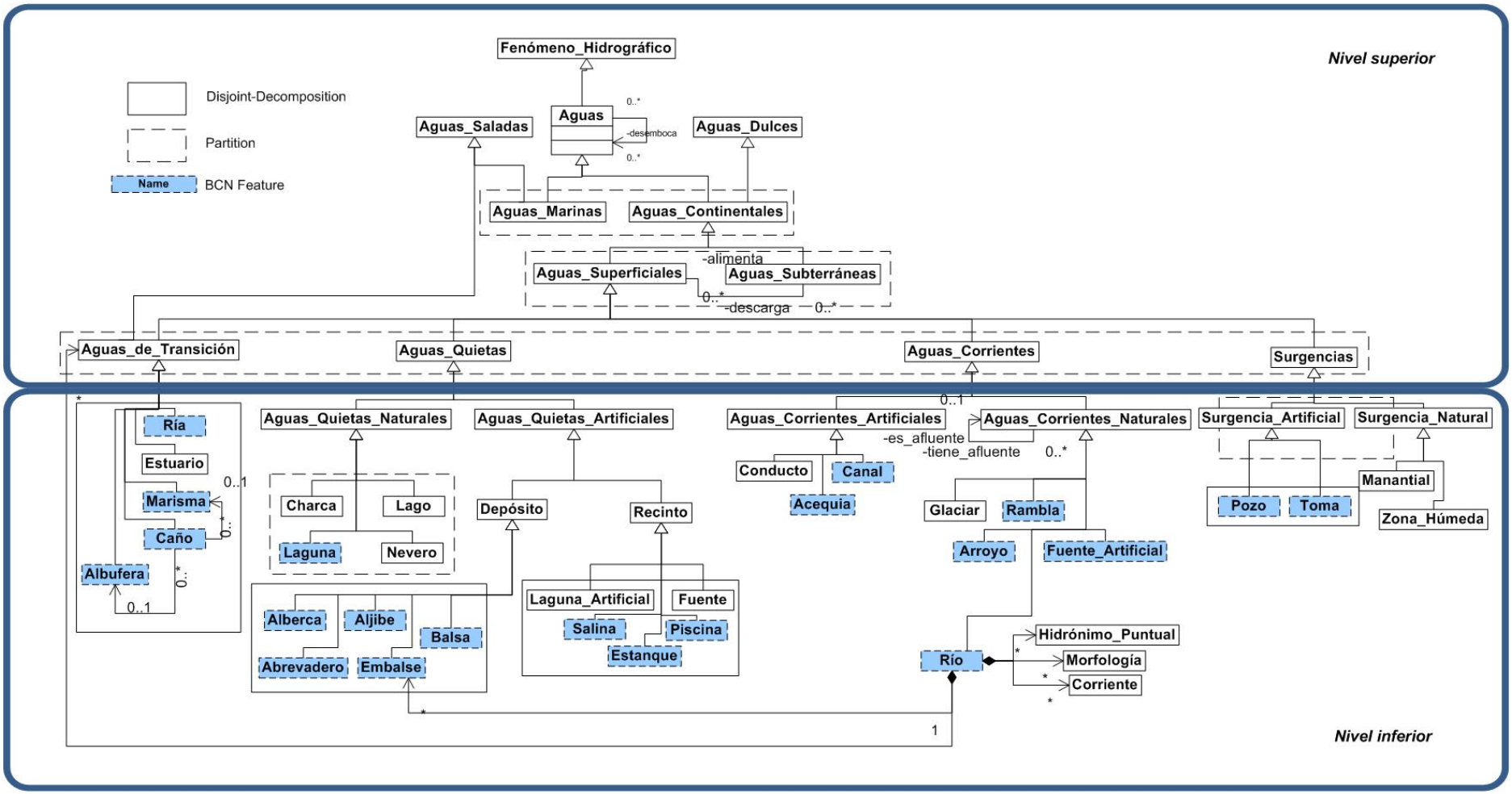


Figura 6.8. Visión general del modelo de *hydrOntology*

Finalmente, con respecto a los datos estadísticos (métrica), esta ontología posee 154 clases, 47 relaciones, 64 atributos y 259 axiomas.

6.4. Red de ontologías en torno a *hydrOntology*

En esta sección se realiza una descripción detallada de las principales características del proceso de desarrollo de la red de ontologías en torno a *hydrOntology*.

6.4.1. Especificación de requisitos

En la descripción del propósito, alcance y requisitos a cumplir (reflejados a través de las preguntas de competencia (Gruninger *et al.*, 1994)) de la red de ontologías entorno a *hydrOntology* se utiliza la plantilla de especificación de requisitos de ontologías (*Ontology Requirements Specification Document*) descrito en (Suárez-Figueroa *et al*., 2008). En la Tabla 6.3 se recogen los campos de la mencionada plantilla y una descripción de la especificación de requisitos de la red de ontologías.

Tabla 6.3. Documento de especificación de requisitos de la ontología (*Ontology Requirements Specification Document*) descrita en (Suárez-Figueroa *et al*., 2008)

Especificación de requisitos de la ontología	
1	*Propósito*
	El propósito de la red de ontologías en torno a *hydrOntology* es servir como marco de armonización para los diferentes productores de información geoespacial a través del recubrimiento de la mayoría de fenómenos asociados al dominio hidrográfico representables cartográficamente. Asimismo, esta red debe servir como marco de armonización entre los diferentes productores de información geoespacial en el entorno nacional e internacional. Finalmente, esta red de ontologías debe proveer información en castellano, catalán, inglés y francés consecuencia del contexto globalizador de la información geográfica.
2	Alcance
	La red de ontologías recogerá información relativa a los fenómenos del dominio hidrográfico y asimismo se añadirá conocimiento proveniente de diversas características relacionadas con el dominio mencionado, la representación cartográfica, relaciones espaciales, etc.
3	*Nivel de formalidad*
	La red de ontologías desarrollada debe ser implementada conforme a la especificación de OWL (Dean *et al.*, 2003). Este lenguaje es una recomendación técnica del W3C Consortium para la Web Semántica. Asimismo, de este lenguaje se han derivado tres sub-lenguajes con expresividad incremental (OWL Lite, OWL DL y OWL Full) (McGuinness *et al.,* 2004)[1]. La formalización de esta red de ontologías es realizada conforme a OWL DL.
4	*Usuarios previstos*
	Productores cartográficos con objeto de armonizar sus fuentes de información y usuarios interesados en la información hidrográfica.
5	*Usos previstos*
	Esta red de ontologías tiene como objetivo recuperar información sobre los diferentes fenómenos que componen el dominio hidrográfico y cuya información se encuentra en las diferentes bases de datos de los productores de esta información.
6	*Grupos de preguntas de competencia*
	Las preguntas de competencias se recogen en la sección 6.4.2 de este trabajo.

6.4.2. Preguntas de competencia

Las preguntas de competencia que el desarrollo de esta red de ontologías tiene que conseguir responder son:

[1] Una descripción detallada de estos lenguajes puede ser encontrada en http://www.w3.org/TR/owl-features/

Tabla 6.4. Preguntas de competencia de la red de ontologías

1.	¿Qué tipo de aguas saladas aparecen en la Comunidad Autónoma "x"? / ¿Qué tipo de aguas aparecen en la provincia "x"?
2.	¿Qué aguas quietas / corrientes se pueden encontrar en el municipio "x"?
3.	¿Qué afluentes alimentan a los ríos de la cuenca "x" / … a los ríos que desembocan en?
4.	¿Qué afluentes tiene el río "x" y dónde se localizan?
5.	¿Qué objetos hay en la cuenca hidrográfica "x"?
6.	¿Cuáles son los elementos de los que consta el río "x"?
7.	¿Embalses que forman parte del río "x"?
8.	¿Qué fenómenos hidrográficos tiene / atraviesan la provincia "x"?
9.	¿Los ríos "x", "y" y "z" de qué río son afluentes y a qué acuíferos alimentan?
10.	¿Qué afluentes desembocan en el río "x", a qué acuíferos alimenta y en qué embalse se almacenan?
11.	¿Qué ríos son conducidos por una tubería?
12.	¿Qué lagunas son originadas por un glaciar?
13.	¿Qué aguas alimentan a la albufera "x" y qué caños la comunican con el mar?
14.	¿En qué idiomas aparecen la información del fenómeno "x"? / Idiomas de los fenómenos hidrográficos de la Cuenca "x".
15.	¿Qué río alimenta a la marisma "x" y dime si origina algún lucio y cuáles son?
16.	¿Cuál es la localización (latitud/longitud) del embalse "x"?
17.	¿Qué embalse almacena las aguas del río "x" y qué arroyos alimentan a dicho río?
18.	¿Qué esteros son inundados por una laguna / aguas costeras?
19.	¿Qué tipo de representación geométrica tiene la instancia "x" que corresponde al fenómeno "y"?
20.	¿Qué ríos tienen un sumidero que convierten su caudal en una corriente subterránea?

6.4.3. Fuentes de conocimiento

En el desarrollo de la red de ontologías en torno a *hydrOntology* se han utilizado diversos recursos ontológicos como fuentes de adquisición de conocimiento. La recopilación de estos recursos se centró, fundamentalmente, en los principales buscadores semánticos (Swoogle y Watson) y no semánticos existentes (Google). Los criterios de búsqueda estuvieron centrados en conseguir ontologías pertenecientes o vinculadas, con mayor o menor grado de especialización, en los dominios geográfico, hidrológico e hidrográfico. Además, el proceso de búsqueda se extendió a otras fuentes relacionadas con los mencionados dominios (por ejemplo, NASA[2] u Ordnance Survey).

Los recursos ontológicos candidatos a la reutilización durante la construcción de la red de ontologías se muestran en la Tabla 6.5. En esta tabla se recogen diferentes recursos ontológicos relacionados con el dominio hidrográfico y la información geoespacial de manera genérica.

El proceso de búsqueda y selección de recursos ontológicos origina la posibilidad de combinar varios escenarios de construcción de redes de ontologías. Con objeto de decidir los escenarios a seguir en el proceso de creación de la red resulta necesario determinar los recursos a utilizar y las posibles adaptaciones requeridas para conformar la mencionada red. El proceso de selección de recursos ontológicos, procedentes del listado controlado (Tabla 6.5), considera los siguientes criterios:

- Relación con el dominio de conocimiento (dominio hidrográfico) y/o con el área de información geográfica.
- Visión general de la cantidad (en términos cualitativos) de conocimiento reutilizable (Poveda, 2009):

[2] *NASA*. http://sweet.jpl.nasa.gov/ontology/

  - Cantidad de clases, relaciones, atributos y axiomas candidatos a permanecer tras la reutilización de la ontología.
  - Cantidad de clases, relaciones, atributos y axiomas candidatos a ser modificados tras la reutilización de la ontología.
  - Razones por los que se debería eliminar o modificar clases, relaciones, atributos y axiomas.
  - Tratamiento (eliminarlas, renombrarlas, redefinirlas, etc.) a realizar sobre dichas clases, relaciones, atributos y axiomas.
  - Cantidad de clases, relaciones, atributos y axiomas que se tendrían que añadir en la ontología.
- Consistencia de la ontología (Poveda, 2009). Este criterio destaca la importancia de la consistencia de las ontologías disponibles por encima de la cantidad o completitud de las mismas. Si algún caso con estas características aparece se procedería a valorar el esfuerzo de detección de que genera la inconsistencia.

La aplicación de esta serie de criterios al proceso de selección de recursos ontológicos procedentes del listado controlado delimita el conjunto de recursos que van a formar parte de la red de ontologías. Las ontologías seleccionadas aparecen delimitadas con línea discontinua en la Tabla 6.5.

La consideración de estos recursos ontológicos conlleva la selección de los siguientes escenarios propuestos en la metodología NeOn (Suárez-Figueroa *et al.*, 2007):

- Desarrollo de redes de ontologías mediante **reutilización** de recursos ontológicos (Escenario 3), es decir, este proceso de desarrollo se lleva a cabo mediante el uso de parte o totalidad de ontologías ya existentes para el desarrollo de una nueva.
- Desarrollo de redes de ontologías mediante **reutilización y reingeniería** de recursos ontológicos (Escenario 4). La reestructuración de recursos ontológicos se refiere al proceso de corrección y reorganización del conocimiento contenido en un modelo conceptual inicial, así como a la detección de conocimiento no representado (Suárez-Figueroa *et al.*, 2007). Esta actividad se divide en distintas subactividades como muestra la Figura 6.9. Estas subactividades corresponden a modularización, poda y enriquecimiento. La actividad de enriquecimiento de una ontología se divide a su vez en: extensión y especialización.
- Desarrollo de redes de ontologías mediante **localización** de recursos ontológicos (Escenario 9). El proceso de localización de ontologías consiste en la adaptación de una ontología a una comunidad idiomática y cultural concreta (Espinoza *et al.*, 2008).

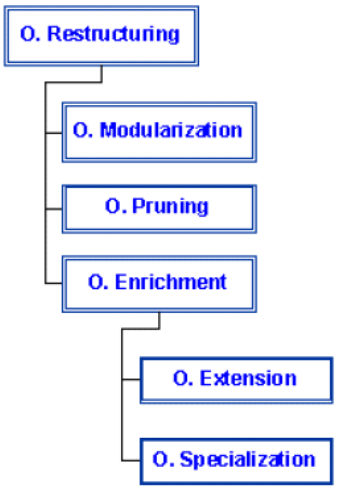


Figura 6.9. Actividades de reestructuración ontológica (Suárez-Figueroa *et al.*, 2007)

Tabla 6.5. Recursos ontológicos candidatos a reutilización en la red en torno a *hydrOntology*

Criterio de búsqueda	Recurso ontológico	Clases	Relaciones	Atributos	Expresividad	Formato
geographic	iso19112	2	0	1	ALN(D)	OWL
	CommonSenseMapping	127	274	2	SHIN(D)	OWL
	D_GeoE	35	7	0	ALE	OWL
	D_Loc	60	60	38	ALCIN(D)	OWL
	swtopics4	3	6	2	ALCO(D)	OWL
	ISO_geographic_service	85	0	0	AL	OWL
	MeasurementBase	119	55	42	ALCIN(D)	OWL
	testbed_v1_1	30	7	52	AL(D)	OWL
	pml-provenance	32	20	38	ALCHIN(D)	OWL
	MGEDOntology	234	84	44	ALEOF(D)	OWL
	Ipronto	113	54	0	ALCHI	OWL
	NTNames	49	26	9	SHIN(D)	OWL
	modifiedO&M_ontology	50	46	1	ALCOIN(D)	OWL
	Geofile	78	6	13	ALUOI(D)	OWL
	TimeAndSpace	186	100	46	ALCIN(D)	OWL
	ogc-gml	173	80	41	ALCON(D)	OWL
hydrographic	mistr-ont-3.2	19	19	74	ALUN(D)	OWL
	Space	171	0	0	AL	OWL
	Government	84	78	3	ALH(D)	OWL
	classes	676	0	0	AL	OWL
	JC3IEDM-CodeClasses	273	0	0	ALUO	OWL
	space0_1	188	0	0	AL	OWL
	FTT	1262	0	0	AL	OWL
	GetRDFByID	187	2	0	AL	OWL
	hydrocv	51	0	0	AL	OWL
	kimo	332	110	0	ALH	RDFS
	space_gis					DAML
	FreeToGovCyc	11748	3733	1	ALCHF(D)	OWL
	Index	187	2	0	AL	RDF
Hydrology	iso-19103-cuahsi	28	2	3	ALEON(D)	OWL
	Modifiedmaterial_thing	1580	102	19	ALCHOIF(D)	OWL
	buggy-sweet-jpl	1537	102	19	ALCHOF(D)	OWL
	AKTiveSAOntology	2343	183	20	SHOIN(D)	OWL
	ontosem	8071	605	0	ALUHI	OWL
NASA	human_activities_partition9	28	0	0	AL	OWL
	sweet-full	1938	84	1	ALCHOIF(D)	OWL
	human_activities	431	13	0	AL	OWL
	human_activities_partition22	10	2	0	AL	OWL
	hydroBodyOfWater	2967	180	28	SHOIN(D)	OWL
	hydroGroundwater	1688	152	28	SHOIN(D)	OWL
	hydroSurface	1534	155	28	SHOIN(D)	OWL
	ocean	2586	181	28	SHOIN(D)	OWL
	units	12	3	5	ALUOF(D)	OWL
	geogCoast	3030	184	28	SHOIN(D)	OWL
	geogFluvial	2174	176	28	SHOIN(D)	
Ordnance Survey	AdministrativeGeography	10	2	9	ALH(D)	RDF
	ForeignLanguageRelations	0	1	1	AL(D)	OWL
	MereologicalRelations	0	14	0	ALHIF+	OWL
	NetworkRelations	1	6	0	ALHI+	OWL
	SpatialRelations	2	28	0	ALHIF+	OWL
	Hydrology0_1	251	107	11	SHOIQ(D)	OWL
Otros	GeoNames	11	15	7	ALHO+(D)	OWL
	ISO19115	146	160	159	ALUOIN(D)	OWL
	hydrology	21	28	5	ALOIF+(D)	OWL
	WGS84 vocabulary	5	1	7	ALH(D)	OWL
	PhenomenOntology	117	0	155	AL(D)	OWL
	FAO_geopolitical	12	6	79	SHIN(D)	OWL

Estos escenarios se combinan con el escenario 1 (Desarrollo de redes de ontologías desde la especificación hasta la implementación), consecuencia de que este escenario se plantea como base en el proceso de construcción de ontologías y redes de ontologías, ya que contiene el núcleo de actividades que deben ser llevadas a cabo en cualquier desarrollo ontológico.

6.4.4. Escenarios

Esta sección recoge los detalles del proceso de desarrollo de la red de ontologías en torno a los diversos escenarios seleccionados. Los procesos de reutilización (escenario 3) y reingeniería (escenario 4) de recursos ontológicos fueron realizados de forma paralela, estos procesos tienen como finalidad acelerar el desarrollo de ontologías. En un proceso posterior se llevó a cabo la localización de ontologías (escenario 9).

Tabla 6.6. Recursos ontológicos reutilizados en la red de ontologías en torno a *hydrOntology*

Ontología	Dominio	Conformidad
SpatialRelations	Relaciones espaciales	RCC8
Mereology	Mereología	
NetworkRelations	Relaciones de redes	
ForeignLanguageRelations	Lingüístico	
Hydrocv		
languagecode	Idiomas	ISO-639-1; ISO-639-2
ISO19115	Metadatos	ISO 19115
OGC-gml	Modelado, transporte y almacenamiento de IG	GML
units	Unidades de medida	SWEET
WGS84 vocabulary	Representación latitud, longitud y altitud	Sistema de coordenadas WGS84

6.4.4.1. Reutilización de recursos ontológicos

La reutilización de recursos ontológicos concierne al uso de parte o totalidad de ontologías existentes para el desarrollo de una nueva ontología. La reutilización directa de recursos ontológicos sin llevar a cabo actividades de reingeniería, mezcla o alineación, es el resultado de la aplicación del escenario 3 propuesto en la metodología NeOn. Los recursos ontológicos reutilizados en el desarrollo de la red de ontologías en torno a *hydrOntology* son listados en la Tabla 6.6.

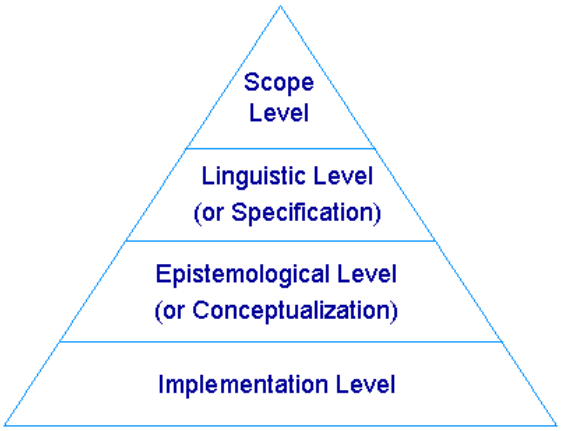


Figura 6.10. Niveles de abstracción para la actividad de Reingeniería Ontológica (Suárez-Figueroa *et al*., 2007)

6.4.4.2. Reutilización y reingeniería de recursos ontológicos

En (Suárez-Figueroa *et al.*, 2007) se define reingeniería de ontologías como el proceso de recuperar y transformar un modelo conceptual existente e implementado en una ontología, en un nuevo modelo, más correcto y completo, que es reimplementado. Asimismo, se proponen niveles de abstracción que subyacen a la actividad de reingeniería ontológica, reflejados en la Figura 6.10.

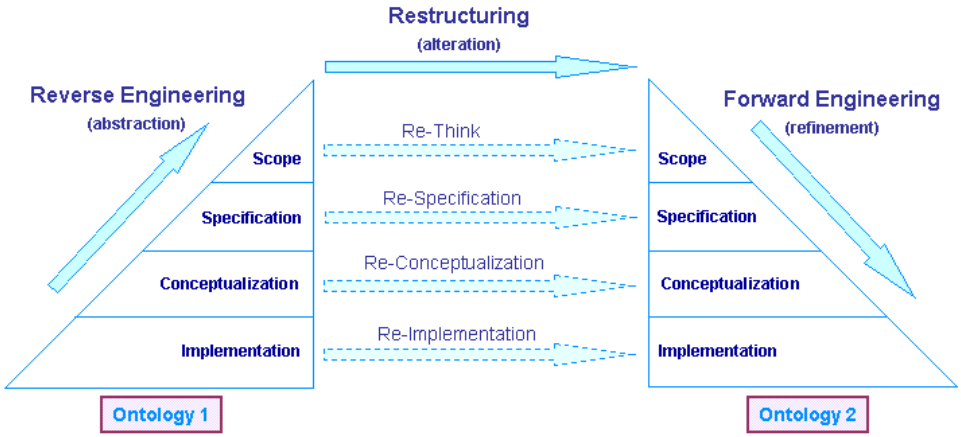


Figura 6.11. Niveles de abstracción para la reingeniería ontológica (Suárez-Figueroa *et al*., 2007)

Considerando estos niveles de abstracción, recogidos en la Figura 6.10, (Suárez-Figueroa *et al*., 2007) propone el modelo de reingeniería ontológica mostrado en la Figura 6.11. Este modelo expone que, con frecuencia, la actividad de reingeniería ontológica parte del código de una ontología y produce el código de una ontología nueva. Este modelo también muestra los diferentes caminos para la transformación de una ontología en otra nueva. Asimismo, se distinguen los siguientes tipos de cambios: reimplementación, reconceptualización, reespecificación y replanteamiento.

Tabla 6.7. Recursos ontológicos sometidos a modificaciones en el desarrollo de la red de ontologías

Ontología	Dominio	Conformidad	Modificación
PhenomenOntology	Topográfico	BTN25	1, 4, 6, 7
Hydrology0.1	Hidrología	Master Map real-world object catalogue	1, 4
FTT (Feature Type Thesaurus)	Nombres geográficos	Alexandria Digital Library	1, 4 ,6
AKTiveSAOntology			1, 4
GetRDFByID	Espacio físico		1, 4, 6
hydrology	Hidrología		1 ,6
hydroBodyOfWater		NASA (SWEET)	1, 4
hydroGroundwater		NASA (SWEET)	1, 4
hydrOntology	Hidrografía	*hydrOntology*	1, 2, 3, 6, 7
hydroSurface		NASA (SWEET)	1, 4
ocean		NASA (SWEET)	1, 4
geogCoast		NASA (SWEET)	1, 4
geogFluvial		NASA (SWEET)	1, 4
FAO_geopolitical	Geopolítico	FAO	2, 5, 7

El proceso de reutilización de recursos ontológicos (concretado en el uso de parte o totalidad de ontologías ya existentes para el desarrollo de una nueva) combinado con actividades de

reingeniería (Escenario 4) conllevan diferentes modificaciones en los recursos ontológicos listados en la Tabla 6.7.

A continuación se recogen algunos ejemplos representativos de los diferentes tipos de modificaciones realizadas sobre los recursos ontológicos que conforman la red.

1. ***Creación y/o reestructuración de jerarquías***. Esta modificación, clasificada dentro del tipo reconceptualización dentro del proceso de reingeniería, consiste en crear una jerarquía a partir de un determinado concepto creando las subclases a partir del mismo y profundizando en la clasificación conforme a la granularidad pretendida. Puede que no sea necesario crear las clases y sea suficiente ordenarlas creando la clasificación. Un ejemplo de reestructuración del conocimiento de PhenomeOntology se muestra en la Tabla 6.8.

Tabla 6.8. Ejemplo de reestructuración de jerarquías

Situación inicial	Situación final
03 ALMACENAMIENTO-DE-AGUA BOCA-DE-HIDROGRAFIA CASCADA CAUCE-ARTIFICIAL CIERRE-CONVENCIONAL DEPOSITO-DE-AGUA EMBALSE HITO-KILOMETRICO-DE-CANAL HUMEDAL ISLA LAGUNA LINEA-DE-COSTA RAMBLA RIA	Aguas_Quietas Aguas_Quietas_Artificiales Aguas_Quietas_Naturales Charca Lago Laguna Albufera Laguna_Salada Masa_de_Hielo

2. ***Compleción de conocimiento representado***. Esta modificación, clasificada dentro del tipo reconceptualización, consiste en el enriquecimiento de ontologías mediante la ampliación del conocimiento representado añadiendo elementos como clases, relaciones, etc. En la Tabla 6.9 se recoge un ejemplo de compleción de conocimiento entre la ontología hydroGroundwater (situación inicial) y su combinación con *hydrOntology* (situación final).

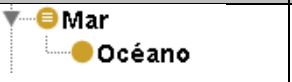
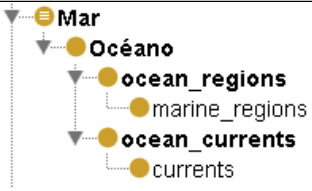
Tabla 6.9. Ejemplo de compleción de conocimiento

Situación inicial	Situación final
Groundwater CascadingWate ConnateWate PerchedGroundwat UndergroundWat	Aguas_Subterráneas ConnateWater CascadingWater PerchedGroundwater UndergroundWater Acuitardo Acuífero Corriente_Subterránea Lago_Subterráneo

3. ***Restablecimiento de granularidad***. Esta modificación, clasificada dentro del tipo reespecificación, consiste en profundizar o podar una jerarquía con el fin de ampliar o

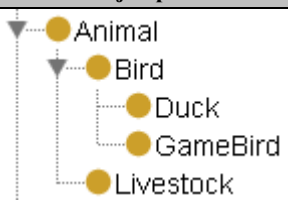
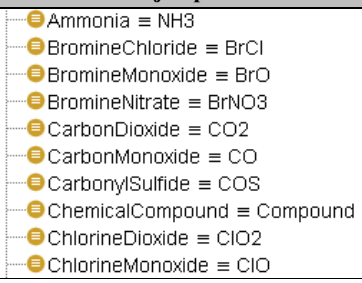
disminuir la granularidad respectivamente. En la Tabla 6.10 se recoge un ejemplo de aumento del grano de una sección de la red de ontologías en torno a *hydrOntology*.

Tabla 6.10. Ejemplo de profundización en el grano de la ontología

Situación inicial	Situación final
	

4. ***Restricción del dominio***. Esta modificación, clasificada dentro del tipo reespecificación, consiste en eliminar el conocimiento que no es propio de un dominio y reubicarlo en el módulo considerado como parte del dominio. Un ejemplo de eliminación de conocimiento no vinculado directamente al dominio hidrográfico se muestra en la Tabla 6.11. Los ejemplos recogidos pertenecen a las ontologías *Hydrology0*.1 (*Ordnance survey*) y *hydroBodyOfWater* (NASA).

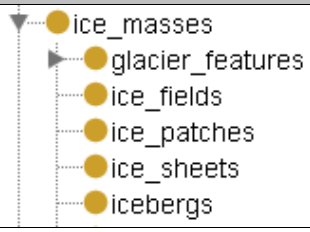
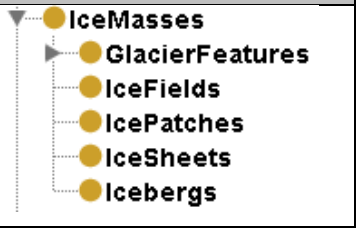
Tabla 6.11. Ejemplos de restricción del dominio

Ejemplo	Ejemplo
	

5. ***Redefinición de clases y axiomas***. Esta modificación, clasificada dentro del tipo reconceptualización, consiste en transformar los axiomas de la ontología. Estos axiomas pueden describir una condición necesaria o una condición necesaria y suficiente.

6. ***Unificación de nombrado***. Esta modificación, clasificada dentro del tipo reimplementación, consiste en adaptar los identificadores de los elementos de la ontología al estilo de nombrado CamelCase[3]. La Tabla 6.12 presenta un ejemplo de unificación de nombrado.

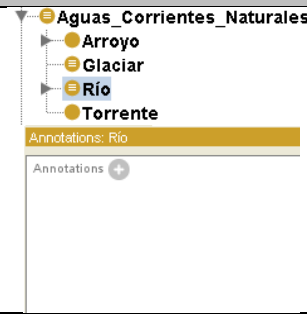
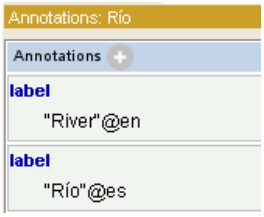
Tabla 6.12. Ejemplo de unificación de nombrado

Situación inicial	Situación final
	

[3] *CamelCase*. Es un estilo de escritura que se aplica a frases o palabras compuestas. Existen dos tipos de CamelCase: *UpperCamelCase*, cuando la primera letra de cada una de las palabras es mayúscula (EjemploDeUpperCamelCase) y *lowerCamelCase*, igual que la anterior con la excepción de que la primera letra es minúscula (ejemploDeLowerCamelCase).

7. ***Unificación de idioma***. Esta modificación, clasificada dentro del tipo reconceptualización, consiste en traducir a un determinado idioma los términos que aparezcan en el campo *rdf:label*. En el desarrollo de esta red de ontologías se establece el idioma inglés como referencia idiomática común del conjunto de ontologías. En la Tabla 6.13 se muestra un ejemplo de unificación de idioma sobre el concepto "Río".

Tabla 6.13. Ejemplo de unificación de idioma

Situación inicial	Situación final
	

6.4.5. Características de la red de ontologías

La combinación de escenarios descrita con anterioridad, junto con el seguimiento de las pautas metodológicas para la construcción de redes de ontologías propuestas en (Poveda, 2009), genera como resultado la red de ontologías reflejada en la Figura 6.12. Entre los recursos que componen esta red aparecen diversas ontologías incluidas entre las siete categorías de ontologías geoespaciales del *W3C Geospatial Incubator Group* (Lieberman *et al.*, 2007), cuya descripción se recoge en el Capítulo 3 (sección 3.1.3) de este trabajo. Además, en el modelo de alto nivel (ver Figura 6.12) se muestran los diferentes recursos ontológicos que componen la red, sus relaciones globales y diversas agrupaciones temáticas de los recursos (Relaciones espaciales, Mereología, División administrativa, Metadatos, Unidades de medida, Multilingüismo y Representación espacial). Asimismo, esta figura también recoge el conjunto de ontologías que pasan a formar parte de *hydrOntology*.

En la Tabla 6.14, a modo de resumen, se recogen las métricas de las diferentes ontologías que componen la red resultante. Esta red está constituida por 4143 clases, 776 propiedades y 447 atributos. Además, en la mencionada tabla, se recoge información sobre el tipo de ontología fundamental según la clasificación recogida por el *W3C Geospatial Incubator Group*.

6.4.6. Documentación y localización de recursos ontológicos

Como se comentó en la descripción general del proceso de desarrollo (sección 6.1) el proceso de documentación y localización se lleva a cabo con posterioridad al proceso de modelado, formalización e implementación de la ontología individual (*hydrOntology*) y la red de ontología.

El proceso de **documentación** se centra en la ontología individual, adoptándose en el caso de los recursos ontológicos que constituyen la red la documentación que acompaña a cada una de esas ontologías reutilizadas. Este proceso está caracterizado por la recogida de la siguiente información para cada componente de *hydrOntology* (clases, relaciones y atributos):

- **Definición** (*comment*): Se recopila una descripción en lenguaje natural del componente. Entre las diferentes fuentes consideradas en el proceso de establecimiento de definiciones merecen especial consideración WordNet[4], Diccionario de la Real Academia Española de la Lengua, Glosario Hidrológico Internacional de la UNESCO,

[4] http://wordnet.princeton.edu/

Diccionario de usos del español (Moliner, 1984), etc. y diversos diccionarios geográficos, como por ejemplo (Fontanillo, 1986; Vázquez Maure *et al.*, 1987).

- **Fuente** (*source*): Se recoge información perteneciente a la procedencia de la descripción en lenguaje natural.
- **Etiqueta** (*label*): Se recopila información perteneciente al nombre alternativo o variantes del componente principal. Asimismo, se recoge el nombre del componente principal sin caracteres adicionales.
- **Procedencia** (*provenance*): Se recoge la información correspondiente al origen del componente principal y alternativos (sinónimos).

Tabla 6.14. Listado de recursos y métricas de la red de ontologías

	Clases	**Relaciones** ***(object property)***	**Atributos** ***(data property)***	**W3C Geospatial**
AKTiveSAOntology	706	110	12	Tipos de fenómenos
ForeignLanguageRelations	0	1	1	-
FTT	217	0	-	Topónimos
geogCoast	42	5	28	Tipos de fenómenos
GetRDFByID	117	2	-	Tipos de fenómenos
Hydrocv	51	-	-	Tipos de fenómenos
Hydrology0.1	229	69	4	Tipos de fenómenos
ISO19115	146	160	159	Metadatos geoespaciales
languagecode	5	4	7	-
Mereology	0	14	-	-
NetworkRelations	1	6	-	Relaciones geo(espaciales)
OGC-gml	169	80	41	Fenómenos geoespaciales
WGS84 vocabulary	5	1	7	Sistema de coordenadas
PhenomenOntology	28	-	-	Tipos de fenómenos
SpatialRelations	2	28	-	Relaciones geo(espaciales)
hydroGroundwater	74	5	-	Tipos de fenómenos
hydroBodyOfWater	2026	195	36	Tipos de fenómenos
hydroSurface	39	3	-	Tipos de fenómenos
ocean	56	8	-	Tipos de fenómenos
hydrOntology	154	47	64	Tipos de fenómenos
hydrology	18	27	4	Tipos de fenómenos
geogFluvial	29	-	-	Tipos de fenómenos
units	12	3	5	-
FAO_geopolitical	17	8	79	Tipos de fenómenos
Total	**4143**	**776**	**447**	

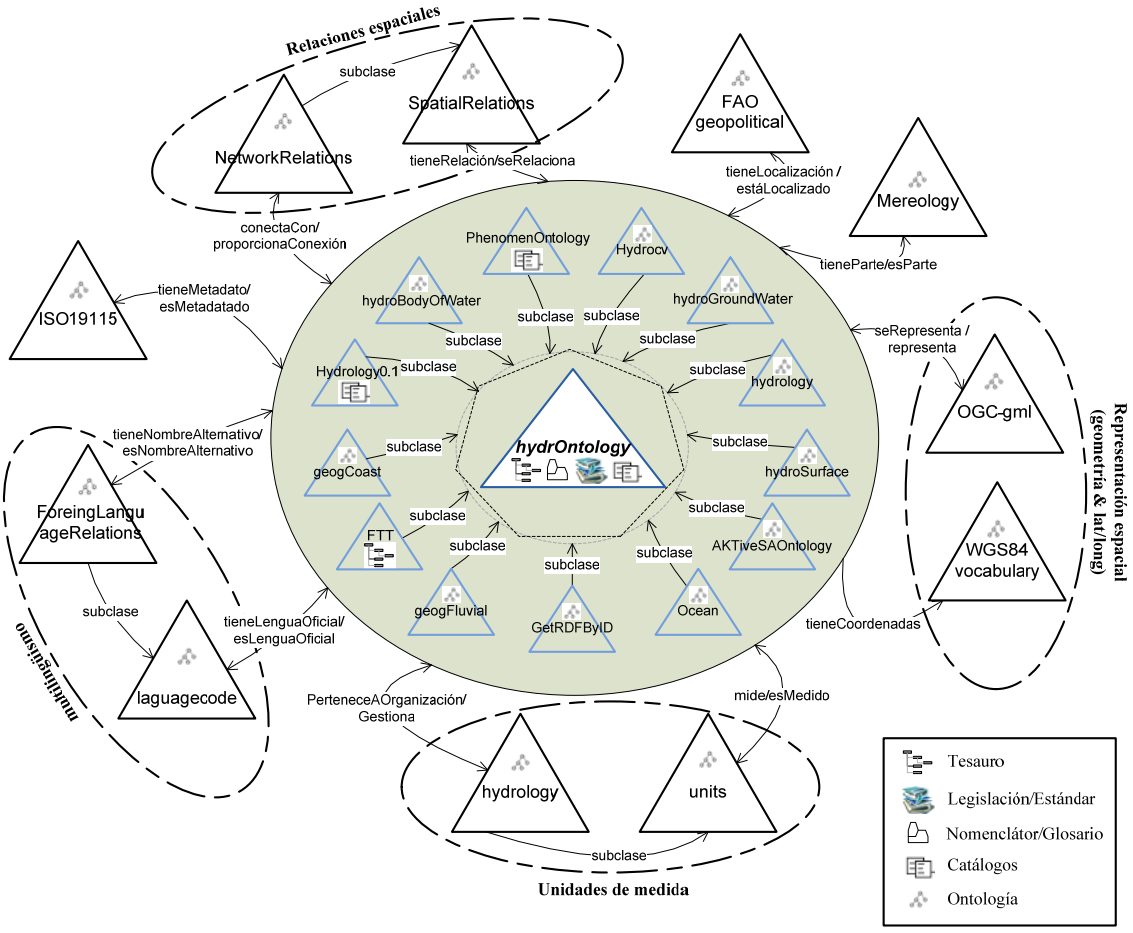


Figura 6.12. Modelo de alto nivel de la red de ontologías en torno a *hydrOntology*

A continuación, en la Figura 6.13, se muestra un ejemplo de documentación extraído de *hydrOntology* perteneciente a la clase "Aguas_Continentales".

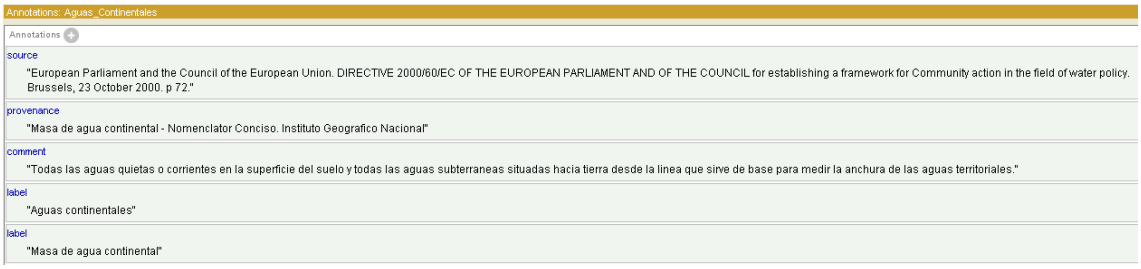


Figura 6.13. Ejemplo de documentación extraído de *hydrOntology*

El proceso de localización de recursos ontológicos (Peters *et al.,* 2007; Espinoza *et al*., 2008), al igual que en el proceso de documentación, se centra en *hydrOntology*. En lo que respecta al resto de recursos ontológicos que conforman la red se adopta el lenguaje de procedencia (inglés). Con respecto a la localización llevada a cabo en la ontología individual, en un primer momento esta ontología fue desarrollada únicamente en español, consecuencia de su contexto local. Con posterioridad, debido al proceso de desarrollo de la red de ontología y al cambio en el propósito y el alcance esta ontología pasa a recoger información en castellano (identificado mediante @es), inglés (identificado mediante @en), catalán (identificado mediante @cat) y francés (identificado mediante @fr) de multitud de componentes definidos en la ontología. En la Figura 6.14 se muestra un ejemplo de localización extraído de hydrOntology perteneciente a la clase "Acuífero". Este proceso de localización de ontologías no sigue el enfoque expuesto en el modelo *Linguistic Information Repository* (LIR) (Montiel-Ponsoda *et al.,* 2008).

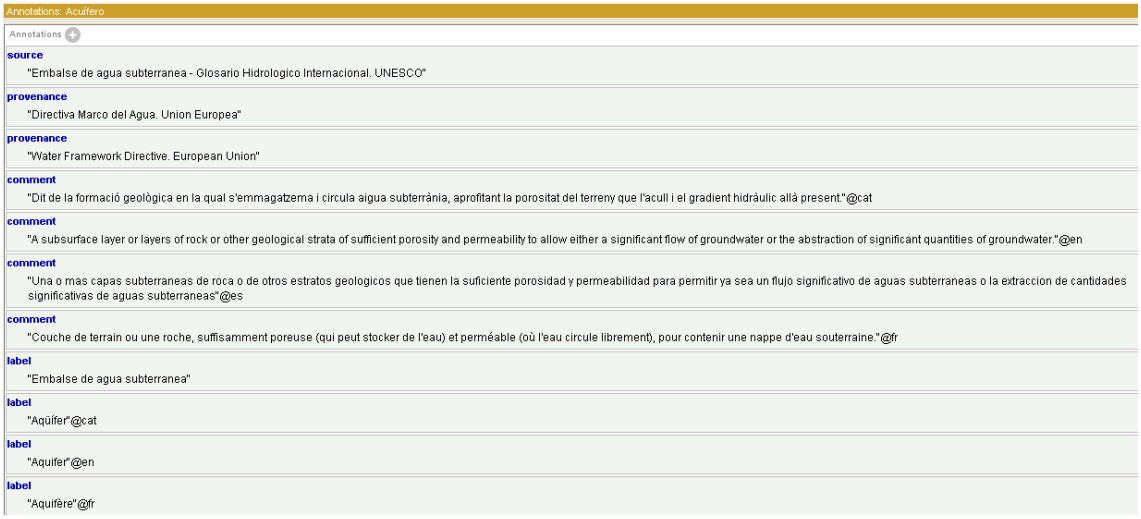


Figura 6.14. Ejemplo de localización extraído de *hydrOntology*

6.5. Conclusiones

En este capítulo se ha presentado la propuesta desarrollada para la armonización del conocimiento hidrográfico representable cartográficamente. Esta propuesta se lleva a cabo a través de la construcción de diferentes recursos ontológicos.

Por un lado, se desarrolla *hydrOntology* como ontología individual (sección 6.3). Su proceso de desarrollo sigue los pasos propuestos en METHONTOLOGY. El propósito de esta ontología es conseguir una visión armonizada de las diferentes fuentes de información del Instituto Geográfico Nacional. *hydrOntology* se desarrolla utilizando únicamente recursos no ontológicos (catálogos de fenómenos, nomenclátores, etc.) seleccionados conforme a una serie de criterios –dependientes e independientes del dominio–. Con respecto al proceso de modelado del conocimiento hidrográfico, este sigue una estrategia *top-down*, un enfoque *Global-as-View* y se rige por diferentes criterios.

Por otro lado, consecuencia de un cambio en el propósito inicial, se realiza el proceso de desarrollo de la red de ontologías en torno a *hydrOntology* (sección 6.4). Su propósito es servir como marco de armonización para los diferentes productores de información geoespacial a través del recubrimiento de la mayoría de fenómenos asociados al dominio hidrográfico. El desarrollo de esta red, basado en la metodología NeOn, utiliza múltiples recursos ontológicos (seleccionados en base a diferentes criterios) sobre los que se aplican diferentes escenarios de la metodología utilizada (reutilización, reutilización y reingeniería de recursos ontológicos). Entre estos recursos aparecen diversos tipos de ontologías geoespaciales acorde al *W3C Geospatial Incubator Group*. Asimismo, se realiza un proceso de documentación y localización de los diferentes elementos que forman parte de *hydrOntology* como ontología individual.

Los recursos ontológicos descritos suponen la propuesta de un vocabulario común, compartido y multilingüe sobre los fenómenos hidrográficos propios del contexto nacional español. Esto supone una contribución para que la comunidad geoespacial pueda sentar las bases de la armonización de esta información. Asimismo, estas ontologías se diferencian del resto de propuestas aparecidas en el hecho de que las ontologías existentes recogen una visión particular y/o limitada del dominio, puesto que son producto de la transformación de determinadas fuentes de conocimiento. Sin embargo, las fuentes de información controladas por las ontologías desarrolladas le proporcionan una visión global del dominio tratado.

Capítulo 7

Generación de instancias y publicación de *Linked Data* en el dominio hidrográfico

Según aparece en Barrasa (2007), para el colectivo de investigadores en bases de datos el nuevo desafío no es ya almacenar más información en menos espacio y tiempo, sino dar el salto[1] de la información ordinaria a la "información con significado" o "semántica", consiguiendo que ésta información sea intercambiable, combinable, integrable entre sistemas a escala global y, sobre todo, procesable de manera automática, de forma que sea posible llevar a cabo inferencias o verificaciones entre otras características. En este contexto se ubica la generación de *Resource Description Framework* (RDF) para el dominio hidrográfico.

En la actualidad, son diversos los conjuntos de datos RDF con información geoespacial presentes en el contexto de la Web de los Datos (Bizer *et al.*, 2009) (DBpedia, CIA World Factbook, GeoNames, etc.). Sin embargo, ninguna de estas fuentes describe adecuadamente la hidrografía del ámbito nacional español con información proveniente de productores oficiales. La consideración de esta circunstancia y la instanciación de la propuesta metodológica de integración de este trabajo conducen a la generación de RDF para la información de este dominio sobre las bases de datos estudiadas.

Este capítulo presenta el proceso seguido para la identificación de los datos, la generación de instancias (RDF) para el conjunto de bases de datos estudiadas y su publicación como datos integrados (*Linked Data*). Este proceso se realiza conforme a las diversas tareas recogidas en las actividades 4 y 5 de la metodología de integración de datos geográficos (Capítulo 5). Así, en la sección 7.1 se describen diferentes aspectos de la identificación de los datos mediante la propuesta de un patrón de URI. Los recursos de información (bases de datos y ontología) utilizados para la generación de RDF se citan en la sección 7.2. Asimismo, en esta sección, se ofrecen algunos detalles sobre las correspondencias (*mappings*) establecidas y se muestran a modo de ejemplo, algunos extractos de los resultados generados durante el proceso de generación de RDF. En la sección 7.3 se describe el proceso de generación de RDF de la información geométrica. En la sección 7.4 se realiza un análisis del RDF resultante y se proponen diversas soluciones a los problemas detectados. La sección 7.5 se dedica a describir el proceso de publicación de los conjuntos de datos RDF obtenidos conforme a los principios de *Linked Data*. Finalmente, en la sección 7.6, se recogen las conclusiones de este capítulo.

[1] El término inglés equivalente es *upgrade*, ascenso, mejora, subida de rango o categoría.

7.1. Identificación de datos

Una de las principales decisiones previa a la generación de instancias, independientemente del sistema utilizado en este proceso, es el formato o patrón en que los identificadores de las instancias van a ser generados. Así, el propósito de esta tarea (tarea 4.1) es la identificación inequívoca de los datos geoespacial considerados en el marco de este trabajo.

En este caso, para la identificación de las instancias se propone la utilización de *Uniform Resource Identifiers* (URI)[2], consecuencia de que como resultado de la generación de instancias se obtendrán tripletas RDF. Las URI son extremadamente relevantes en este proceso, ya que estas contribuirán de manera clave en el alineamiento de instancias provenientes de diferentes fuentes de información. Asimismo, la utilización de URI para la identificación de recursos (instancias) es conforme con uno de los principios en los que se basa *Linked Data*.

Para el diseño del patrón de las URI que identifican los datos en el contexto de este trabajo se adoptan las recomendaciones y buenas prácticas recogidas en (Ayers *et al.*, 2008; Davidson, 2009). A continuación se recogen los principales detalles del patrón:

Raíz de las URI

Se adopta como raíz de las URI `<< http://geo.linkeddata.es>>`. A su vez, este será el dominio donde se publicará toda la información generada en la integración de datos geográficos propuesta en este trabajo.

Codificación

La utilización del español en las URI conlleva un problema de legibilidad de las mismas. Este problema se produce porque, a pesar de que RDF permite la codificación UTF-8, en *Linked Data* las URI deben ser también URL. Esto obliga a que los caracteres no ASCII-US sean codificados utilizando %. A continuación se muestra un ejemplo de URI que incluye caracteres específicos del idioma español y su codificación asociada.

```
http://geo.linkeddata.es/ontology/Río
http://geo.linkeddata.es/ontology/R%C3%ADo
```

Modelo (ontología) y recurso (instancia)

El patrón adoptado para la identificación de un recurso (fenómeno geográfico) modelado en las diferentes ontologías utilizadas es el siguiente:

```
http://geo.linkeddata.es/ontology/<<feature type>>
```

Asimismo, este patrón se utiliza para las propiedades (*object properties*) y atributos (*data properties*) que aparecen en los recursos ontológicos utilizados.

Para identificar a los recursos asociados a los datos (instancias) se adopta el siguiente patrón:

```
http://geo.linkeddata.es/resource/
```

Tipo de fenómenos

Resulta frecuente que diferentes instancias puedan tener un nombre similar y que estas pertenezcan a distintos tipos de fenómenos geográficos. Para distinguir esta posible duplicidad de URI se procede a integrar en el patrón el tipo de fenómeno asociado a cada instancia. De esta manera el patrón adoptado es el siguiente:

```
http://geo.linkeddata.es/resource/<<type>>/<<resourcename>>
```

[2] Una URI es una cadena de caracteres que identifica inequívocamente un recurso (servicio, página, fotografía, documento, dirección de correo electrónico, etc.).

Un ejemplo de URI para la identificación de una instancia asociada al fenómeno "Río"se muestra a continuación:

```
http://geo.linkeddata.es/resource/R%C3%ADo/Arag%C3%B3n%2C%20R%C3%AD

http://geo.linkeddata.es/resource/Río/Aragón, Río
```

Asimismo, en el contexto de este trabajo, este patrón de URI se adopta para la identificación de una instancia de un fenómeno geográfico en su concepción global y única, es decir, este patrón va a identificar a las instancias sin dependencias de los datos capturados en una determinada fuente de información y/o de una determinada visión sobre la instancia en cuestión.

Fuente de información

La inclusión de la fuente de procedencia de las instancias en la URI se debe a la existencia de múltiples fuentes de información geoespacial que capturan la realidad a diferentes escalas y para diferentes objetivos. Al mismo tiempo, esta variedad de información conlleva diversidad en la calidad de los datos. Por tanto, en ocasiones, resulta conveniente conocer la procedencia de dicha información. A continuación se presenta un ejemplo del patrón de URI establecido.

```
http://geo.linkeddata.es/resource/<dataset>/<type>/<resourcename>
http://geo.linkeddata.es/resource/BTN25/Rio/Río_Ebro
```

Geometría (*Hash*)

Una opción contemplada fue la inclusión en la URI de un valor numérico, generado de forma automática, extraído de la geometría del fenómeno. De esta manera, se pretendía identificar de manera única a un determinado recurso.

```
http://geo.linkeddata.es/resource/hash/<resourcename>
http://geo.linkeddata.es/resource/0aede55f6d98/Río_Ebro
```

Una práctica similar se considera en la información publicada como *Linked Data* por el Ordnance Survey[3]. Finalmente, esta opción se descartó por considerar que esta forma de generar las URI no era conforme con las recomendaciones y buenas prácticas recogidas en (Ayers *et al.*, 2008; Davidson, 2009).

Finalmente, en la Figura 7.1 se muestra una visión general de la estructura del patrón de URI adoptado para la identificación global de datos geoespaciales (instancias) en el contexto de este trabajo. La información asociada a estos recursos es general y objetiva en la descripción de su realidad (independiente de cualquier fuente de información).

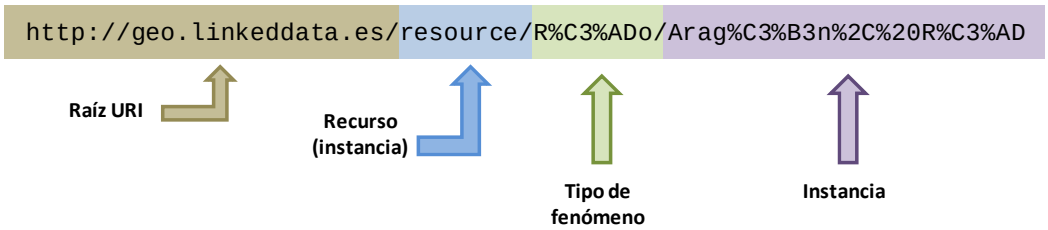


Figura 7.1. Estructura del patrón de URI adoptado para la identificación global de instancias

Mientras, en la Figura 7.2 se recoge la estructura del patrón adoptado para la identificación de las instancias de una determinada fuente de información. La información asociada a estos recursos describe una abstracción particular de la realidad geoespacial de una instancia.

[3] Un ejemplo de esta práctica es la URI del recurso Southampton publicada por el Ordnance Survey http://data.ordnancesurvey. co.uk/id/50kGazetteer/218013

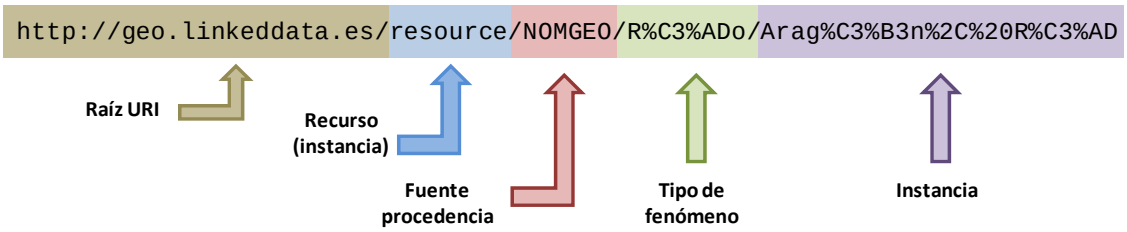


Figura 7.2. Estructura del patrón de URI adoptado para la identificación de instancias asociadas a una fuente de datos geográficos

7.2. Generación de RDF

En esta sección se realiza un proceso de integración estructural y semántica de las diferentes instancias existentes en el escenario de múltiples bases de datos tratadas en este trabajo. El objetivo de este proceso de integración es poblar de instancias una ontología con el contenido de las bases de datos. Paralelamente, con este objetivo se consigue establecer una armonización estructural de las diferentes bases de datos en torno a la referencia semántica de una ontología global de dominio. El resultado de este proceso será la transformación de las bases de datos relacionales estudiadas a tripletas RDF, es decir, en bases de conocimiento. A continuación se presentan los diferentes recursos de información considerados en este proceso.

7.2.1. Recursos de información

El proceso de generación de RDF está basado en la utilización de un conjunto de recursos de información. Estos recursos están compuestos de una ontología y diversas bases de datos. A continuación se describen algunas de las características de estos componentes.

7.2.1.1. hydrOntology

La ontología elegida para el establecimiento de *mappings* es *hydrOntology* (Vilches-Blázquez *et al.*, 2009) como ontología individual. Una descripción detallada de las características de esta ontología se recoge en el Capítulo 6 de este trabajo.

La herramienta que da soporte al plugin ODEMapster++ (NeOn toolkit) trata las ontologías que forman parte de una red como ontologías individuales. Por tanto, al carecer de una visión conjunta, la generación de *mappings* se convierte en un proceso individualizado para cada una de las ontologías participantes. Este hecho resultó determinante en la decisión de seleccionar únicamente a *hydrOntology* para la generación de correspondencias con las diferentes bases de datos.

7.2.1.2. Bases de datos

En este trabajo se consideran diferentes bases de datos geográficas, con diferente granularidad, escalas de captura (desde 1:10.000.000 a 1:5.000) e información. Dichas bases de datos proceden de instituciones o productores a diversos niveles (europeas, nacionales y locales). Con respecto a su contenido, este trabajo se centra en la información hidrográfica relacionada con instancias de fenómenos geográficos españoles (componente común del conjunto de bases de datos). Además, aunque con algunas excepciones, estas fuentes poseen información multilingüe en los diferentes idiomas oficiales nacionales. Una información más detallada sobre las bases de datos estudiadas aparece en el Anexo.

7.2.2. Establecimiento de *mappings*

Desde una perspectiva general una correspondencia o *mapping* se define como una explicitación formal de una relación entre elementos, o conjuntos de elementos, de diferentes recursos de conocimiento (Ramos *et al.*, 2008). La explicitación de relaciones entre elementos

convierte la generación de *mappings* en una tarea crucial en muchos dominios de aplicación, por ejemplo: integración de ontologías o esquemas, almacenes de datos (*data warehouses*), integración de datos, etc.

De entre los diferentes tipos de *mappings*, recopilados en (Ghawi *et al.*, 2007), este trabajo se centra en la generación de *mappings* entre bases de datos y ontologías. Esta tipología se caracteriza por el hecho de que una base de datos y una ontología están semánticamente relacionadas con el nivel conceptual.

Para la generación de instancias de *hydrOntology* se establecen *mappings* (tarea 4.2) entre los diferentes componentes de la referida ontología con instancias de los esquemas de las bases de datos descritas con anterioridad, donde se almacenan los topónimos vinculados con la información hidrográfica y el resto de información asociada. Con este acercamiento se mantienen separadas las instancias o individuos de la estructura de la ontología. Así, *hydrOntology* únicamente contiene las clases y propiedades pero no las instancias, que permanecen en las diversas bases de datos. Estas instancias son recuperadas y traducidas como respuesta a las consultas de los usuarios. En la Figura 7.3 se puede observar la propuesta de integración de información realizada por *hydrOntology*. Ésta propuesta es conforme con el enfoque de integración basado en una única ontología presentado en (Wache *et al*., 2001).

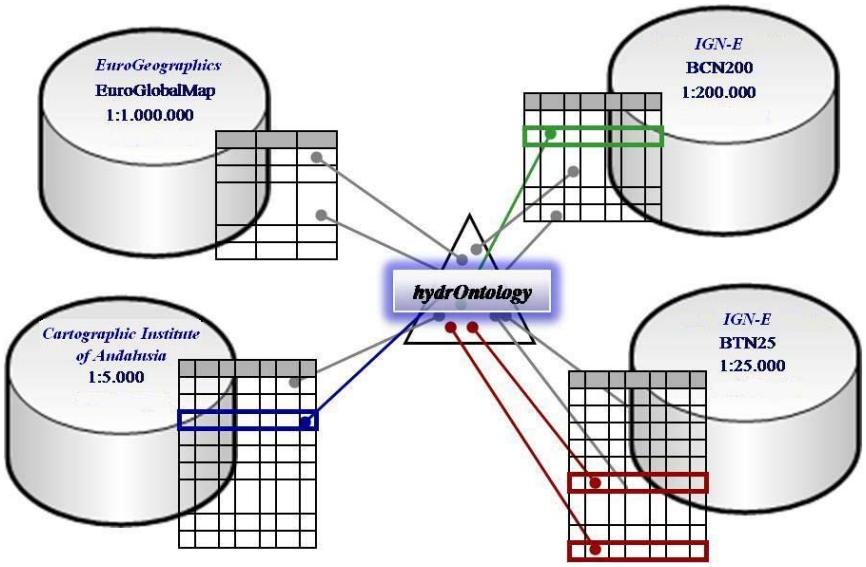


Figura 7.3. Visión de alto nivel de los *mappings* entre *hydrOntology* y algunas de las bases de datos estudiadas

En la realización de esta tarea un aspecto importante a considerar es el hecho de que la información puede residir en múltiples fuentes de información, en ocasiones, de forma redundante. Por este motivo, la generación de *mappings* de todas las fuentes de información estudiadas puede implicar creación de información (tripletas) repetida.

Dentro de los diversos lenguajes y motores de inferencia se opta por la utilización del lenguaje R_2O+ para el establecimiento de *mappings* y del plugin ODEMapster++ como motor de inferencia. La utilización de estos elementos permite la generación de instancias de *hydrOntology*, es decir, la transformación de las bases de datos relacionales a RDF.

En la Figura 7.4 se identifican los elementos fundamentales en el escenario de enriquecimiento o "*upgrade*" del contenido de una base de datos relacional de acuerdo a una ontología, conforme a lo expuesto en (Barrasa, 2007). En la mencionada figura pueden observarse los elementos claves anteriormente descritos: un procesador de consultas descritas en términos de la ontología (ODEMapster++) y una descripción declarativa de las correspondencias (expresada en lenguaje

R_2O+) existentes entre la ontología y el modelo relacional de la base de datos sobre la que se lleva a cabo el "*upgrade*". En base a dicha descripción el procesador reescribe la consulta en otra consulta equivalente en términos del modelo relacional de la base de datos.

7.2.3. Generación de resultados

Una vez expuestos algunos de los aspectos teóricos de los elementos utilizados en el establecimiento de *mappings* y la inferencia de resultados entre la ontología y las bases de datos, a continuación se ofrecen algunos detalles del proceso realizado para la generación de tripletas RDF.

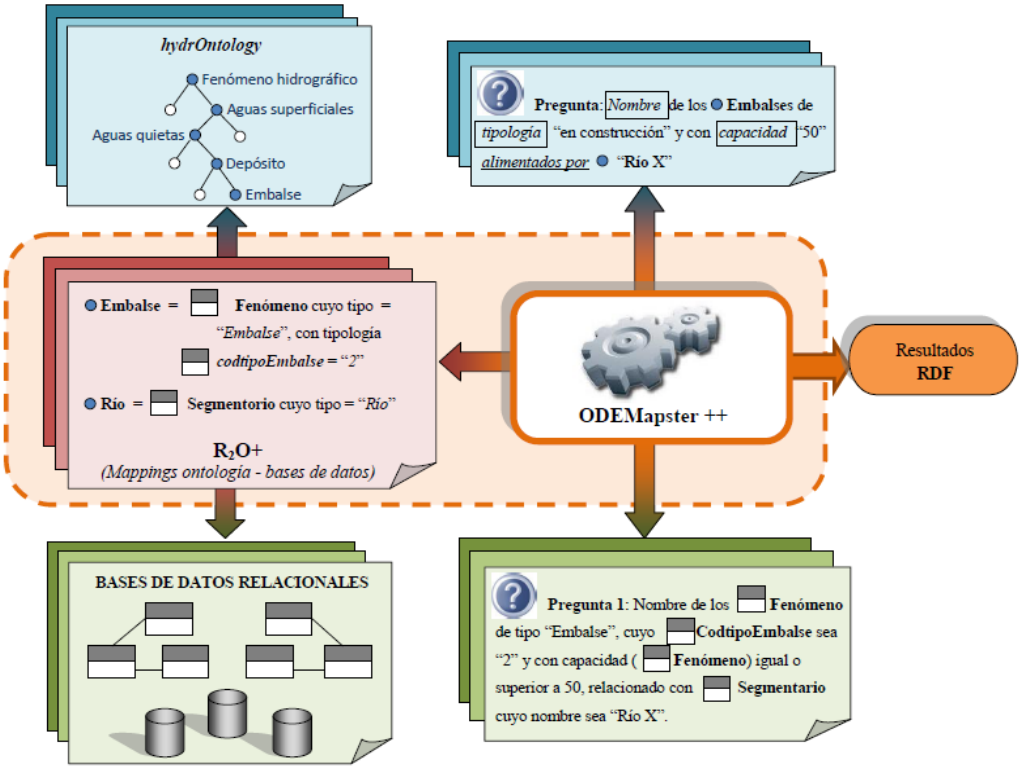


Figura 7.4. Escenario global de establecimiento de *mappings* entre *hydrOntology* y múltiples bases de datos para upgrade

La generación de *mappings* se puede realizar a través de la interfaz gráfica del plugin ODEMapster. Este interfaz permite a los usuarios crear *mappings* gráficamente (expresados en lenguaje R2O+) ejecutar y consultar dichos mappings a múltiples bases de datos. En la Figura 7.5 se muestra un extracto de algunos *mappings* establecidos entre la ontología y la base de datos de la CHE.

Esta interfaz consta de cuatro componentes principales, que son:

- Visor de bases de datos. Este componente muestra el esquema de una o múltiples bases de datos mediante una representación en árbol de las diferentes tablas y campos conformantes. Esta sección del plugin ODEMapster++ aparece numerada en la Figura 7.5 con el número 1.
- Visor de ontologías. Este componente recoge en un árbol la ontología generada o importada en NeOn Toolkit. En la Figura 7.5 esta sección del plugin aparece con el número 2.
- Panel de *mapping*. Este panel es el lugar donde el usuario realiza los *mappings* entre los elementos de la base de datos y los componentes de la ontología. Este panel aparece en la Figura 7.5 señalado con el número 3.

- Panel de control. En este panel se localizan las condiciones y operaciones que ODEMapster++ permite al usuario. Estas condiciones y operaciones son sólo un subconjunto del lenguaje R_2O, aunque conforman las primitivas más utilizadas del mencionado lenguaje. Este panel de control se encuentra en la sección 4 de la Figura 7.5. Una descripción detalla de las operaciones y condiciones posibles dentro de ODEMapster++ se lista en (Priyatna, 2009).

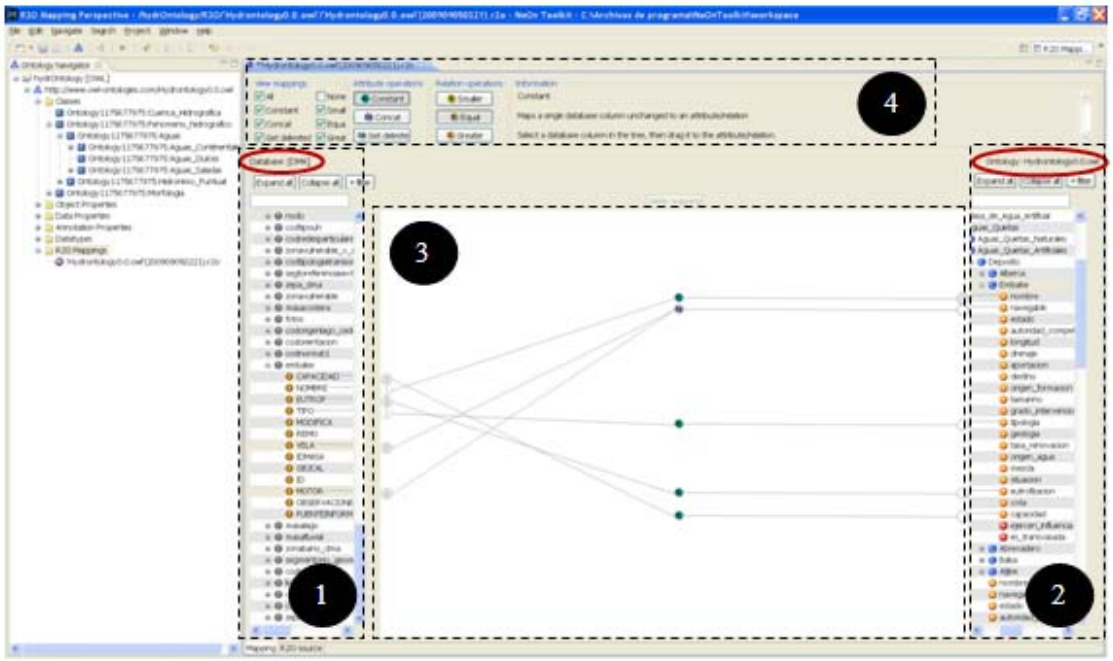


Figura 7.5. Visión general de *mappings* establecidos con el plugin ODEMapster++

Mappings R$_2$O+

El establecimiento de *mappings* R_2O+ genera un documento que recoge las relaciones entre los componentes de la ontología y los de la base de datos. Asimismo, incluye una completa descripción del esquema de la base de datos, un conjunto de definiciones del *mapping* de conceptos que consta de los nombres de los conceptos con sus columnas identificadoras y un conjunto de definiciones de *mappings* de relaciones y atributos. A modo de ejemplo, en la Tabla 7.1 se muestra un extracto representativo del código R_2O+ generado en el establecimiento de *mappings* entre la base de datos del Nomenclátor Geográfico Conciso de España y *hydrOntology*.

La codificación del lenguaje R_2O recogida en la Tabla 7.1 muestra los *mappings* correspondientes al fenómeno "Embalse" con algunos de sus atributos (*DataProperties*). A continuación, con objeto de conseguir un mayor entendimiento de la referida codificación se describen algunos de sus componentes principales.

- En la sección 1 se encuentra el elemento `<conceptmap-def>`. Este elemento es el encargado de explicitar la relación entre el concepto de la ontología y las instancias de la base de datos que también lo son de dicho concepto. En términos procedurales se podría decir que describe cómo obtener instancias del concepto de la ontología a partir de la base de datos.
- El elemento `<attributemap-def>`, presente en la sección 2, describe una correspondencia para un atributo entre base de datos y ontología. Asimismo, en esta sección, aparecen dos ejemplos de las operaciones de transformación de primitivas, propuestas en el lenguaje R_2O, presentes con mayor frecuencia en este trabajo, son:

Tabla 7.1. Ejemplo de codificación de los *mappings* R_2O+

```
<conceptmap-def name=" http://www.oeg-upm.net/ontologies/hydrOntology.owl
#Embalse">
        <uri-as>
           <operation oper-id="concat">
            <arg-restriction on-param="string1">
                        <has-value>http://www.oeg-
              upm.net/ontologies/hydrOntology.owl# Embalse
              </has-value>
            </arg-restriction>
            <arg-restriction on-param="string2">
                    <has-column>ngce.nomenclatorgce.ID</has-column>
            </arg-restriction>
           </operation>
        </uri-as>
 <described-by>
           <attributemap-def name="http://www.oeg-
           upm.net/ontologies/hydrOntology. owl# nombre">
            <selector>
                <aftertransform>
                  <operation oper-id="concat">
                   <arg-restriction on-param="string1">
                  <has-column>ngce.nomenclatorgce.Tipo_Entidad</has-column>
                   </arg-restriction>
                   <arg-restriction on-param="string2">
                     <has-column>ngce.nomenclatorgce.Nombre
                     </has-column>
                   </arg-restriction>
                  </operation>
                </aftertransform>
            </selector>
 </attributemap-def>

           <attributemap-def name="http://www.w3.org/2000/01/rdf-
           schema#label">
                      <has-language>
                          <has-value>es</has-value>
                      </has-language>
                      <selector>
                          <aftertransform>
                              <operation oper-id="constant">
                                  <arg-restriction on-param="string1">
                          <has-
           column>ngce.nomenclatorgce.nombreOficial</has-column>
                                  </arg-restriction>
                              </operation>
                          </aftertransform>
                      </selector>
 </attributemap-def>
 </described-by>
</conceptmap-def>
```

- `"concat".` Esta operación concatena dos cadenas alfanuméricas presentes en dos columnas de la base de datos con un atributo/relación de la ontología. En el ejemplo recogido en la Tabla 7.1 (sección 2) se realiza la concatenación del "tipo de fenómeno" y "nombre" para obtener, a través de la consulta, la información completa del nombre de la instancia.
- `"constant".` Esta operación establece una correspondencia entre una columna de la base de datos y un atributo/relación de la ontología, devolviendo el valor sin modificación alguna. En el ejemplo de la Tabla 7.1 (sección 2) esta operación establece correspondencias con los atributos Latitud y Longitud.

Más detalles de los componentes del lenguaje R_2O pueden ser encontrados en (Barrasa, 2007).

Consultas y generación de RDF

El plugin ODEMapster++ también permite consultar la ontología y conseguir obtener como salida resultados en RDF. El lenguaje de consulta es ODEMQL (ODEMapster Query Language) (Barrasa, 2007), diseñado específicamente para implementar el tipo de consultas sobre el motor ODEMapster. Este lenguaje permite expresar consultas basadas en la terminología de la ontología. Tiene sintaxis XML y como resultado genera emparejamientos variable-valor (*bindings*) o bien conjuntos de individuos en forma de modelo RDF. Una descripción exhaustiva de la gramática de este lenguaje aparece en (Barrasa, 2007).

Continuando con el ejemplo anterior (*mappings* R_2O+ establecidos entre la base de datos de la NGCE e *hydrOntology*) en la Tabla 7.2 se muestra un extracto de consulta generada con el plugin ODEMapster++ sobre los alineamientos establecidos.

Tabla 7.2. Ejemplo de consulta ODEMQL

```
<query>
 <onConcept conceptUri="http://www.oeg-upm.net/ontologies/hydrOntology.owl#
 Embalse">


     <attSelect>
        <onAtt attName=" http://www.oeg-upm.net/ontologies/hydrOntology.owl#
 nombre"/>
     </attSelect>
     <attSelect>
        <onAtt attName=" http://www.oeg-upm.net/ontologies/hydrOntology.owl#
 lat."/>
     </attSelect>
     <attSelect>
        <onAtt attName=" http://www.oeg-upm.net/ontologies/hydrOntology.owl#
 long."/>
     </attSelect>
 </onConcept>
 </query>
```

La expresión `onConcept conceptUri` (sección 1, Tabla 7.2) equivale a solicitar todos los individuos que verifican: $x|Embalse(x)$. Asimismo, con el elemento `<attSelect>` se extraen algunos atributos (nombre, latitud y longitud) para dichos individuos (sección 2, Tabla 7.2). El resultado de la consulta de este ejemplo, como se mostrará con posterioridad, contendrá los identificadores (URI) de los individuos del fenómeno "Embalse" y los valores correspondientes a los atributos Nombre, Latitud y Longitud.

Como resultado del proceso de generación de *mappings* y consultas se extraen diferentes conjuntos de datos RDF iniciales correspondientes a las diversas bases de datos estudiadas. En la siguiente sección se procede a analizar los datos RDF resultantes.

7.3. Generación de RDF de la información geométrica

Un aspecto característico y diferenciador a considerar en el proceso de transformación de la información geográfica es la definición de la información geométrica en RDF. A continuación se presentan las diversas formas de definir la geometría en el contexto de la información geoespacial y, posteriormente, se describe la propuesta desarrollada para la creación de RDF para geometría contenida en bases de datos geospaciales.

7.3.1. Geometría: WKT, WKB y GML

De acuerdo con (Schade, 2010), los estándares en la comunidad geoespacial son, por diseño, compatibles con la filosofía propuesta por *Linked Data*. Una buena prueba de esto son las especificaciones del Open Geospatial Consortium (OGC). Esta sección se centra en aquellas que definen la representación geométrica de los datos geoespaciales.

Por un lado, OGC define **WKT** (*Well-Know Text*), un lenguaje de marcado de texto para la representación de objetos geométricos vectoriales sobre un mapa, los sistemas de referencia espacial de los objetos espaciales y las transformaciones entre los sistemas de referencia espacial, y **WKB** (*Well-Know Binary*), un equivalente binario de WKT que se utiliza para transformar y almacenar la misma información en las bases de datos. WKT y WKB presentan las siguientes formas geométricas: Punto, LineString, Polígono, MultiPunto, MultiLineString, MultiPolígono y GeomCollection.

Por otro lado, **GML**, como se menciona en el Capítulo 3 de este trabajo, es la instanciación de XML definida por OGC para el almacenamiento, transporte, intercambio, procesamiento y transformación de la IG. La implementación de este lenguaje, como se describe en (Schade, 2010), estuvo fuertemente influenciada por el lenguaje RDF. La primera versión de GML tenía una implementación RDF/XML explícita y las versiones posteriores implementan un patrón objeto-propiedad e incluyen vínculos XML (W3C, 2001) de manera isomorfa a RDF.

En cualquier caso, la decisión de utilizar una representación GML o RDF de los datos geoespaciales depende del uso previsto. Ambos lenguajes están optimizados para propósitos distintos, mientras RDF permite consultas complejas y razonamiento, numerosos clientes de Sistemas de Información Geográfica y servicios de Infraestructuras de Datos Espaciales utilizan GML.

7.3.2. Geometría y RDF

En esta sección se describe el proceso desarrollado para la transformación de la información geométrica almacenada en diferentes bases de datos geoespaciales a RDF (De Leon *et al.*, 2010). Este trabajo se relaciona con la generación de instancias propuesta en la tarea 4.3 de la metodología descrita. En la Figura 7.6 se muestra una visión general del proceso de transformación.

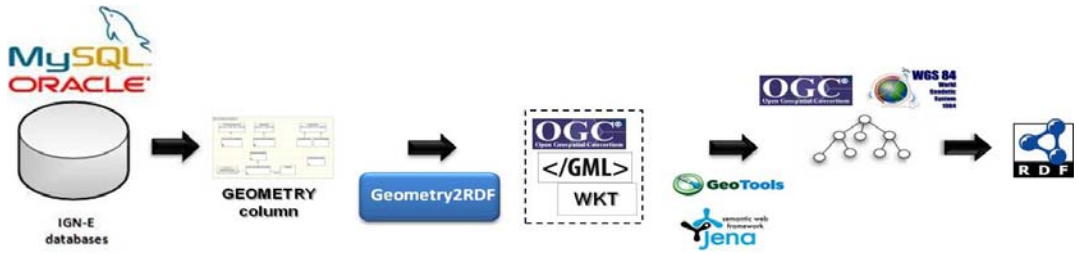


Figura 7.6. Transformación de la información geométrica en RDF

Para aquellas bases de datos que utilizan **Oracle** como SGBD, se utiliza el paquete Oracle SDO UTIL para la transformación de los datos geométricos almacenados en las bases de datos a GML. La generación de GML se lleva a cabo sobre la columna GEOMETRY, donde diferentes filas de una tabla almacenan la información geométrica de cada fenómeno. En la Tabla 7.3 se muestra un extracto del GML generado.

Tabla 7.3. Fragmento del GML generado

```
<gml:LineString srsName="EPSG: 4230"
xmlns:gml="http://www.open gis.net/gml">
<gml:coordinates decimal="." cs="," ts=" ">
-2.76263331,38.9686579,0 -2.7632481,38.96774903,0
-2.76497904,38.96788769,0 -2.76722341,38.96633326,0
-2.76884046,38.96599405,0 -2.77255198,38.96759632,0
-2.77433045,38.96731141,0 -2.77838617,38.9622369,0
-2.77957737,38.96146419,0 -2.78495751,38.96088851,0
…
</gml:coordinates>
</gml:LineString>
```

Para aquellas bases de datos que utilizan **MySQL** como SGBD, se trabaja con el lenguaje WKT para extraer la información de la columna GEOMETRY. En este caso, se trabaja directamente con el lenguaje WKT debido a que no existe ninguna función para generar GML en estas bases de datos. La Tabla 7.4 recoge un ejemplo del WKT generado.

Tabla 7.4. Un ejemplo del WKT generado

```
POINT(39.4666666666667  -0.366666666666667)
```

Tras la extracción de la información geométrica se procede a convertir el GML y WKT generado en RDF. Para esto se ha desarrollado una librería de software, denominada **GEOMETRY2RDF**, que define un conjunto de tripletas RDF para la geometría (que puede estar disponible en GML o WKT). El GML y WKT generados con anterioridad se manipula con GeoTools[4] para recuperar la geometría y también puede ser utilizar, si resulta necesario, para la transformación de coordenadas. Finalmente, se utiliza Jena[5] para generar el RDF geoespacial resultante de este proceso (ver Figura 7.7).

```
<rdf:Description rdf:about="http://geo.linkeddata.es/resource/wgs84/39.158523176194414_-1.6341427210305737">
  <geontology:order rdf:datatype="http://www.w3.org/2001/XMLSchema#int">205</geontology:order>
  <geo:long rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">-1.6341427210305737</geo:long>
  <geo:lat rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">39.158523176194414</geo:lat>
  <rdf:type rdf:resource="http://www.w3.org/2003/01/geo/wgs84_pos#Point"/>
</rdf:Description>
<rdf:Description rdf:about="http://geo.linkeddata.es/resource/wgs84/38.65501872007066_-2.6158910085778153">
  <geontology:order rdf:datatype="http://www.w3.org/2001/XMLSchema#int">6</geontology:order>
  <geo:long rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">-2.6158910085778153</geo:long>
  <geo:lat rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">38.65501872007066</geo:lat>
  <rdf:type rdf:resource="http://www.w3.org/2003/01/geo/wgs84_pos#Point"/>
</rdf:Description>
<rdf:Description rdf:about="http://geo.linkeddata.es/resource/wgs84/38.37015012539368_-1.6964507982671462">
  <geontology:order rdf:datatype="http://www.w3.org/2001/XMLSchema#int">141</geontology:order>
  <geo:long rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">-1.6964507982671462</geo:long>
  <geo:lat rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">38.37015012539368</geo:lat>
  <rdf:type rdf:resource="http://www.w3.org/2003/01/geo/wgs84_pos#Point"/>
</rdf:Description>
```

Figura 7.7. Extracto del RDF geoespacial generado con GEOMETRY2RDF

El RDF generado es compatible con WSG84 *vocabulary* y la ontología GML, ambos integrados en la red de ontologías conformada en torno a *hydrOntology*, descrita en detalle en el Capítulo 6 de este trabajo. La Figura 7.8 ilustra la información que describe a la instancia "Río Ebro" junto con su geometría asociada.

[4] *GeoTools*. Librería Java de código abierto que proporciona herramientas para el tratamiento de la información geoespacial. http://www.geotools.org/

[5] *Jena*. Es un *framework* de Java para la construcción de aplicaciones de la Web Semántica. Éste *framework* proporciona un entorno programático para RDF, RDFS, OWL, SPARQL e incluye un motor de inferencia basado en reglas. Mas detalles en: http://jena.sourceforge.net/

Los conjuntos de información RDF generados de las distintas fuentes de información tratadas están disponibles para consulta en la Web[6]. Asimismo, la calidad de estos datos se trata en la evaluación realizada en el Capítulo 9 de este trabajo.

7.4. Revisión del RDF resultante

Tras la generación de *mappings* con el plugin ODEMapster++ se realiza un proceso compuesto de tres fases dependientes:

- Por un lado, se realiza un análisis de los resultados obtenidos para detectar posibles problemas en los *mappings* establecidos, conforme a la tarea 4.4 (*Revisión y retroalimentación de expertos*) de la metodología propuesta en este trabajo. En esta fase de análisis de los resultados obtenidos se selecciona una muestra de cada una de las fuentes. Esta muestra está compuesta por un conjunto de instancias, en torno al 20-25% del número total, seleccionadas aleatoriamente. En este análisis, fundamentalmente, se realiza una comprobación manual de la correspondencia de las diferentes instancias con el tipo de fenómeno asociado a través de los *mappings* generados.
- Posteriormente, se lleva a cabo la depuración de la codificación de los *mappings* establecidos con objeto de solventar diversos problemas detectados en la generación inicial de RDF y otros provenientes de la heterogeneidad existentes en las bases de datos estudiadas.
- Finalmente, se generan nuevos conjuntos de datos RDF de gran calidad.

A continuación se describen los principales detalles de las distintas fases del proceso de revisión.

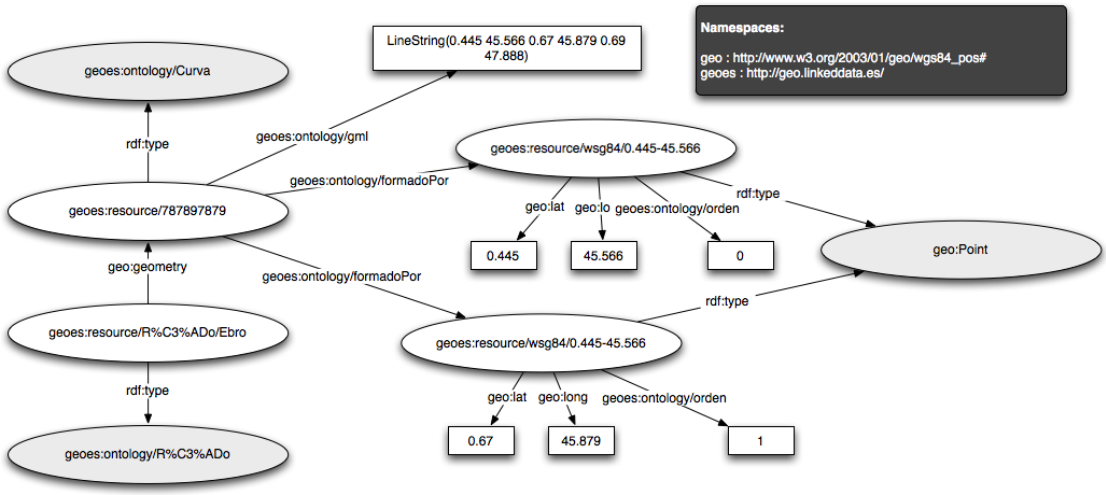


Figura 7.8. Río Ebro y su información geométrica

7.4.1. Análisis. Problemas detectados

Durante la fase de análisis, desde una perspectiva general, se ha comprobado que gran parte de los problemas que generan estos datos están causados por las diferencias léxico sintácticas entre las diferentes bases de datos, es decir, por problemas relacionados con aspectos de acentuación, mayúsculas y/o minúsculas. Estos problemas conllevan la pérdida de información en la generación de tripletas RDF, consecuencia de las diferencias existentes entre las cadenas (*strings*) presentes en el conjunto de las bases de datos estudiadas y la ontología.

[6] http://geo.linkeddata.es/web/guest/endpoints

Asimismo, se han detectado diversos problemas relativos a la calidad de los *mappings* establecidos. Estos problemas, descritos a continuación, son clasificados como simples y complejos.

Los ***problemas*** clasificados como ***simples*** están vinculados con leves modificaciones en la codificación de los *mappings* generados. A este nivel se encuentran:

- **Identificadores.** La existencia de URI conformes con los identificadores (ID) presentes en las diferentes bases de datos provoca los siguientes problemas:

  - ***Duplicados***. La presencia de instancias duplicadas dentro de las diferentes bases de datos persiste en los conjuntos de tripletas RDF generadas (ver Tabla 7.5). Esto se produce a pesar de que ODEMapster genera tripletas RDF con URI únicas, mediante la operación SQL `Select distinct`. Por tanto, se puede afirmar que no se están generando verdaderas URI para la identificación de las instancias.

Tabla 7.5. Ejemplo de problemas de duplicados

```
 <rdf:Description rdf:about="http://www.oeg-upm.net/ontologies/hydrOntology
 .owl#Río319599">
        <rdf:type rdf:resource="http://www.oeg-
    upm.net/ontologies/hydrOntology .owl#Río"/>
   <j.0:nombre>RÍO TAJO/RÍO TEJO(EMBALSE DE CEDILLO) </j.0:nombre>
    <j.0:superficie>0</j.0:superficie>
    <j.0:longitud>0</j.0:longitud>
</rdf:Description>
 <rdf:Description rdf:about="http://www.oeg-upm.net/ontologies/hydrOntology
 .owl#Río319790">
    <j.0:longitud>0</j.0:longitud>
    <rdf:type rdf:resource="http://www.oeg-upm.net/ontologies/hydrOntology.owl
#Río"/>
   <j.0:nombre>RÍO TAJO/RÍO TEJO(EMBALSE DE CEDILLO) </j.0:nombre>
    <j.0:superficie>0</j.0:superficie>
</rdf:Description>
```

  - ***Registros sin información***. La presencia de registros sin información resulta común entre las diferentes bases de datos estudiadas. Esta ausencia de datos junto a la generación de URI acordes a los IDs de las fuentes de datos conlleva la generación de tripletas RDF con registros en blanco. Esta situación genera a su vez un aumento innecesario del volumen de la información. En la Tabla 7.6 aparece un ejemplo representativo de este problema.

Tabla 7.6. Ejemplo de registros sin información

```
<rdf:Description rdf:about="http://www.oeg-upm.net/ontologies/hydrOntology
.owl#Río55274">
       <rdf:type rdf:resource="http://www.oeg-upm.net/ontologies/hydrOntology
   .owl#Río"/>
   <j.0:nombre></j.0:nombre>
 </rdf:Description>
```

- **Tipo fenómeno**. La práctica extendida de incluir el tipo de fenómeno en el nombre de las instancias (bases de datos) y su inserción en el establecimiento del *mapping* como elemento proveniente de la ontología genera reiteración de las tipologías de fenómeno en el nombre de las instancias. Un ejemplo de esta problemática se muestra en la Tabla 7.7.

Tabla 7.7. Ejemplo de repetición del tipo de fenómeno

```
<rdf:Description rdf:about="http://www.oeg-upm.net/ontologies/hydrOntology
.owl#Balsa Balsa del Planeron">
       <rdf:type rdf:resource="http://www.oeg-upm.net/ontologies/hydrOntology
   .owl#Balsa"/>
 </rdf:Description>
```

En el ***nivel complejo*** los problemas existentes están relacionados con la presencia de instancias de tipología diversa dentro de una misma tabla, provocando la necesidad de generar expresiones de *mappings* avanzadas entre bases de datos y ontología. Entre los problemas encontrados destacan los siguientes:

- **Conectores**. Este problema se caracteriza por la existencia de instancias, asociadas a un mismo fenómeno, con y sin conectores dentro del nombre (por ejemplo, "Laguna de Carravalseca" y "Laguna Larga").
- **Inclusión**. Este problema se caracteriza por la presencia de tripletas alineadas con fenómenos distintos del que aparece en el nombre de la instancia. Se produce por la inclusión (total o parcial) del tipo de fenómeno en el nombre de la instancia. En la Tabla 7.8 se recoge un ejemplo de esta problemática (la instancia "Embalse de Navacerrada" incluye en su nombre el tipo de fenómeno "Nava").

Tabla 7.8. Ejemplo de inclusión del fenómeno en el nombre de instancia

```
<rdf:Description rdf:about="http://www.oeg-upm.net/ontologies/hydrOnto
logy.owl#Embalse de Navacerrada.id">

     <rdf:type rdf:resource="http://www.oeg-upm.net/ontologies/hydrOnto
     logy.owl#Nava"/>

</rdf:Description>
```

- **No inclusión**. Esta problemática se caracteriza por la existencia de tripletas con fenómenos distintos del que aparece en el nombre de la instancia. A diferencia del caso anterior, en este problema no aparece inclusión (total o parcial) del tipo de fenómeno en el nombre de la instancia. La Tabla 7.9 incluye un ejemplo de este problema (la instancia "Laguna La Albuera" aparece asociada al tipo de fenómeno "Embalse".

Tabla 7.9. Ejemplo de tripletas con correspondencias distintas entre el tipo de fenómeno y el nombre de instancia

```
<rdf:Description rdf:about="http://www.oeg-upm.net/ontologies/hydrOntology
.owl#/Laguna La Albuera.id">

     <rdf:type rdf:resource="http://www.oeg-upm.net/ontologies/hydrOntolo
gy.owl#Embalse"/>

</rdf:Description>
```

- **Valor de un campo**. Los problemas derivados de la ausencia de tipología de fenómenos, algunas de las fuentes de información no incluyen el tipo de fenómeno en el campo (columna) destinado al nombre de las instancias. En ocasiones, la tipología (fenómeno) asociada a las instancias aparece en una columna diferente de la relacionada con el nombre de la instancia. También resulta frecuente que la tipología aparezca identificada en el propio nombre de la tabla, a pesar de que en su contenido pueda presentarse una miscelánea de fenómenos. Estas situaciones conllevan la posibilidad de que el establecimiento directo de *mappings* al nombre de las instancias incluya

cualquier tipo de información, incluso que aparezca información no adscrita al dominio hidrográfico. Este hecho resulta más probable cuando la bases de datos contiene información de diversos dominios. En la Tabla 7.10, donde se muestra un ejemplo de este problema, aparece una instancia (Aeropuerto de Menorca) asociada a un tipo de fenómeno hidrográfico (Canal).

Tabla 7.10. Ejemplo de tripletas fuera del dominio hidrográfico relacionadas con el fenómeno "Canal"

```
<rdf:Description rdf:about="http://www.oeg-upm.net/ontologies/hydrOntology
.owl#Canal Menorca, Aeropuerto de">

   <rdf:type rdf:resource="http://www.oeg-upm.net/ontologies/hydrOntology
.owl#Canal"/>

   <j.1:nombre>AeropuertoMenorca, Aeropuerto de</j.1:nombre>
   <j.0:idioma>es</j.0:idioma>
   <j.0:lat.>39º 52'</j.0:lat.>
   <j.0:long.> 4º 13'</j.0:long.>
</rdf:Description>
```

- **Valor de un campo + inclusión**. La combinación de los problemas relacionados con el valor de un campo y la inclusión de fenómenos origina un nuevo tipo de problema a considerar. Un ejemplo de esta problemática se muestra en la Tabla 7.11 (el nombre de instancia Canals es asociado al tipo de fenómeno "Canal" únicamente por similitud entre cadenas (*strings*), a pesar de que esta instancia ("Población Canals") está fuera del dominio hidrográfico.

Tabla 7.11. Ejemplo de información de otro dominio vinculada al dominio hidrográfico

```
<rdf:Description rdf:about="http://www.oeg-upm.net/ontologies/hydrOntology
.owl#Canals">
   <rdf:type rdf:resource="http://www.oeg-upm.net/ontologies/hydrOntology
.owl#Canal"/>
   <j.0:idioma>cat</j.0:idioma>
   <j.1:nombre>Población 2Canals</j.1:nombre>
   <j.0:long.>-0º 35'</j.0:long.>
   <j.0:lat.>38º 57'</j.0:lat.>
</rdf:Description>
```

7.4.2. Problemas solventados

Con objeto de solventar los problemas descritos con anterioridad, en el contexto de los conjuntos de recursos RDF generados se lleva a cabo una tarea de depuración de la codificación de los *mappings* establecidos. En esta depuración se solucionan problemas creados durante el proceso de generación de *mappings* entre la ontología y las diversas bases de datos y, también, se salvan problemas cuyo origen es la heterogeneidad existente en las fuentes de información estudiadas. A continuación se describen las diferentes soluciones propuestas a los problemas presentados.

Conjuntos de datos RDF

Dentro de los conjuntos de datos RDF se realizaron desde leves modificaciones en la codificación de los *mappings* generados (problemas simples) hasta la utilización de ciertas primitivas (operaciones) del lenguaje R_2O (problemas complejos). De esta manera, se alcanza la expresividad necesaria para solucionar los problemas encontrados.

Las modificaciones relativas a los **problemas leves** están vinculadas con la codificación presente en la sección 1 de la Tabla 7.1. A continuación se concretan las modificaciones realizadas:

- **Tipo fenómeno**. Para las fuentes de información que contienen en el nombre de la instancia el tipo de fenómeno, se procede a eliminar de la codificación la tipología del fenómeno procedente de la ontología. Un ejemplo de esta modificación se presenta en la sección 1 de la Tabla 7.12.

- **Duplicados y Registros sin información**. La solución a estos problemas se centra en dejar de crear URI conformes a los ID de las fuentes de datos y proceder a generarlos conformes a los nombres de las instancias. De esta manera, se consiguen eliminar los duplicados de instancias dentro de un conjunto de tripletas RDF, consecuencia de la generación de auténticos identificadores únicos de recursos (instancias). Asimismo, se consigue eliminar la presencia de tripletas sin información procedente de los registros sin información. Un ejemplo de la sustitución del identificador de la fuente por el campo donde se recoge el nombre de la instancia se muestran en la sección 2 de la Tabla 7.12.

Tabla 7.12. Ejemplo de modificación de la codificación de los *mappings* R_2O+

```
<conceptmap-def name="http://www.oeg-upm.net/ontologies/hydrOntology.owl#
Embalse">
   <uri-as>
     <operation oper-id="concat">
       <arg-restriction on-param="string1">
        <has-value>http://www.oeg-
       upm.net/ontologies/hydrOntology.owl/</has-value>
                    </arg-restriction>
       <arg-restriction on-param="string2">
        <has-column>ngce.nomenclatorgce.Nombre</has-column>
         </arg-restriction>
     </operation>
   </uri-as>
```

Para solucionar los **problemas complejos** se utilizan ciertas operaciones del lenguaje R_2O destinadas a enriquecer la expresividad de las operaciones de correspondencia entre la ontología y las bases de datos.

- **Inclusión y No inclusión**. Estas problemáticas se solucionan con la inclusión del operador `"match-regexp"` en la codificación de los *mappings*. Este operador permite probar la conformidad de una cadena de caracteres con respecto a un conjunto de reglas sintetizadas por medio de un metalenguaje llamado "expresión regular". El operador `"match-regexp"` recibe 2 parámetros: 1) la cadena de entrada en la que buscar la expresión regular, y 2) la expresión regular a buscar. Un ejemplo de la utilización de esta condición para el establecimiento de *mappings* con el fenómeno "Embalse" se muestra en la Tabla 7.13.

Tabla 7.13. Ejemplo de condición "match-regexp"

```
<condition oper-id="match-regexp">
        <arg-restriction on-param="string"><has-
    column>ngce.nomenclatorgce.Nombre</has-column></arg-restriction>
              <arg-restriction on-param="regexp"><has-value>Embalse</has-
          value></arg-restriction>
</condition>
```

- **Conectores**. Para solventar esta problemática (instancias con y sin conectores) se especifican condiciones adicionales dentro de la operación `"match-regexp"`. A continuación, a modo de ejemplo, en la Tabla 7.14 se muestra un extracto del código utilizado para solventar este problema presente en la extracción del fenómeno "Laguna" en la base de datos de la CHE.

Tabla 7.14. Ejemplo de codificación para solventar los problemas de conectores

```
<condition oper-id="match-regexp">
        <arg-restriction on-param="string"><has-
    column>dma.MASALAGO.DENOMINACION</has-column></arg-restriction>
         <arg-restriction on-param="regexp"><has-value>((Laguna
         de)|(Laguna))(.*)</has-value></arg-restriction>
</condition>
```

- **Valor de un campo**. Este problema se soluciona con la utilización de la condición `"equals"` dentro de la codificación de los *mappings*. Esta condición permite comparar si dos cadenas son iguales. Este operador recibe dos parámetros de entrada, constituidos por las cadenas (*string*) a comparar. Este operador retorna *true* si las cadenas son iguales y *false* si no lo son. La Tabla 7.15 presenta un ejemplo de la codificación añadida para el establecimiento de *mappings* considerando un valor (*string*) de un campo de una determinada columna de una tabla de la base de datos.

Tabla 7.15. Ejemplo de codificación para solventar los problemas de la ausencia del tipo de fenómeno

```
<condition oper-id="equals">
        <arg-restriction on-param="value1"><has-
    column>ngce.nomenclatorgce.Grupo</has-column></arg-restriction>
             <arg-restriction on-param="value2"><has-value>HIDRO</has-
          value></arg-restriction>
          </condition>
```

- **Valor de un campo + inclusión**. Este problema requiere la combinación de las condiciones `"match-regexp"` y `"equals"` dentro de la expresión `<applies-if>` (que describe declarativamente el conjunto de instancias de la base de datos que darán lugar a sus equivalentes en la ontología) en la codificación de los *mappings*. Un ejemplo de ésta combinación aparece en la Tabla 7.16.

Tabla 7.16. Ejemplo de combinación de las condiciones "match-regexp" y "equals"

```
<applies-if-top>
        <and>
         <condition oper-id="equals">
             <arg-restriction on-param="value1"><has-column>
           ngce.nomenclatorgce.Grupo</has-column></arg-restriction>
             <arg-restriction on-param="value2"><has-value>HIDRO</has-
           value>
           </arg-restriction>
          </condition>
         <condition oper-id="match-regexp">
             <arg-restriction on-param="string"><has-
           column>ngce.nomenclatorgce.Nombre
           </has-column></arg-restriction>
             <arg-restriction on-param="regexp"><has-value>Río</has-value>
           </arg-restriction>
          </condition>
        </and>
 </applies-if-top>
```

Bases de datos

En esta sección se listan las soluciones propuestas, en el proceso de generación de *mappings* R_2O, a diversos problemas originados por la heterogeneidad existente en las bases de datos estudiadas. Estos problemas están vinculados con las diferencias léxico-sintácticas, la ausencia de tipología de fenómenos y la presencia de nombres oficiales y alternativos. Una descripción detallada del conjunto de problemas existentes en las fuentes de información se recoge en el Capítulo 8 (sección 8.2) de este trabajo.

- **Diferencias léxico-sintácticas**. Los problemas derivados de las diferencias léxico-sintácticas entre las instancias de las múltiples bases de datos se salvan utilizando la condición `"match-regexp"`. Entre los valores de esta condición se añaden las diferentes formas (acentuación, mayúsculas y/o minúsculas) que un determinado fenómeno presenta en una base de datos concreta. De esta manera, se consigue evitar la pérdida de información en la generación de tripletas RDF. Un ejemplo de la codificación utilizada para solventar este problema aparece en la Tabla 7.17.

Tabla 7.17. Ejemplo de uso de la condición "match-regexp" para solventar problemas léxico-sintácticos

```
<condition oper-id="match-regexp">
                  <arg-restriction on-param="string"><has-
         column>DMA.segmentorio.
         DENOMINACION</has-column></arg-restriction>
                 <arg-restriction on-param="regexp"><has-value>((RIO)|(Río)
           |(Rio)) (.*)
           </has-value></arg-restriction>
</condition>
```

- **Ausencia de tipología**. Para las fuentes de información que carecen de la descripción del tipo de fenómeno dentro del nombre de la instancia, la generación de *mappings* R_2O permite la inclusión del tipo de fenómeno dentro de la URI de la instancia. Esta inclusión se produce automáticamente al establecer la correspondencia entre una determinada clase de la ontología y un campo (columna) de la base de datos. La Tabla 7.18 presenta un ejemplo de inclusión del tipo de fenómeno en la URI de la instancia.

Tabla 7.18. Ejemplo de inclusión del tipo de fenómeno

```
<conceptmap-def name="http://www.oeg-upm.net/ontologies/hydrOntology.owl#Río">
   <uri-as>
      <operation oper-id="concat">
        <arg-restriction on-param="string1">
          <has-value>http://www.oeg-upm.net/ontologies/hydrOntology.owl/Río
          </has-value>
        </arg-restriction>
        <arg-restriction on-param="string2">
          <has-column>IGN.ANE_RIOS.RIO</has-column>
        </arg-restriction>
      </operation>
   </uri-as>
</conceptmap-def>
```

- **Nombres oficiales/alternativos**. La utilización del atributo "nombre alternativo", presente en la ontología, permite asociar a los nombres oficiales las variantes existentes dentro de las fuentes de información. En la codificación de los *mappings* esto se traduce en la inclusión de un elemento `<attributemap-def>` para vincular el mencionado atributo a las respectivas clases de la ontología. Un ejemplo aparece en la Tabla 7.19.

Tabla 7.19. Ejemplo de elemento "attributemap-def" para la inclusión de nombres alternativos

```
<attributemap-def name="http://www.oeg-upm.net/ontologies/hydrOntology.owl#
nombre_alternativo">
           <selector>
                <aftertransform>
                     <operation oper-id="constant">
                      <arg-restriction on-param="const-val">
                        <has-column>ngce.nomenclatorgce.Variante</has-column>
                      </arg-restriction>
                     </operation>
                </aftertransform>
           </selector>
</attributemap-def>
```

- **Espacios en blanco**. La presencia de espacios en blanco en los nombres de las instancias genera problemas para la creación de URI. Estos espacios, procedentes de las fuentes originarias, se ha eliminado a través de un post-procesamiento de los conjuntos de datos RDF resultantes. En la Tabla 7.20 se recoge un ejemplo donde se muestra un recurso del NGCE con y sin espacio en blanco.

Tabla 7.20. Recursos RDF con y sin espacios en blanco

```
<rdf:Description rdf:about="http://www.oeg-upm.net/ontologies/hydrOntology
.owl#Salado, Río ">

<rdf:Description rdf:about="http://www.oeg-upm.net/ontologies/hydrOntology
.owl#Salado,_Río">
```

Generación de RDF

En esta sección se muestra un extracto de los resultados obtenidos tras las mejoras aplicadas en la codificación de los *mappings* R_2O generados (ver Tabla 7.21). La representación de los resultados se realiza conforme a la anotación N3 para su mejor visualización.

En la tabla anterior se recogen diversas instancias y sus respectivos atributos generados del proceso descrito con anterioridad. Por ejemplo, la instancia "`Judío, Bodón del`", pertenece al tipo de fenómeno "`Bodón`" (`Bod%C3%B3n`). El idioma de dicha instancia es español ("`es`") y su posición espacial está fijada en "`41.1235708903828`" Latitud y -4.47315019507167" Longitud.

Los conjuntos de información RDF generados de las distintas fuentes de información tratadas están disponibles para consulta en la Web. Asimismo, la calidad de estos datos se trata en la evaluación realizada en el Capítulo 9 de este trabajo.

7.5. Publicación de datos integrados

En esta sección se describe el proceso de publicación (Vilches-Blázquez *et al.*, 2010a) de los conjuntos de datos RDF conforme a los principios de *Linked Data*. Este proceso se realiza acorde con las diversas tareas descritas en la Actividad 5 de la metodología propuesta, cuyos resultados finales se presentan en el sitio web de GeoLinked Data[7], donde, además, se publica toda la información relacionada con la propuesta de integración realizada en este trabajo (ver Figura 7.9).

En este sitio web, para la publicación de datos RDF se utiliza como *triple store* a Virtuoso Universal Server[8], un *middleware* y motor de bases de datos, que combina la funcionalidad de un Sistema Gestor de Bases de Datos tradicional, una base de datos virtual, RDF, XML, texto libre, servidor de aplicaciones web y *Linked Data*, y la funcionalidad de servidor de ficheros.

De este componente se produce la generación de SPARQL Endpoints, es decir, servicios conformes al protocolo SPARQL (Seaborne *et al.,* 2008). Estos servicios permiten al usuario consultar la base de conocimiento (conjuntos de datos almacenados en el repositorio RDF) a través del lenguaje de consulta SPARQL. En el sistema desarrollado los SPARQL Endpoints desplegados se ofrecen a los usuarios para que puedan ser consultados directamente en la web.

[7] http://geo.linkeddata.es/
[8] http://virtuoso.openlinksw.com/

Tabla 7.21. Extracto de tripletas RDF procedente del Nomenclátor Geográfico Conciso de España

```
     @prefix : <http://www.w3.org/2003/01/geo/wgs84_pos#> .
     @prefix j.0: <http://geo.linkeddata.es/ontology/> .
     @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
     @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://geo.linkeddata.es/resource/Arroyo/Ca%C3%B1er%C3%ADa%2C%20Arroy
o%20de%20la>     a j.0:Arroyo;
         rdfs:label " Cañería, Arroyo de la"@es;
          :geometry <http://geo.linkeddata.es/resource/wgs84/37.4221216
          387048_-6.37676820605391> .


<http://geo.linkeddata.es/resource/Bod%25C3%25B3n/Jud%C3%ADo%2C%20Bod%
C3%B3n%20del>     a j.0:Bod%C3%B3n;
         rdfs:label " Judío, Bodón del"@es;
          :geometry <http://geo.linkeddata.es/resource/wgs84/41.1235708
          903828_-4.47315019507167> .


<http://geo.linkeddata.es/resource/Playa/Cala%20Gogo%2C%20Platja%20de>
a j.0:Playa;
         rdfs:label "Cala Gogo, Platja de"@ca;
          :geometry <http://geo.linkeddata.es/resource/wgs84/41.8277767
          103577_3.0842923046658> .

    <http://geo.linkeddata.es/resource/Playa/Perchel%2C%20Playa%20del>
a j.0:Playa;
         rdfs:label "Perchel, Playa del"@es;
          :geometry <http://geo.linkeddata.es/resource/wgs84/28.5821238
          203424_-13.8288504290811> .

   <http://geo.linkeddata.es/resource/wgs84/28.5821238203424_-13.8288
   504290811>     a :Point;
         :lat 28.5821238203424;
         :long -13.8288504290811 .

   <http://geo.linkeddata.es/resource/wgs84/37.4221216387048_-6.3767
   6820605391>     a :Point;
         :lat 37.4221216387048;
         :long -6.37676820605391 .

   <http://geo.linkeddata.es/resource/wgs84/41.1235708903828_-4.4731
   5019507167>     a :Point;
         :lat 41.1235708903828;
         :long -4.47315019507167 .

   <http://geo.linkeddata.es/resource/wgs84/41.8277767103577_3.0842923
   046658>     a :Point;
         :lat 41.8277767103577;
         :long 3.0842923046658 .
```

Figura 7.9. Portal web de GeoLinked Data

Asimismo, para la visualización de los datos también se utiliza Pubby[9] (tarea 5.3) (ver Figura 7.10), un *FrontEnd* de *Linked Data* para SPARQL Endpoints que está diseñado para proporcionar un interfaz para la visualización y navegación de los conjuntos de datos RDF. Asimismo, Pubby consigue que las URI de los conjuntos de datos RDF sean URI *deferenciables*, permitiendo que las instancias (URI) del SPARQL Endpoint de GeoLinked Data sean accesibles por un navegador de la Web Semántica, también maneja diversos detalles de la interacción HTTP, tales como la redirección 303 requerida por la Arquitectura Web y la negociación de contenido entre HTML, RDF/XML y las descripciones N3 de los mismos recursos. En la Figura 7.11 se presenta un ejemplo de la visualización de los recursos RDF ofrecida por Pubby. En la visualización de la información ofrecida por Pubby también se han añadido diversos elementos de metadatos (tarea 5.1) que incluyen información sobre el tipo de licencia de los datos, información sobre la procedencia de la información, cobertura espacial, etc. Para ello, se han utilizado algunos elementos de metadatos procedentes de Dublin Core y voID.

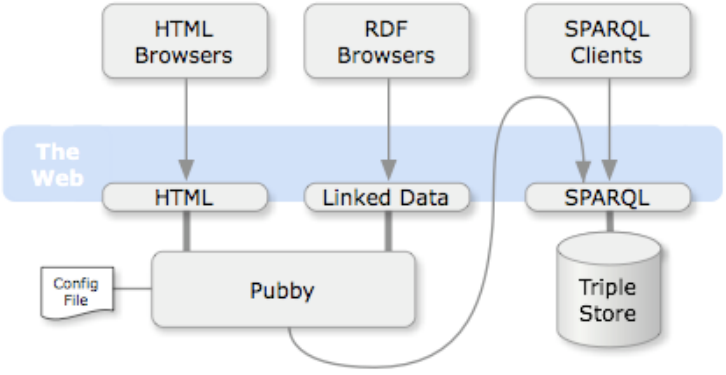


Figura 7.10. Funcionamiento de Pubby

[9] http://www4.wiwiss.fu-berlin.de/pubby/

Figura 7.11. Visualización de recursos con Pubby

Para la visualización gráfica (Vilches-Blázquez *et al.*, 2010b) de la información se ha desarrollado una aplicación web[10] (tarea 5.3) para enriquecer la visualización del RDF generado. Esta interfaz combina el paradigma de la navegación por facetas (Oren *et al.*, 2006) con la visualización cartográfica soportada en la utilización de Google Maps API. Así, por ejemplo, la aplicación permite la representación de distintas características geométricas para los fenómenos hidrográficos recogidos en los diversos conjuntos de datos RDF, tales como, la utilización puntos para mostrar la información sobre los embalses de España (ver Figura 7.12) o *LineStrings* para la representación de "Ríos" (ver Figura 7.13). La representación geométrica de estos fenómenos resulta dependiente de la información geométrica recogida en la fuente original.

Figura 7.12. Visualización de los embalses de España

[10] http://geo.linkeddata.es/web/guest/visualizacion-beta

Finalmente, se ha implementado un navegador de facetas (*faceted browser*) para GeoLinked Data (ver Figura 7.13). Este navegador es un ejemplo de interfaz exploratorio de búsqueda (*Exploratory Search Interface*), cuyo diseño ha sido investigado en algunas recientes investigaciones del área de Interfaces Persona-Ordenador (*Human Computer Interaction* - HCI) para dar soporte a los usuarios con necesidades más complejas (Wilson *et al.*, 2007).

Figura 7.13. Visualización del fenómeno hidrográfico "Río" como *LineStrings*

7.6. Conclusiones

En este capítulo se ha descrito el proceso de generación de información RDF para el dominio hidrográfico. Este trabajo se desarrolla consecuencia de la ausencia de una descripción de la hidrografía del ámbito nacional español con información proveniente de productores oficiales en el contexto de la Web de los Datos.

En primer lugar, se presenta el **patrón** generado para la identificación de los datos a través de **URI**, así como las principales decisiones de diseño adoptadas para la identificación de los datos.

Asimismo, se describe el proceso de **generación de *mappings* entre bases de datos y ontología**. Los recursos de información utilizados son *hydrOntology*, como ontología de referencia, y múltiples bases de datos con información sobre el dominio a nivel europeo, nacional y regional. En este proceso se utiliza el lenguaje R_2O+, como lenguaje de *mappings*.

Posteriormente, se detalla el **proceso para la transformación de la información geoespacial a RDF**. En este proceso de generación de instancias se utiliza ODEMaspter++, a través del plugin ODEMapster de la herramienta NeOn Toolkit, y la librería GEOMETRY2RDF.

Con el objetivo de comprobar la calidad de los resultados generados, se seleccionan y analizan unas muestras de instancias de los conjuntos de datos resultantes. De este análisis se detecta una serie de problemas, descritos en la sección 7.4.1. Estos problemas se clasifican en dos niveles, en base a su complejidad. Para solventar estos problemas se realiza un **proceso de depuración de los *mappings*** establecidos. Este proceso conlleva desde simples modificaciones hasta la utilización de ciertas primitivas del lenguaje R_2O. Estas modificaciones solucionan los diversos problemas detectados en el RDF generado así como algunos provenientes de las bases de datos (sección 7.4.2). Asimismo, se muestran algunos ejemplos del RDF materializado tras las mejoras aplicadas en la codificación. La evaluación de los conjuntos de datos RDF resultantes se presenta en el Capítulo 9 de este trabajo.

Además, se presenta una visión general de las tareas realizadas (Actividad 5) para la **publicación de conjuntos de datos integrados en formato RDF y conforme a los principios de *Linked Data***.

Los principales logros del trabajo descrito en este capítulo son: 1) la **generación de siete conjuntos de datos hidrográficos en RDF**, 2) el **tratamiento e inclusión de las características geométricas** asociadas a la información geoespacial a través de la incorporación de GML en el RDF generado y 3) la **publicación** de la información hidrográfica (conjuntos de datos RDF) según los principios de ***Linked Data***. Asimismo, la información publicada ha contribuido al **enriquecimiento de la Web de los Datos** con información geoespacial en español y ha servido para poner en marcha la **iniciativa GeoLinked Data**.

Capítulo 8

Detección de instancias duplicadas en conjuntos de datos heterogéneos

Como se describe en el Capítulo 3 (sección 3.3), la detección de duplicados es un área de investigación ampliamente desarrollada, incluso en diferentes dominios de aplicación, donde este problema es conocido con diferentes nombres conforme a los requerimientos específicos que necesitan ser satisfechos y a los objetivos perseguidos en cada comunidad. Sin embargo, resultan escasos los trabajos que se centran en el estudio de la detección de duplicados en el proceso de integración de múltiples y heterogéneos conjuntos de datos geoespaciales.

Con el reciente surgimiento de la iniciativa *Linked Data* los problemas para la detección de duplicados aparecen también en el escenario de la Web de los Datos, donde diversos tipos de enlaces semánticos entre conjuntos de datos pueden ser descubiertos, a pesar de que con frecuencia los términos de diferentes vocacularios están mezclados y no se produce una consistencia entre los esquemas (RDF Schema o OWL) de dichos conjuntos de datos (Volz *et al.*, 2009). En este nuevo contexto las propuestas existentes para la detección de duplicados resultan aún más escasas y limitadas para su aplicación a conjuntos de datos con características propias de la infomación geoespacial.

Este capítulo describe las técnicas desarrolladas para la detección de instancias duplicadas en un mismo y entre distintos conjuntos de datos RDF heterogéneos con información hidrográfica. Este proceso se realiza conforme a las tareas especificadas en la Actividad 6 de la metodología de integración de datos geográficos propuesta. Así, el capítulo se estructura de la siguiente manera: En la sección 8.1 se recoge algunas de las características principales de la duplicidad de instancias. La sección 8.2 analiza y clasifica los principales problemas presentes en el contexto de las instancias de la información geoespacial. Distintas enfoques heurísticos para detección de instancias duplicadas en un mismo y entre múltiples conjuntos de datos se describen en la sección 8.3. Finalmente, la sección 8.4 recoge las conclusiones al trabajo descrito en este capítulo.

8.1. Duplicidad de instancias

Uno de los mayores problemas de la información geoespacial es la diversidad de los conjuntos de datos (qué información es almacenada, cómo es representada y estructurada, qué calidad tiene, a qué datos se refiere, cuál es la escala de recogida de información, etc.) (Devogele *et al.*, 1998). Sin embargo, según Menestrina *et al.* (2006), cuando los datos de diversas fuentes se integran resultan frecuentes los solapamientos de información, debido a que múltiples registros se refieren a la misma entidad del mundo real y estos pueden ser abstraídos

en diferentes repositorios. Asimismo, cuantos más conjuntos de datos formen parte del proceso de integración mayor es el potencial de solapamiento de la información.

Los solapamientos entre diferentes fuentes también afectan a la descripción de los datos. En ocasiones, las múltiples representaciones normalmente no son idénticas, ya que pueden existir diferencias, por ejemplo: errores ortográficos, alteraciones o ausencia de valores, etc. Por tanto, resulta frecuente encontrar descripciones similares para diferentes fenómenos (ver Tabla 8.1) y descripciones diferentes para un mismo fenómeno (ver Tabla 8.2).

Tabla 8.1. Descripciones similares para diferentes fenómenos

Río Salado en Navarra (afluente del río Arga)
Río Salado en Andalucía
Río Salado en Castilla y León
Río Salado en Guadalajara
Río Salado, en Guadalajara (también llamado Río Linares)
Río Salado en La Rioja (afluente del río Jubera)
Río Salado de Espera en Cádiz
Río Salado de Rota en Cádiz

En el contexto de *Linked Data* con la explosión del número de fuentes de información generadas también se ha comenzado a afrontar problemas estrechamente relacionados con aquellos presentes en la integración de información tradicional. En este nuevo contexto los problemas para la detección de duplicados provienen de la utilización de diversas URI para identificar a un mismo recurso (ver Tabla 8.2). Asimismo, resulta frecuente que datos en el contexto de la Web de los Datos presenten información conforme a entidades idénticas, pero con diferentes identificadores, lo que provoca dificultades de relación (*linkage*) y de integración de información. Un ejemplo de esta situación se refleja en (Jaffri *et al.*, 2007), donde fuentes como DBpedia, GeoNames, CIA Factbook y Eurostat poseen diferentes URI para el mismo país.

Tabla 8.2. Descripciones diferentes para un mismo fenómeno

`http://dbpedia.org/resource/Spain`
`http://www4.wiwiss.fuberlin.`
`de/factbook/resource/Spain`
`http://sws.geonames.org/2510769/`
`http://www.daml.org/2001/09/countries/fips#SP`
`http://www4.wiwiss.fuberlin.de/eurostat/resource/countries/Espa%C3%B1a`

Según Jaffri *et al.* (2008) la multiplicidad de URI conduce al problema de la co-referencia. La co-referencia es el problema de asegurar que dos entidades diferentes no comparten el mismo nombre o identificador pero, en cambio, se refieren a la misma entidad. Este problema puede ocurrir en dos formas: Por un lado, cuando una única URI identifica más de un recurso, por ejemplo, existen diversos topónimos con el mismo nombre pero estos topónimos son localizados en diferentes lugares. Por otro, cuando múltiples URI identifican el mismo recurso (por ejemplo, la instancia "España" tiene diferentes URI que dependen de la fuente de procedencia; así, en DBpedia aparece como `http://dbpedia.org/resource/Spain`, mientras en GeoNames se identifica como `http://sws.geonames.org/2510769`). En la Web Semántica Geoespacial (Egenhofer, 2002) esto presenta un problema cuando existe la necesidad de relacionar conocimiento compilado dentro de diversas fuentes de información (bases de datos) procedentes de proveedores dispares.

8.2. Problemas en el ámbito de las instancias

La comparación de los datos (tarea 6.1), recogidos en las fuentes geoespaciales consideradas (ver Anexo), pone de manifiesto la existencia de diferentes problemas en el proceso de integración de información. Como se refleja en (Vilches-Blázquez *et al.*, 2009), estos problemas son clasificados conforme al enfoque basado en niveles o capas (léxica, sintáctica, semántica y pragmática) distinguido en (Euzenat, 2001; Corcho, 2005). Este enfoque es común en trabajos que tratan con interoperabilidad semántica. Los diferentes problemas recogidos se han clasificado, acorde a sus características vinculadas o desvinculadas del dominio, en independientes del dominio y dependientes del dominio. Dentro de los problemas dependientes del dominio la clasificación se subdivide en problemas vinculados al dominio geográfico (problemas generales) y/o relacionados con el dominio hidrográfico (problemas particulares). Asimismo, estos problemas son clasificados conforme a diferencias internas (una base de datos) y diferencias inter-fuente (múltiples bases de datos), considerando similitud y diferencias de escala. A continuación se presentan los diferentes problemas de duplicación de instancias dentro de las fuentes de datos geoespaciales consideradas.

8.2.1. Nivel léxico

Esta capa trata con la capacidad para segmentar la representación en caracteres y palabras (o símbolos) (Euzenat, 2001). En esta capa, se han observado diversos tipos de problemas en diferentes niveles de análisis.

Una única base de datos. En el ámbito de una única base de datos los problemas identificados son los siguientes:

- **Acentuación**. Este problema está relacionado con los caracteres propios de la lengua española. En este caso, la diferencia entre intancias es motivada por la presencia o ausencia de acentuación ortográfica. En el contexto de este problema aparecen instancias como por ejemplo, “Río Tajo” y “Rio Tajo”, presentes en la base de datos de EuroRegionalMap.
- **Caracteres especiales**. Al igual que ocurre con el problema de acentuación, este problema está vinculado a las características de la lengua española. Así la presencia de caracteres como “ñ”, “ç”, “ º ”, “ ª ”, etc., son frecuentes en las múltiples instancias de las diversas bases de datos. Por ejemplo, las instancias “Caño de Guadiamar” y “Cano de Guadiamar” aparecen en la base de datos de EuroRegionalMap.
- **Abreviaturas**. La presencia de representaciones gráficas reducidas de una palabra mediante supresión de letras finales o centrales y que suele cerrarse con un punto está presente en el contexto de las bases de datos tratadas. Un ejemplo de esta problemática es “S. Lorenzo”, referida a la instancia “Embalse de San Lorenzo”, presente en la base de datos de la CHE.
- **Espacios en blanco**. Resulta frecuente la presencia de espacios en blanco al final de los nombres de las instancias. Un ejemplo de la inclusión de estos espacios finales es la instancia “Embalse de Cabrerón ", presente en BTN25.

Múltiples bases de datos. En el contexto de varias bases de datos los problemas recogidos son resultado del proceso de comparación de diversas bases de datos con idéntica o desigual escala.

- **Mayúsculas y minúsculas**. La diferenciación en el uso de letras mayúsculas y minúsculas se hace patente, por ejemplo, entre las bases de datos de la CHE y el Nomenclátor Conciso. Este problema aparece al analizar una o varias bases de datos a la misma o diferentes escalas.

Tabla 8.3. Diferenciación de caracteres (mayúsculas y minúsculas) entre dos bases de datos

Nomenclátor Conciso	CHE
Alhama, Río	ALHAMA
Alcanadre, Río	ALCANADRE
Arga, Río	ARGA

Los problemas descritos en el análisis de las bases de datos de forma individualizada (acentuación, caracteres especiales y abreviaturas) repercuten especialmente en el proceso de integración del conjunto de fuentes estudiadas. A modo de ejemplo, a continuación, se describe el problema de las abreviaturas en la integración de diversas fuentes.

- **Abreviaturas**. La presencia de abreviaturas se identifica como un problema al integrar información de diferentes bases de datos, por ejemplo en el caso de la información perteneciente a la CHE y el Nomenclátor Conciso. La información relacionada con la instancia "Santa Ana", en la primera base de datos mencionada, aparece en la tabla EMBALSE nombrada como "STA. ANA", mientras en la referida al Nomenclátor Conciso la instancia es nombrada como "Santa Ana, Embalse de".
- **Espacios en blanco**. La existencia de instancias con y sin espacios en blanco al final de los nombres de las mismas es frecuente entre las múltiples intancias del conjunto de bases de datos estudiadas. Este elemento puede resultar un problema por el ruido que genera en el proceso de comparación de las instancias.

El conjunto de problemas detectados en el nivel léxico pertenecen al dominio general, por tanto, son frecuentes en cualquier base de datos y son clasificados como independientes del dominio.

8.2.2. Nivel sintáctico

Este nivel trata con la capacidad para estructurar la representación en las frases estructuradas, formulas o afirmaciones (Euzenat, 2001). Los problemas recogidos en este nivel se caracterizan por estar ubicados entre los límites de los niveles léxico y sintáctico. A continuación se describen diversos tipos de problemas en diferentes niveles de análisis.

Una única base de datos. A continuación, se exponen brevemente diferentes problemas comunes que aparecen en las diferentes bases de datos objeto de estudio.

- **Errores tipográficos**. Este tipo de error está relacionado con los errores de codificación o grabación recogidos en (Chatterjee *et al*., 1995). En el contexto de estudio aparecen algunos errores tipográficos, por ejemplo, en la base de datos de BCN200. Un listado no exhaustivo de estos errores se muestra en la Tabla 8.4.
- **Proximidad del nombre** (*distancia de edición*). Este problema se caracteriza por la similitud o proximidad entre dos nombres de topónimos diferentes, es decir, similitud entre diferentes cadenas (*strings*) (por ejemplo, las instancias de los ríos "Arga" y "Arfa" presentes en la base de datos de la CHE o "Perrera, Arroyo de la" y "Perreras, Arroyo de las" presentes en la base de datos del Nomenclátor Conciso).
- **Orden de nombrado**. Los cambios de orden o alteraciones en nombres de instancias son frecuentes cuando se analizan diversas bases de datos de manera individualizada. La causa de este problema proviene de escribir nombres de instancias de manera diferente, siendo las formas habituales [nombres de instancia, tipo de fenómeno] o [tipo de fenómeno + nombre de instancia]. Así, en la información asociada al Nomenclátor NOMGEO se encuentran indistintamente ejemplos como "Castaño, Fuente del" y "Fuente Amarguilla".

Tabla 8.4. Lista de errores tipográficos y equivalencias en la base de datos de BCN200

CADENA_ORIGINAL	Equivalencia
S?quia	Acequia
do	de
Lagoa	Lago
Encoro	Embalse
Aigua	Agua
R?u	Riu
X?quer	Júcar
da	de
das	de
Assut	Presa
Bra?	Brazo
Conca	Cuenca
Curs	Curso
Doll	Chorro
els	los
Embassament	Embalse
Engorjat	Garganta
Espona	Margen
Gelera	Nevero
Glacera	Glaciar
Gual	Vado
Llac	Lago
Llacuna	Laguna
Pant?	Embalse
Rebals	Embalse
Rial	Arroyuelo
Riera	Arroyo
Riu	Río
Segla	Acequia
Torrent	Torrente
Arroio	Arroyo
Cal	Canal
Leito	Cauce
Quenlla	Acequia
Rega	Riego
Ribeira	Ribera
Vao	Vado
Veiga	Vega
Barranc	Barranco

Múltiples bases de datos. Desde una perspectiva general, cabe resaltar que los problemas mencionados en el contexto individual de una fuente dificultan el proceso de integración del conjunto de la información desde la perspectiva sintáctica de los datos. A continuación, a modo de ejemplo, se describen los siguientes problemas:

- **Orden de nombrado**. Los cambios de orden o alteraciones en nombres de instancias es un problema presente en el proceso de integración de múltiples fuentes de información. Diversos muestras de este problema aparecen en las bases de datos estudiadas, como por ejemplo: "Mequinenza, Embalse de"/"Embalse de Mequinenza" o "Fondo, Estany"/ "Estany Fondo".
- **Proximidad del nombre** (*distancia de edición*). La similitud en el nombre de instancia entre múltiples conjuntos de información es un problema recurrente. Así, en las fuentes estudiadas se puede encontrar "Ladra, Río" y "Adra, Río" pertenecientes a la base de datos del NOMGEO e ICA, respectivamente.

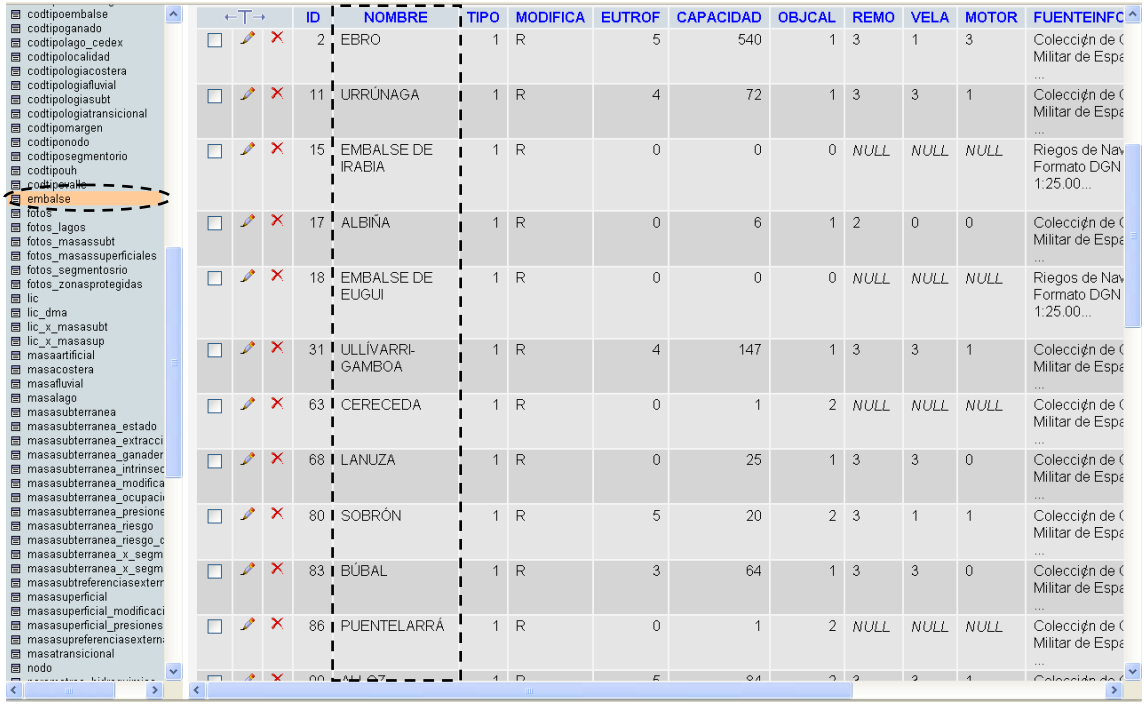
Los problemas detectados en el nivel sintáctico, descritos en el ámbito de una y múltiples fuentes, pertenecen al dominio general. Esto les hace frecuentes en cualquier base de datos y, por este motivo, son clasificados como independientes del dominio.

8.2.3. Nivel semántico

Este nivel trata con la capacidad de construir el significado proposicional de la representación (Euzenat, 2001). En este nivel se listan los siguientes problemas:

Una única base de datos. En el contexto de una única base de datos los problemas identificados dentro del nivel semántico son:

- **Diferentes instancias en un mismo campo.** Resulta relativamente frecuente encontrar más de un nombre de instancia dentro de un mismo registro de la base de datos. De esta manera, se almacena en un registro los nombres oficiales y alternativos de un determinado fenómeno geográfico. Así, por ejemplo, en la base de datos de BTN25 aparece en un mismo campo "RIA DO RIBADEO / RIA DO EO" y en la base de datos de Waterbase, concretamente en EEA_Water base river, aparecen "RIO CUBILLAS EMBALSE DE CUBILLAS" o "RIO GENIL EMBALSE DE IZNAJAR".
- **Ausencia de tipología de fenómenos.** En ocasiones, en algunas de las bases de datos estudiadas aparecen topónimos sin referencia a la tipología de fenómeno. Este problema se deriva de las diferentes formas de escribir la información, tales como [tipo de fenómeno + nombre de instancia] o [nombre de instancia]. Así, en el Nomenclátor Conciso aparecen ejemplos de esta problemática. Se encuentran instancias como "Besòs, el" o "Freser, el" en cuyo nombre no aparece la tipología de fenómeno asociada. La ausencia de la tipología de fenómenos, en ocasiones, se ve solventada por la presencia de:
    - Una tipología de fenómenos determinada como nombre de una tabla de la bases de datos. Un ejemplo de esta situación aparece en la base de datos de la CHE (ver Figura 8.1).



ID	NOMBRE	TIPO	MODIFICA	EUTROF	CAPACIDAD	OBJCAL	REMO	VELA	MOTOR	FUENTEINFO
2	EBRO	1	R	5	540	1	3	1	3	Colección de C Militar de Espa ...
11	URRÚNAGA	1	R	4	72	1	3	3	1	Colección de C Militar de Espa ...
15	EMBALSE DE IRABIA	1	R	0	0	0	NULL	NULL	NULL	Riegos de Nav Formato DGN 1:25.00...
17	ALBIÑA	1	R	0	6	1	2	0	0	Colección de C Militar de Espa ...
18	EMBALSE DE EUGUI	1	R	0	0	0	NULL	NULL	NULL	Riegos de Nav Formato DGN 1:25.00...
31	ULLÍVARRI-GAMBOA	1	R	4	147	1	3	3	1	Colección de C Militar de Espa ...
63	CERECEDA	1	R	0	1	2	NULL	NULL	NULL	Colección de C Militar de Espa ...
68	LANUZA	1	R	0	25	1	3	3	0	Colección de C Militar de Espa ...
80	SOBRÓN	1	R	5	20	2	3	1	1	Colección de C Militar de Espa ...
83	BÚBAL	1	R	3	64	1	3	3	0	Colección de C Militar de Espa ...
86	PUENTELARRÁ	1	R	0	1	2	NULL	NULL	NULL	Colección de C Militar de Espa ...

Figura 8.1. Extracto de la tabla "Embalse" de la base de datos de la CHE

- La tipología de fenómenos se contempla en un campo (columna) diferente al del nombre de la instancia. Un ejemplo de esta situación también aparece en la base de datos de la CHE (ver Figura 8.2).

Múltiples bases de datos. En el contexto de integración de las diversas fuentes consideradas aparecen problemas similares a los descritos con anterioridad (granularidad y ausencia de tipología de fenómenos) y problemas particulares (diferentes clasificaciones). A continuación se describen sus características principales.

- **Diferentes clasificaciones**. Este problema se presenta cuando diferentes bases de datos recogen fenómenos similares y éstos son clasificados conforme a diversos puntos de vista. Por ejemplo; "Arroyo de las Rozas" (Dehesa) y "Arroyo de las Rozas" (Arroyo). En el primer caso, la base de datos del Nomenclátor NOMGEO clasifica esta instancia como un lugar, concretamente, como un tipo de paraje (Dehesa). Mientras la segunda instancia, perteneciente a BTN25, se clasifica como un fenómeno hidrográfico (Arroyo). Este problema está relacionado con la heterogeneidad cognitiva contemplada en (Bishr, 1998).
- **Ausencia de tipología de fenómenos.** Los problemas generados por la ausencia de la tipología de fenómenos en el ámbito de una única base de datos repercuten de forma directa en la comparación de instancias presentes en el conjunto de fuentes de datos estudiadas. Fundamentalmente, esto ocurre al comparar diferentes instancias del Nomenclátor Conciso (donde en ocasiones aparecen instancias sin tipología de fenómeno en el nombre del topónimo) con el resto de bases de datos controladas. También se encuentran ejemplos como "Río Ulla" y "Ulla" en la base de datos de EuroRegionalMap y Waterbase, respectivamente.

CDR1	NOMBRE	TIPO	ID	NOMBRERIO2	LONGITUD	FUENTE	APORTACION	SUPCU
90129140602	TRISTANÍ	RIO	1706	NULL	5	NULL	0	
9012710	TRISTE	RIO	1707	NULL	14	NULL	0	
9010503	TRUEBA	RIO	1708	NULL	46.5	E.R.S.	243	
9011613	TUERTO	RIO	1709	NULL	21.8	NULL	0	
9014402	ULLDEMO	RIO	1710	NULL	17.3	E.R.S.	8	
901231602	ULZAMA	RIO	1711	NULL	36	NULL	0	
901292803	UNARRE	RIO	1712	NULL	9	NULL	0	
901231202	URBELCHA	RIO	1713	NULL	4	NULL	0	
9011401	URBION	RIO	1714	NULL	5.5	NULL	0	
9011604	URBION	RIO	1715	NULL	21.2	NULL	0	
901231201	URCHURIA	RIO	1716	NULL	10	NULL	0	
9012103	UREDERRA	RIO	1717	NULL	18.8	E.R.R.	205	
901150201	URQUIOLA	RIO	1718	URQUIOLA	22.3	NULL	0	
901231204	URRIO	RIO	1719	NULL	6	NULL	0	
901231206	URROBI	RIO	1720	NULL	25.5	NULL	0	
901293409	USIA	RIO	1721	NULL	19	NULL	0	
901231002	USTARROZ	RIO	1722	NULL	9.8	NULL	0	
9012004	VADILLOS	RIO	1723	NULL	11.5	NULL	0	
9012601	VAL	BARRANCO	1724	NULL	28	NULL	0	
NULL	VAL DE RECORDIN	ARROYO	1725	NULL	NULL	NULL	NULL	
9012914	VALIRA	RIO	1726	NULL	44	NULL	0	
9012928080201	VALIRI	RIO	1727	NULL	8.7	NULL	0	
9011201	VALLARTA	RIO	1728	NULL	5	NULL	0	
90129280501	VALLFARRERA	RIO	1729	NULL	27	NULL	0	
90129341103	VALLIBIERNA	RIO	1730	NULL	9.5	NULL	0	

Figura 8.2. Extracto de la tabla "Rios" de la base de datos de la CHE

De forma general, los problemas detectados en el nivel semántico presentes tanto en una como en múltiples bases de datos son dependientes del dominio geográfico. De forma particular, la problemática referida a diferentes instancias de fenónemos en un mismo campo (aunque resulta

muy característica en la información geográfica) puede presentarse en otros dominios, por tanto, es clasificada como independiente del dominio.

8.2.4. Nivel pragmático

Este nivel trata con la capacidad de construir el significado pragmático de la representación (o su significado en el contexto) (Corcho, 2005). A continuación se listan los problemas reflejados en este nivel.

Una única base de datos. En el análisis de las características de las diversas fuentes de manera individualizada se presentan los siguientes problemas:

- **Nombres oficiales y alternativos**. La presencia de nombres oficiales y alternativos dentro de una base de datos genera diferentes problemas. Estos pueden ser subdivididos en tres tipos distintos:

    - Nombres oficiales y alternativos pueden ser encontrados en la misma entrada de una tabla de la base de datos. Por ejemplo; "Río Izfalada o de Huéneja" o "Perepuig, Barranc de /Padrepuig, Barranco de" o "Arroyo Somero <->Rio Despeñaperros" pertenecen a las bases de datos del ICA, Nomenclátor NOMGEO y BCN200 respectivamente.

Tabla 8.5. Ejemplo de nombres oficiales y/o alternativos en diferentes entradas de una tabla

ID	Name	Alternative	Length	Coordinate
190	*RÍA DE ORTIGUEIRA*	*Ría de Ortigueira e Ladrido*	2121,3	42,7612972; 42,7686864
307	Río Palancia	Riu Palància	53574,7	43,1791692; 43,1608346

    - Nombres oficiales y alternativos pueden ser encontrados en diferentes entradas de una tabla de la base de datos. Por ejemplo, "RÍA DE ORTIGUEIRA" y "Ría de Ortigueira e Ladrido" en BTN25. Un ejemplo de este tipo de representación de nombres oficiales y alternativos aparece en la Tabla 8.5.

Tabla 8.6. Ejemplo de nombres oficiales y/o alternativos en diferentes tablas de una base de datos

ID	Name	Province	Latitude	Longitude
32	Río Turia	Castellón	39° 27'	-0° 19'
74	Río Mijares	Valencia	39° 55'	-0° 01'

ID	Name	Slope	Basin	Language
11	Túria, Riu	ATLÁNTICA	JÚCAR	cat
15	Millars, Riu	ATLÁNTICA	JÚCAR	cat

    - Nombres oficiales y alternativos que pueden ser encontrados en diferentes tablas de una base de datos. Un ejemplo de esta problemática aparece en la Tabla 8.6.

- **Topología**[1] **y Geometría**[2]. Algunas de las fuentes estudiadas presentan duplicidad de instancias. Esta duplicidad está vinculada a la existencia de diferentes secciones geométricas de un fenómeno geográfico o a la implementación de topología mediante la utilización de herramientas relacionadas con los Sistemas de Información Geográfica (SIG). Un ejemplo de este problema se muestra en la Tabla 8.7. En relación a los problemas de heterogeneidad semántica, este tipo de duplicidad está relacionada con los problemas de codificación o grabación recogidos en (Chatterjee *et al*., 1995).
- **Información incompleta**. Existen diferentes entradas en una base de datos cuyo valor resulta desconocido. Ejemplos de esta incompletitud son los campos LONGITUD, PERIMETRO y/o SUPERFICIE de la base de datos de BTN25 o la base de datos de EuroRegionalMap (Ver Tabla 8.8).

Tabla 8.7. Duplicidad de información presente en un extracto de la base de datos de EGM

OBJECTID	Shape	FCsubtype	gfid	F_CODE	EXS	HOC	HYC	NHI	NAMN1	NAMN2	NAMA1	NAMA2	NLN1	NLN2	Shape_Length	Shape_Area
8	[BLOB - 1.3 KB]	1	NULL	BH502	998	5	8	N_P	Río Ulla	N_A	Río Ulla	N_A	SPA	N_A	0.241585222699906	0.000495159995250335
9	[BLOB - 12.2 KB]	1	NULL	BH502	998	5	8	N_P	Río Ebro	N_A	Río Ebro	N_A	SPA	N_A	3.23782443794389	0.00495177613599993
10	[BLOB - 1.2 KB]	1	NULL	BH502	998	5	8	N_P	Río Miño	N_A	Río Mino	N_A	SPA	N_A	0.476555126149795	0.000995917687999499
11	[BLOB - 10.1 KB]	1	NULL	BH502	998	5	8	N_P	Río Duero	N_A	Río Duero	N_A	SPA	N_A	2.85509457099298	0.00432759552000145
12	[BLOB - 2.5 KB]	1	NULL	BH502	998	5	8	N_P	Río Duero	N_A	Río Duero	N_A	SPA	N_A	0.599240259047293	0.000388936496749685
13	[BLOB - 8.7 KB]	1	NULL	BH502	998	5	8	N_P	Río Ebro	N_A	Río Ebro	N_A	SPA	N_A	2.34590910101081	0.00377939952400128
14	[BLOB - 1.2 KB]	1	NULL	BH502	998	5	8	N_P	Río Duero	N_A	Río Duero	N_A	SPA	N_A	0.304300301971079	0.000547354228125392
15	[BLOB - 4.4 KB]	1	NULL	BH502	998	5	8	N_P	Río Duero	N_A	Río Duero	N_A	SPA	N_A	1.50728532500329	0.00233091248400066
16	[BLOB - 2.0 KB]	1	NULL	BH502	998	5	8	N_P	Río Duero	N_A	Río Duero	N_A	SPA	N_A	0.609169261844702	0.000963948662000101
17	[BLOB - 1.3 KB]	1	NULL	BH502	998	5	8	N_P	Río Duero	N_A	Río Duero	N_A	SPA	N_A	0.305854290708324	0.000267268964875253
18	[BLOB - 1.3 KB]	1	NULL	BH502	998	5	8	N_P	Río Ebro	N_A	Río Ebro	N_A	SPA	N_A	0.391593388905404	0.00076943976949976
19	[BLOB - 2.4 KB]	1	NULL	BH502	998	5	8	N_P	Río Duero	N_A	Río Duero	N_A	SPA	N_A	0.56242569325892	0.000499580664250498
20	[BLOB - 8.2 KB]	1	NULL	BH502	998	5	8	N_P	Río Ebro	l'Ebre	Río Ebro	l'Ebre	SPA	CAT	2.67645435713158	0.00465759825712593
21	[BLOB - 1.3 KB]	1	NULL	BH502	998	5	8	N_P	Río Duero	N_A	Río Duero	N_A	SPA	N_A	0.320874573058362	0.000232084841375176
22	[BLOB - 24.4 KB]	1	NULL	BH502	998	5	8	N_P	Río Tajo	N_A	Río Tajo	N_A	SPA	N_A	4.67601690584512	0.00642030944249917
23	[BLOB - 11.3 KB]	1	NULL	BH502	998	5	8	N_P	Río Tajo	N_A	Río Tajo	N_A	SPA	N_A	2.11628999103004	0.00272311978849989
24	[BLOB - 5.6 KB]	1	NULL	BH502	998	5	8	N_P	Río Tajo	N_A	Río Tajo	N_A	SPA	N_A	1.43296368634324	0.00265016924175068
25	[BLOB - 2.7 KB]	1	NULL	BH502	998	5	8	N_P	Río Tajo	N_A	Río Tajo	N_A	SPA	N_A	0.709279994679565	0.00113137021912505
26	[BLOB - 1.6 KB]	1	NULL	BH502	998	5	8	N_P	Río Tajo	N_A	Río Tajo	N_A	SPA	N_A	0.508693903565033	0.000915111082875807
27	[BLOB - 13.4 KB]	1	NULL	BH502	998	5	8	N_P	Río Guadiana	N_A	Río Guadiana	N_A	SPA	N_A	4.31555209663688	0.00644558503950188
28	[BLOB - 26.7 KB]	1	NULL	BH502	998	5	8	N_P	Río Guadalquivir	N_A	Río Guadalquivir	N_A	SPA	N_A	7.46342175239308	0.0121585657682486
29	[BLOB - 3.3 KB]	1	NULL	BH502	998	5	8	N_P	Río Guadiana	N_A	Río Guadiana	N_A	SPA	N_A	1.01100674639907	0.00105417835125013
30	[BLOB - 6.0 KB]	1	NULL	BH502	998	5	8	N_P	Río Odiel	N_A	Río Odiel	N_A	SPA	N_A	1.65331247300766	0.00261000578512421

- **Valores de atributos**. Se encuentran diferentes valores de atributos asociados a un mismo fenómeno geográfico, vinculados a los diferentes elementos geométricos de la representación gráfica de un determinado fenómeno. Así, pueden apreciarse diferentes valores de un mismo atributo (por ejemplo del atributo "longitud") vinculados a las diferentes geometrías. Diversos ejemplos de esta problemática pueden apreciarse en la base de datos Waterbase, concretamente en EEA_Water base river, o en la base de datos de EuroGlobalMap. Algunos de estos ejemplos son; valores de longitud asociados al Río Deva (47,22; 48,34; etc.), Río Miño (66,57; 185,82), etc.
- **Idioma**. La presencia de información en diferentes idiomas es un hecho frecuente en contextos multilingüe. En el ámbito de las bases de datos estudiadas, este problema se presenta por la existencia de cinco lenguas oficiales (castellano, catalán, gallego, vasco y aranés) en el ámbito nacional español. Este hecho conlleva la construcción de bases de datos monolingües o multilingües (caracterizadas por la mezcla de información de

[1] *Topología*. Trata con las características de las figuras geométricas que permanecen invariables si el espacio se deforma elásticamente y de forma continua, por ejemplo, cuando los datos se transforman de un sistema de coordenadas a otro. En el contexto de la información geoespacial, la topología se utiliza para describir la conectividad de un grafo n-dimensional (Buehler, 2003).

[2] *Geometría*. Proporciona el significado de la descripción cuantitativa a través de coordenadas y funciones matemáticas de las características espaciales de los fenómenos, incluyendo información sobre dimensión, posición, tamaño, forma y orientación (Buehler, 2003).

distintos idiomas dentro de una misma fuente de información). Estas situaciones generan cuatro escenarios en los que se pueden presentar aspectos multilingües y, en consecuencia, diversos tipos de problemas:

- **Tipos de fenómenos**. Con respecto a la multilingualidad que afecta únicamente a los tipos de fenómenos, en las bases de datos estudiadas, pueden encontrarse instancias que presentan el mismo nombre de instancia y tipos de fenómenos multilingües (por ejemplo; "Turia, Río" y "Turia, Riu" presentes en la base de datos del Nomenclátor Conciso).
- **Instancias**. En relación a la presencia de multilingualidad en el nombre de las instancias se pueden encontrar nombres de instancia multilingües y tipos de fenómenos monolingües (por ejemplo, "Ría Mundaca" y "Ría Mundaka" en la base de datos de EuroRegionalMap)

Tabla 8.8. Extracto de la base de datos de ERM con información incompleta

F_CODE	ICC	SN	HYC	NHI	NAMN1	NAMN2	NAMA1	NAMA2	NLN1	NLN2	TID	ZV2	Shape_Length
BH080	ES	5110	8	N_P	N_P	N_A	N_P	N_A	N_A	N_A	1	-29997	0.02282
BH080	ES	5110	8	N_P	N_P	N_A	N_P	N_A	N_A	N_A	1	-29997	0.02173
BH080	ES	5110	8	N_P	Balsa del Planerón	N_A	Balsa del Planeron	N_A	SPA	N_A	1	-29997	0.01748
BH080	ES	5110	8	N_P	N_P	N_A	N_P	N_A	N_A	N_A	1	-29997	0.01275
BH080	ES	5110	8	N_P	N_P	N_A	N_P	N_A	N_A	N_A	1	-29997	0.03842
BH080	ES	5110	8	N_P	N_P	N_A	N_P	N_A	N_A	N_A	1	-29997	0.02548

- **Tipos de fenómenos e instancias**. En este caso, el componente multilingüe se presenta en los tipos de fenómenos y en los nombres de las instancias (por ejemplo; "Mequinenza, Embalse de" y "Mequinensa, Pantà de" presentes en la base de datos del Nomenclátor Conciso).
- **Nexos**. Los nexos, también denominados enlaces o *stop words*[3] (Hu *et al*., 2007), son un componente común en el nombre de las instancias. Desde la perspectiva lingüística están conformados principalmente por determinantes, preposiciones y conjunciones. En ocasiones, únicamente, se recoge presencia de multilingüismo en los enlaces entre los tipos de fenómenos y nombres de instancia (por ejemplo; "RÍA DO RIBADEO/RÍA DO EO" y "RÍA DEL EO" presentes en BTN25).

En los problemas detectados en el nivel pragmático, en el contexto de una única base de datos, se da una mezcla de dependencia e independencia del dominio. Los problemas dependientes del dominio están vinculados a una visión general del área, por tanto, aunque en este caso se encuentran particularizados en el dominio hidrográfico, pertenecen al dominio geográfico. En este ámbito se encuentran los problemas referidos a nombres oficiales y alternativos, así como, los pertenecientes a topología y geometría. Por otro lado, en los problemas independientes del dominio aparecen los referenciados a información incompleta y valores de atributos. Finalmente, los problemas vinculados al idioma, dadas las características presentadas se encuentran en un punto de intersección entre la vinculación al dominio y la independencia del mismo.

[3] *Stop words*. En recuperación de información las *stop words* son palabras que no tienen significado semántico (por ejemplo: el, de, la, etc.).

Múltiples bases de datos. En el análisis del conjunto de las bases de datos estudiadas, los problemas descritos en el ámbito de una única base de datos repercuten en el conjunto de la información durante el proceso de integración, apareciendo los siguientes problemas:

- **Duplicidad**. Los solapamientos de información entre diferentes fuentes provoca duplicidad de información (instancias repetidas) como resultado del proceso de integración. Esta problemática se trata en detalle en el Capítulo 8 de este trabajo.
- **Identificador**. La integración de diferentes fuentes de información conlleva la presencia de diferentes identificadores para diferentes instancias, es decir, en diversas tablas de diferentes bases de datos contienen instancias similares, aunque éstas presentan diferentes identificadores (IDs) asociados.
- **Valores de atributos**. Este problema es causado por las diferentes abstracciones de los fenómenos geográficos y sus características relacionadas. Así, por ejemplo, resulta frecuente encontrar diversos valores de longitud o superficie asociados a una misma instancia (la longitud del "Río Bidasoa" es 67 km en la base de datos del Atlas Nacional, mientras en la correspondiente al proyecto WISE el valor es de 59,48 km). Este problema está relacionado con la heterogeneidad semántica contextual reflejada en (Worboys *et al*., 1991).
- **Unidades de medida**. La utilización de diversas unidades de medida en las fuentes de información estudiadas se deriva, entre otros factores, de las diferentes abstracciones e instrumentos utilizados en la captura de los fenómenos geográficos. Así, entre las bases de datos se encuentran ejemplos como: Río Ulla", cuyo valor de longitud (241585222699906) en EuroRegionalMap está expresado en grados decimales, mientras en el proyecto WISE se utiliza el kilómetro como unidad de medida (124,84 km).
- **Idioma**. La presencia de bases de datos monolingües y multilingües en el proceso de integración lleva los problemas descritos a nivel individual al escenario global de las fuentes estudiadas. A continuación, a modo de ejemplo, se describe los problemas relacionados con los tipos de fenómenos.

    - **Tipos de fenómenos**. En las bases de datos estudiadas pueden encontrarse instancias que presenten el mismo nombre de instancia y tipos de fenómenos multilingües (por ejemplo; "Itsasadarra Oria" y "Río Oria" presentes en la base de datos de EuroRegionalMap y la correspondiente al proyecto WISE, respectivamente).

- **Granularidad**. La existencia de diferencias en el volumen y nivel de detalle en fuentes de información con idéntica y/o diferente escala, es decir, la cantidad de información recogida encargada de describir un conjunto de fenómenos geográficos o sus características varía de una fuente a otra. Por ejemplo, esta situación de disparidad se manifiesta en el proceso de integración de BTN25 y la base de datos correspondiente al Nomenclátor NOMGEO. Este hecho incide de manera directa en el nivel de solapamiento de la información. En (Gómez-Pérez *et al*., 2008) el problema de la granularidad se trata desde la perspectiva del desigual número de conceptos que describen un dominio geográfico dentro de un catálogo de fenómenos. En el contexto de las bases de datos reflejado con anterioridad este problema se presenta a nivel conceptual y de atributos:

    - **Granularidad conceptual**. En este nivel, por ejemplo, se puede encontrar la siguiente información en la base de datos del Nomenclátor NOMGEO: "Cubillas, Río", "Cubillas, Barranco de", "Cubillas, Rambla de", "Cubillas, Arroyo" y "Cubillas, Valle del". Mientras en BTN25 la única información que aparece es la perteneciente a la instancia "Río Cubillas".

- **Granularidad de atributos**. En este nivel las diferencias están vinculadas a la cantidad de atributos relacionados con un fenómeno geográfico. Por ejemplo, existen diferentes atributos para conjuntos de instancias similares. Así para la instancia del "Embalse de Escales" en la base de datos de EGM aparecen atributos como: nombre de instancia, longitud, área o idioma de codificación. La base de datos de la CHE presenta atributos como: nombre de instancia, tipología, capacidad o coordenadas. Mientras, la información contenida en la base de datos del Nomenclátor Conciso es más descriptiva, recogiendo atributos como nombre de instancia, nombre oficial o alternativo, coordenadas, etc.

Este problema es completamente diferente al problema de heterogeneidad semántica relacionado con sinónimos, homónimos o heterogeneidad designativa presentes en (Naiman *et al.*, 1995; Bishr, 1998; Chatterjee *et al.*, 1995). La diferencia se circunscribe al hecho de que en el contexto de la información geoespacial existen diferentes organizaciones legales a diferentes niveles (por ejemplo, Comisión Especializada de Nombres Geográficos[4], EuroGeonames, Conferencias de las Naciones Unidas sobre la Normalización de los Nombres Geográficos[5] y Grupo de Expertos de las Naciones Unidas en Nombres Geográficos (GENUNG)[6]) encargadas de la normalización y oficialización de los nombres geográficos.

- **Generalización y agregación**. Los tipos de problemas asociados con la generalización y agregación (Naiman *et al.*, 1995) están relacionados con la representación cartográfica, el factor escala y la abstracción conceptual. En este contexto los problemas encontrados se vinculan a las clasificaciones de super-clase y subclase.

  - *Super-clase*. Existen individuos en las bases de datos que tienen el mismo nombre de instancia, pero su tipo de fenómeno relacionado es una super-clase de otro tipo de fenómeno (por ejemplo, "Acequia de la Fábrica de Luz" es una super-clase de "Fábrica de Luz, Caz de la"). Estas instancias aparecen en las bases de datos de BTN25 y del Nomenclátor NOMGEO, respectivamente.
  - *Subclase*. Existen algunos individuos que tienen el mismo nombre de instancia, pero su tipo de fenómeno relacionado es una subclase de otro tipo de fenómeno (por ejemplo, "Riachuelo de la Cañada" es subclase de "Arroyo de la Cañada").

De forma general, dentro de los problemas pertenecientes a varias bases de datos detectados en el nivel pragmático, aparecen problemas clasificados como dependientes e independientes del dominio. Con respecto a los independientes, aparecen los relativos a los identificadores, duplicidad y valores de atributos. Por otro lado, los referidos a los dependientes del dominio están vinculados a los problemas de granularidad, generalización y agregación. Estos problemas están estrechamente vinculados al dominio hidrográfico, aunque por sus características fácilmente podrían encontrarse en un modo más general, es decir, en todo el dominio geográfico. Con respecto a los problemas relacionados con el idioma, al igual que ocurría a nivel de una única base de datos, se encuentran en un punto de intersección entre la vinculación al dominio y la independencia del mismo.

[4] *Comisión Especializada de Nombres Geográficos*.
http://www.fomento.es/MFOM/LANG_CASTELLANO/DIRECCIONES_GENERALES/ORGANOS_COLEGIADOS/CSG/TOPONIMIA/ComNomGeo.htm
[5] *United Nations Conferences on the Standardization of Geographical Names*.
http://unstats.un.org/unsd/geoinfo/default_geoinfo.htm
[6] *United Nations Group of Experts on Geographical Names (UNGEGN)*.
http://unstats.un.org/unsd/geoinfo/ungegn.htm

Finalmente, como resultado de la tarea 6.1 (Actividad 6), a modo de resumen, en la Tabla 8.9 se recogen los diferentes problemas presentados en el proceso de integración del conjunto de las instancias presentes en el contexto de las bases de datos estudiadas. Estos problemas son organizados conforme a los diferentes niveles (léxico, sintáctico, semántico y pragmático) expuestos con anterioridad.

8.3. Un acercamiento heurístico para la detección de duplicados

En esta sección se recoge el desarrollo de las tareas descritas en la Actividad 6 de la metodología propuesta (Capítulo 5, sección 5.5.6). Así, en la sección 8.3.1 se presenta el planteamiento e influencias del proceso de detección de duplicados. En la sección 8.3.2 y 8.3.3 se describen las diferentes aproximaciones desarrolladas para la detección de instancias duplicadas en un mismo y en múltiples conjuntos de datos, respectivamente.

8.3.1. Planteamiento

La duplicidad de instancias aparece consecuencia de que diferentes fuentes capturan una abstracción particular de la realidad. En el contexto de las fuentes de información estudiadas, a pesar de que dichas fuentes pertenecen a un mismo productor, éstas no tienen un tratamiento común ya que se gestionan de manera independiente. Sin embargo, las fuentes estudiadas consideran un mismo ámbito de recogida de datos (ámbito nacional español), por tanto, existe información común, aunque ésta puede presentar diferencias por los mecanismos de captura y las diferencias de escala de representación.

Considerando el conjunto de fuentes estudiadas (ver Anexo) y su transformación a RDF (Capítulo 7), cada conjunto de datos contiene un volumen determinado de recursos (URI) y descripciones.

$$D_1 = \{a_1, a_2, \dots\ a_n\}, D_2 = \{b_1, b_2, \dots\ b_m\}, \dots D_n = \{n_1, n_2, \dots n_n\}$$

Donde D_n corresponde a cada una de las fuentes de información en RDF,
mientras $\{n_1,\ n_2, \dots\ n_n\}$ corresponde a los diferentes recursos de cada fuente.

Dadas dos fuentes de información D_1 y D_2 aparecen conjuntos de recursos $D_1 \in \{a_1, a_2, \dots a_n\}$ y $D2 \in \{b1, b2, \dots bm\}$ con particulares tipos de fenómenos.

Como se afirma en el Capítulo 3 (sección 3.3), el objetivo de la detección de instancias duplicadas consiste en identificar entidades referidas al mismo fenómeno del mundo real, incluso si los registros no son idénticos, y relacionar o fusionar dichas entidades. Para ello, se requiere identificar todas las correspondencias existentes entre recursos $\{a1, a2, \dots an\} \times \{b1, b2, \dots bm\}$ que representen a la misma instancia del mundo real. La identificación de similitud (s) conlleva el establecimiento del producto cartesiano $c = (a_i, b_j, s)$ entre los recursos de fuentes D_1 y D_2.

El proceso de detección de duplicados está inspirado en la noción de emparejamiento de instancias (*instance matching*). Esto permite describir el problema dentro del reconocido *framework* de emparejamiento de ontologías (*ontology matching*). El *ontology matching* (Euzenat *et al.*, 2007) se define como el proceso de detección de enlaces entre entidades de diferentes ontologías. Estos enlaces se anotan mediante un valor de confianza y una etiqueta que describe el tipo de enlace. Dicho enlace se denomina correspondencia y un conjunto de estas correspondencias se denomina alineamiento.

Tabla 8.9. Problemas presentes en las instancias en el proceso de integración del conjunto de bases de datos estudiadas organizados en diferentes niveles

Niveles	Problemas	Independiente	Dependiente		Fuente		Escala	
			Geográfico	Hidrográfico	Única	Múltiples	=	≠
Léxico	Acentuación	✓			✓		✓	✓
	Caracteres especiales	✓			✓		✓	✓
	Abreviaturas	✓			✓	✓	✓	✓
	Mayúsculas / Minúsculas	✓				✓	✓	✓
	Espacios en blanco	✓			✓	✓	✓	✓
Sintáctico	Errores tipográficos	✓			✓		✓	
	Proximidad del nombre	✓			✓	✓	✓	✓
	Orden de nombrado	✓			✓	✓	✓	✓
Semántico	Diferentes instancias (campo)	✓			✓			✓
	Clasificaciones		✓			✓	✓	✓
	Ausencia de fenómeno		✓		✓	✓		✓
Pragmático	Nombres oficiales y alternativos		✓	✓	✓		✓	✓
	Duplicidad	✓				✓	✓	✓
	Topología y Geometría		✓	✓	✓		✓	✓
	Información incompleta	✓			✓		✓	✓
	Valores de atributos	✓			✓	✓	✓	✓
	Idioma	✓	✓		Fenómenos | Instancias | Ambos | Nexos	Fenómenos	✓	✓
	Identificador	✓				✓	✓	✓
	Unidades de medida					✓		✓
	Granularidad		✓	✓		Conceptual | Atributos	✓	✓
	Generalización y Agregación		✓	✓		Super-clase | Sub-clase	✓	✓

Asimismo, en la detección de duplicados se considera que no sólo los nombres de los recursos (topónimos) resultan relevantes, sino que el conocimiento del dominio posee un rol clave en dicho proceso. Esta importancia del dominio resulta vital para guiar el proceso de detección (en grandes cantidades de información), en la depuración de instancias candidatas incorrectas y en la evaluación de las instancias seleccionadas (Doan *et al.*, 2005). Este hecho motiva el desarrollo de un enfoque heurístico que combina información de diferentes niveles de dependencia e independencia del dominio, también denominada información espacial y no-espacial (Beeri *et al.*, 2004; Sehgal *et al.*, 2006). Respecto a las características dependientes a nivel hidrográfico (por ejemplo, un río es similar a una corriente de agua, pero es bastante diferente de una laguna), mientras en el nivel geográfico (una línea y/o polilínea pueden representar al mismo individuo geográfico en diferentes escalas, mientras una línea y/o un punto siempre van a representar individuos diferentes). Asimismo, desde la óptica de las características independientes del dominio se hace patente la consideración de diferentes aspectos (por ejemplo, las entidades con el mismo nombre inicialmente podrían ser consideradas similares entre sí).

La consideración de la relevancia de la información relacionada con el dominio está inspirada en trabajos como la métrica *Fellegi and Sunters (SFS)* (Fellegi *et al.*, 1969) y en la mejora de la precisión de *structure exploitation* (Morris *et al.*, 2008).

8.3.2. Enfoques heurísticos

Conforme a lo expuesto en la Actividad 6 de la metodología propuesta (Capítulo 5) el proceso de detección de duplicados se realiza mediante diversos enfoques heurísticos.

Según Rich *et al*. (1994), una heurística es una técnica que aumenta la eficiencia de un proceso de búsqueda, posiblemente sacrificando demandas de completitud. Las heurísticas resultan adecuadas porque generalmente suelen indicar datos interesantes, pero son malas porque pueden olvidar información de interés para ciertas personas. Algunas heurísticas ayudan a guiar el proceso de búsqueda sin sacrificar ninguna demanda de completitud que el proceso haya podido tener previamente. Otras (en realidad, muchas de las mejores) pueden ocasionalmente causar que una información sea pasada por alto. Pero, en promedio, mejoran la calidad de la información que explotan. Existen algunas heurísticas de propósito general que son adecuadas para una amplia variedad de dominios de problemas. Asimismo, es posible construir heurísticas de propósito especial que exploten el conocimiento específico del dominio para resolver problemas particulares.

En las heurísticas desarrolladas se consideran diversas características dependientes e independientes del dominio de conocimiento. Así, se utiliza diversos atributos representativos del dominio geográfico (coordenadas, geometría e información temática (área, perímetro, etc.)), mientras para la información independiente del dominio o no espacial se consideran URI (nombre), *label* e idioma. Además, las heurísticas se clasifican en distintas capas (léxica, sintáctica, semántica y pragmática), acorde con los problemas recogidos en el ámbito de las instancias (Capítulo 8, sección 8.2).

Conjuntos de datos RDF

Los elementos de entrada de este sistema están formados por conjuntos de datos RDF accesibles a través de un SPARQL Endpoint. En este trabajo los datos provienen del RDF generado con la tecnología R2O+, ODEMapster++ y por la librería GEOMETRY2RDF. Para más detalles sobre el proceso de generación de los conjuntos de datos RDF se recomienda revisar el Capítulo 7 de este trabajo.

Los conjuntos de datos de entrada se caracterizan por poseer el mismo sistema de coordenadas (WGS84) y por tener superposición de información entre los diversos conjuntos comparados.

Almacenamiento y consulta

Este componente está formado por el *triple store* Virtuoso Universal Server. Como se mencionó con anterioridad, es un middleware y motor de bases de datos que combina la funcionalidad de un Sistema Gestor de Bases de Datos tradicional, una base de datos virtual, RDF, XML, texto libre, servidor de aplicaciones web y *Linked Data*, y la funcionalidad de servidor de ficheros.

En la arquitectura del sistema este componente se utiliza para almacenar, consultar y servir conjuntos de datos RDF. Además, entre otras funcionalidades, Virtuoso está diseñado para dar soporte a SPARQL (Prud'hommeaux *et al.*, 2008), lenguaje de consulta desarrollado por el W3C RDF Data Access Working Group. De este componente se produce la generación de SPARQL Endpoints, es decir, servicios que permiten al usuario consultar la base de conocimiento (conjuntos de datos RDF almacenados en un repositorio) a través del lenguaje de consulta SPARQL. Asimismo, se puede generar un único SPARQL Endpoint y definir tantos grafos (*graph*) como conjuntos de datos a publicar. En el sistema desarrollado los SPARQL Endpoints desplegados se conectan a la red (Internet). De esta manera, se ofrece a los usuarios la posibilidad de consultar directamente estos datos sin necesidad de pasar por todo el proceso que ofrece este sistema.

A continuación se procede a la descripción de los distintos enfoques heurísticos desarrollados para la detección de instancias duplicadas en el ámbito de un mismo y entre múltiples conjuntos de datos geoespaciales en formato RDF.

8.3.2.1. Detección de duplicados en un mismo conjunto de datos

Dentro de una misma fuente de información geoespacial, con frecuencia, aparecen instancias de fenómenos (topónimos) que poseen el mismo nombre y, sin embargo, tienen distinta localización. En consecuencia, estos topónimos representan distintos fenómenos de la realidad. Para detectar estas instancias duplicadas en el contexto de un mismo conjunto de datos se ha desarrollado un enfoque heurístico semiautomático de comparación de datos geográficos, cuya visión general se recoge en la Figura 8.3.

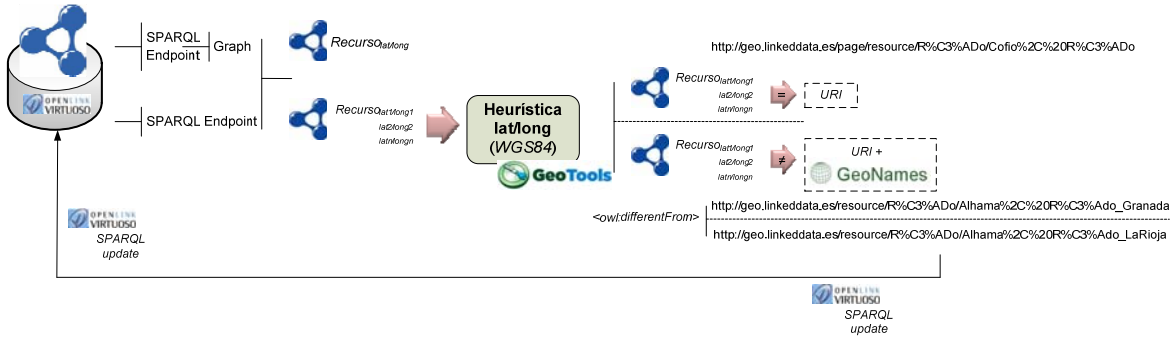


Figura 8.3. Detección de instancias duplicadas en un mismo conjunto de datos

En primer lugar, esta heurística se centra en descubrir si un recurso que posee más de una geometría de tipo puntual (latitud/longitud) representa un mismo fenómeno geográfico o, por el contrario, recoge distintos fenómenos con idéntico nombre.

$$d\ (a_1\ lat/long_1, a_1\ lat/long_2 \ldots\ a_n\ lat/long_n) \geq\ distance_n \Rightarrow a_1 = a_1$$
$$d\ (b_2\ lat/long_1, b_2\ lat/long_2 \ldots\ \ b_m\ lat/long_m) \geq\ distance_m \Rightarrow b_2 \neq b_2$$

Así, esta heurística recorre los recursos disponibles en un SPARQL Endpoint o en un

grafo determinado[1], recuperando aquellos recursos que tienen más de un valor de latitud /longitud asociado.

Tras recuperar esta información, esta heurística utiliza la librería GeoTools, que proporciona librerías para el tratamiento de la información geoespacial, para la comparación de coordenadas conformes con WGS84 o WKT. Esta comparación se realiza a través de la siguiente función:

$$R_1[p1(lat, long)].isWithinDistance\ R_1\{[p2(lat, long)], distance\}$$

Donde R_1 representa a un recurso (instancia) en un mismo conjunto de datos.
`p1(lat,long)` corresponde a un par de valores de una geometría de tipo puntual asociados al recurso.
`p2(lat,long)` corresponde a otro par de valores de una geometría puntual asociados con el mismo recurso.
`Distance` corresponde a la longitud del radio considerada en la comparación.

Los resultados obtenidos de la comparación de los distintos valores de la geometría puntual (latitud/longitud) dividen a los recursos en dos niveles:

- **Valores similares**. Los valores obtenidos se encuentran dentro de los umbrales de distancia definidos. Por tanto, las distintas coordenadas pertenecen a un mismo fenómeno geográfico caracterizado por tener presencia en diversas localizaciones espaciales.

```
http://geo.linkeddata.es/page/resource/R%C3%ADo/Cofio%2C%20R%C3%ADo

geo:geometry
      <http://geo.linkeddata.es/resource/wgs84/40.4166666666667_-
   4.31666 666666667>
      <http://geo.linkeddata.es/resource/wgs84/40.525866633996_-
   4.263189 32340339>
      <http://geo.linkeddata.es/resource/wgs84/40.5979760892974_-
   4.28380 890015799>
rdfs:label
   Cofio, Río (es)
```

- **Valores diferentes**. Los valores obtenidos no se encuentran dentro de los umbrales de distancia definidos. En este caso, a pesar de que los recursos poseen la misma URI, los resultados demuestran que existe información asociada a varios recursos. Así, en este tipo de recursos se está recogiendo información sobre distintos fenómenos geográficos caracterizados por poseer nombre idéntico y distinta localización espacial.

```
http://geo.linkeddata.es/resource/R%C3%ADo/Alhama%2C%20R%C3%Ado

geo:geometry
   <http://geo.linkeddata.es/resource/wgs84/36.9516312627864_3.955720
78918559>
   <http://geo.linkeddata.es/resource/wgs84/37.0166666666667_-3.96666
666666667>
   <http://geo.linkeddata.es/resource/wgs84/41.8814627364368_-2.20867
947403874>
```

[1] En el contexto de este trabajo los SPARQL Endpoints y/o grafos se generan con el tri*ple store* Virtuoso Universal Server (descrito en el Capítulo 7 de este trabajo).

```
   <http://geo.linkeddata.es/resource/wgs84/42.0144274433984_-1.94388
130850726>
   <http://geo.linkeddata.es/resource/wgs84/42.0672891088227_-1.82308
151663923>
   <http://geo.linkeddata.es/resource/wgs84/42.1833333333333_-1.71666
666666667>
rdfs:label
   Alhama, Río (es)
```

Una vez se ha identificado la categoría de un recurso con múltiples valores de latitud/longitud se procede a:

- **Mantenimiento de URI**. Si los valores extraídos resultan similares, se mantiene la URI del recurso como está almacenada en el SPARQL Endpoint o grafo correspondiente.

```
http://geo.linkeddata.es/page/resource/R%C3%ADo/Cofio%2C%20R%C3%ADo
```

- **Modificación de URI**. Si los valores extraídos resultan diferentes, se procede a la distinción de las URI de los distintos recursos mediante la incorporación de un topónimo para los recursos identificados. Este topónimo proviene de la respuesta ofrecida por el servicio de geocodificación inversa (*reverse geocoding*)[2] externo utilizado. Concretamente, se utiliza un servicio de GeoNames para asignar topónimos a los distintos valores de latitud/longitud de un mismo recurso. Además, se establece una relación de tipo `<owl:differentFrom>` entre las URI generadas para reflejar explícitamente la diferencia entre dichos recursos.

```
http://geo.linkeddata.es/resource/R%C3%ADo/Alhama%2C%20R%C3%Ado_Granada

geo:geometry
      <http://geo.linkeddata.es/resource/wgs84/36.9516312627864_-
   3.95572 078918559>
      <http://geo.linkeddata.es/resource/wgs84/37.0166666666667_-
   3.96666 666666667>
http://geo.linkeddata.es/resource/R%C3%ADo/Alhama%2C%20R%C3%Ado_LaRioja

geo:geometry
      <http://geo.linkeddata.es/resource/wgs84/41.8814627364368_-
   2.20867 947403874>
      <http://geo.linkeddata.es/resource/wgs84/42.0144274433984_-
   1.94388 130850726>
      <http://geo.linkeddata.es/resource/wgs84/42.0672891088227_-
   1.82308 151663923>
      <http://geo.linkeddata.es/resource/wgs84/42.1833333333333_-
   1.71666 666666667>
```

Finalmente, se realiza la actualización de los nuevos recursos generados en el repositorio RDF (*triple store*). Para ello, se utiliza la API del *triple store*[3] desplegado en el contexto de este trabajo (OpenLink Virtuoso). Esta API permite ejecutar sentencias SPARQL *update* para actualizar automáticamente los recursos identificados.

[2] Geocodificación inversa. Es el proceso de encontrar una localización textual (topónimo) de unas coordenadas geográficas (normalmente expresadas como latitud y longitud).

[3] En el contexto de este trabajo se utiliza la API de OpenLink Virtuoso para la ejecución de la sentencia SPARQL *update* por las limitaciones, en el momento de este desarrollo, en la especificación SPARQL.

La aplicación de esta heurística sobre un mismo conjunto de datos contribuye a la depuración o curación de los datos provenientes de las distintas fuentes de información. Asimismo, los resultados obtenidos de esta heurística pueden servir como preprocesamiento de los recursos (RDF) publicados. Este proceso previo contribuirá a mejorar los resultados de la detección de instancias duplicadas entre múltiples conjuntos de datos geoespaciales heterogéneos.

Dadas las características de esta heurística, ésta se clasifica en el nivel pragmático y dependiente del dominio.

8.3.2.2. Detección de duplicados en múltiples conjuntos de datos

En esta sección se describen las principales características (preprocesamiento y métricas de similitud) y los diversos enfoques heurísticos desarrollados para la detección de instancias duplicadas en múltiples y heterogéneos conjuntos de datos RDF geoespaciales.

Preprocesamiento

De forma previa a la aplicación de las heurísticas sobre los conjuntos de datos se realiza la transformación de la codificación de las URI (caracteres no ASCII-US – codificación %) de los recursos a la codificación UTF8. Esta transformación se encarga de sustituir la codificación % por caracteres más eficientes para las diversas métricas utilizadas en el proceso de comparación. Este proceso previo se realiza como consecuencia de que en análisis previos realizados a los resultados obtenidos de las diferentes heurísticas se detectó la pérdida de resultados durante la comparación de recursos por la consideración de los caracteres %. De esta manera, se consigue que la comparación entre los diferentes recursos ofrezca unos resultados más depurados y conformes a la realidad de sus semejanzas. En la Figura 8.4 se muestra un ejemplo de la transformación realizada en las URI de los recursos.

```
http://geo.linkeddata.es/resource/NOMGEO/R%C3%ADo/Arag%C3%B3n%2C%20R%C3%AD
```

```
http://geo.linkeddata.es/resource/NOMGEO/Río/Aragón, Río
```

Figura 8.4. Transformación de la codificación de las URI

Métricas de similitud

Son numerosos los trabajos que consideran la similitud entre cadenas (*strings*) como medida para la detección de instancias duplicadas. Esto repercute en la existencia de múltiples métricas[4] para medir la similitud (Elmagarmid *et al.*, 2007). Sin embargo, consecuencia de la gran variedad de fuentes de datos y de entidades a emparejar no existe un único algoritmo universalmente reconocido como el mejor (Köpcke *et al.*, 2009).

En esta sección se recogen las métricas utilizadas para calcular la similitud entre cadenas. Estas métricas se dividen en dos categorías: igualdad y similitud. Esta división se utiliza para combinar las métricas y obtener el valor de similitud más óptimo en la comparación de recursos de diferentes conjuntos de datos. Cada métrica ofrece un valor de similitud entre 0 y 1. Cuanto más elevados sean los valores mayor resulta la similitud entre los recursos. A continuación se presentan las métricas de similitud utilizadas en el sistema desarrollado (ver Tabla 8.10 y Tabla 8.12).

- **Igualdad**. Las métricas de esta categoría se caracterizan por ofrecer resultados de similitud de 0 ó 1. Así, cuando se comparan dos instancias exactamente iguales estas métricas ofrecen como resultado un valor de 1. Sin embargo, cuando se comparan dos

[4] Una recopilación exhaustiva de las métricas existentes se recoge en la siguiente dirección web: http://staffwww.dcs.shef.ac.uk/people/S.Chapman/stringmetrics.html

instancias que poseen diferencias, por pequeñas que estas sean, las métricas ofrecen un valor de 0. Las diferentes métricas de esta categoría se listan en la Tabla 8.10.

Tabla 8.10. Conjunto de métricas de similitud utilizadas

Métrica	Descripción
Block distance	Medida de similitud de cadenas basada en la métrica Block distance o L1 distance
Cosine similarity	Medida de similitud entre dos vectores de n dimensiones para encontrar el coseno del ángulo entre ellos
Dice's Coefficient	Medida de similitud relacionada con el índice Jaccard
Jaccard Similarity	Medida estadística utilizada para comparar la similitud y diversidad de conjuntos de muestras
Matching Coefficient	Medida de similitud basada en la métrica Matching Coefficient
Overlap Coefficient	Medida de similitud relacionada con el índice Jaccard que computa la superposición entre dos conjuntos

De forma general, este conjunto de métricas se caracterizan por centrarse en la información relevante excluyendo del proceso de detección de similitud las *stop words* (Hu *et al*., 2007). Las *stop words* son palabras que no tienen significado semántico (por ejemplo: el, de, la, etc.), por tanto, estas palabras no se consideran útiles para la detección de similitud. Asimismo, en el contexto de las fuentes estudiadas hay conjuntos de datos que carecen de estas palabras, por tanto, la consideración de estas métricas no afectará en los resultados de la comparación de pares de recursos.

En el sistema desarrollado si la comparación de una pareja de instancias ofrece en todas estas métricas un valor igual a 1, entonces se registra que esos recursos son iguales. En caso contrario, la similitud entre la pareja de instancias es verificada con el siguiente grupo de métricas. Un ejemplo de los valores ofrecidos por estas métricas durante la comparación de diferentes recursos se muestra en la Tabla 8.11.

Tabla 8.11. Resultados métricas igualdad

$Recurso_{Fuente1}$	$Recurso_{Fuente2}$	Block Distance	Cosine Similarity	Dice Sim.	Jaccard Sim.	Matching Coefficient	Overlap Coefficient
@/Órbigo, Río	@/Órbigo, Río	1,00	1,00	1,00	1,00	1,00	1,00
@/Rio Naval	@/Río Basarua	0.0	0.0	0.0	0.0	0.0	0.0
@/Alagón, Río	@/Alagón, Río	1,00	1,00	1,00	1,00	1,00	1,00
@/Rio Garona	@/Río Baco	0.0	0.0	0.0	0.0	0.0	0.0
@/Albaida, Riu d'	@/Albaida, Riu d'	1,00	1,00	1,00	1,00	1,00	1,00
@/Río Ciurana	@/Río Sorteny	0.0	0.0	0.0	0.0	0.0	0.0
@/Ara, Río	@/Ara, Río	1,00	1,00	1,00	1,00	1,00	1,00
@/Rio Aranda	@/Río Izarilla	0.0	0.0	0.0	0.0	0.0	0.0
@/Ayuda, Río	@/Ayuda, Río	1,00	1,00	1,00	1,00	1,00	1,00
@/Bornova, Río	@/Bornova, Río	1,00	1,00	1,00	1,00	1,00	1,00
@/Cànyoles, Riu	@/Cànyoles, Riu	1,00	1,00	1,00	1,00	1,00	1,00
@/Cabra, Río de	@/Cabra, Río de	1,00	1,00	1,00	1,00	1,00	1,00
@/Cabriel, Río	@/Cabriel, Río	1,00	1,00	1,00	1,00	1,00	1,00

- **Similitud**. Las métricas de este grupo se caracterizan por ofrecer valores de similitud entre 0 y 1. Así, cuando se comparan dos instancias ya sean exactamente iguales o posean diferencias, en mayor o menor medida, estas métricas ofrecen como resultado valores entre 0 y 1. Las diferentes métricas de este grupo se listan en la Tabla 8.12.

Tabla 8.12. Conjunto de métricas de similitud utilizadas

Métrica	Descripción
JaroSimilarity	Similitud de cadenas basada en la métrica de distancia Jaro
Jaro Winkler	Medida de similitud entre dos cadenas, variante de la métrica Jaro
Levenshtein Distance	Mínimo número de edición necesario para transformar una cadena en otra
Monge Elkan distance	Medida de similitud basada en la métrica Monge Elkan distance
Needleman-Wunch distance	Realiza alineamiento global entre dos cadenas
Smith-Waterman-Gotoh	Realiza alineamiento local entre dos cadenas

En el sistema desarrollado el resultado obtenido por estas métricas (ver Tabla 8.13) es comparado con un umbral (*threshold*) δ. Asimismo, se incluye un proceso adicional que permite comparar si los valores obtenidos por un número determinado de métricas pasan un umbral establecido. Este proceso permite afinar el grado de similitud entre pares de instancias y, por tanto, la obtención de mejores resultados.

Esta división propuesta (métricas de igualdad y similitud) está basada en un análisis previo de los diferentes valores arrojados tras la aplicación de estas métricas sobre diversos conjuntos de datos. Este análisis permitió seleccionar esta división y combinación de métricas.

Tabla 8.13. Resultados métricas similitud (umbral 0,96)

RecursoFuente1	RecursoFuente2	Jaro	Jaro Winkler	Levenshtein	Monge Elkan	Needleman Wunch	Smith Waterman Gotoh Windowed Affine
@/Órbigo, Río	@/Órbigo, Río	0,99	0,99	0,92	1,00	0,92	1,00
@/Rio Naval	@/Río Basarua	0.83	0.86	0.56	0.57	0.65	0.57
@/Alagón, Río	@/Alagón, Río	0,99	0,99	0,92	1,00	0,92	1,00
@/Rio Garona	@/Río Baco	0.89	0.90	0.69	0.60	0.77	0.60
@/Albaida, Riu d'	@/Albaida, Riu d'	0,99	0,99	0,94	1,00	0,94	1,00
@/Río Ciurana	@/Río Sorteny	0,87	0.86	0.57	0.50	0.78	0.5
@/Ara, Río	@/Ara, Río	0,99	0,99	0,89	1,00	0,89	1,00
@/Rio Aranda	@/Río Izarilla	0.87	0.89	0.67	0.63	0.77	0.63
@/Ayuda, Río	@/Ayuda, Río	0,99	0,99	0,91	1,00	0,91	1,00
@/Bornova, Río	@/Bornova, Río	0,99	0,99	0,92	1,00	0,92	1,00
@/Cànyoles, Riu	@/Cànyoles, Riu	0,99	0,99	0,93	1,00	0,93	1,00
@/Cabra, Río de	@/Cabra, Río de	0,99	0,99	0,93	1,00	0,93	1,00
@/Cabriel, Río	@/Cabriel, Río	0,99	0,99	0,92	1,00	0,92	1,00

Arquitectura

Junto a los componentes generales descritos con anterioridad (conjuntos de datos de entrada y *triple store*), para la detección de instancias duplicadas en el contexto de múltiples conjuntos de datos se desarrollan diferentes enfoques heurísticos disponibles a través de una serie de servicios web y de una API (ver Figura 8.5).

- **Servicios web**. Este componente está constituido por una plataforma de servicios web que acoge los diferentes servicios desplegados para la detección de duplicados en el contexto de la información geoespacial.

  Este componente utiliza Jena para comunicar y procesar las consultas de los diferentes SPARQL Endpoints o grafos existentes utilizando el lenguaje SPARQL. Además, la

comunicación entre los diferentes servicios web y el cliente final se realiza a través de SOAP (*Simple Object Access Protocol*)[5].

- **API (Interfaz de Programación de Aplicaciones) heurísticas**. El conjunto de heurísticas desarrolladas se ofrece también como una biblioteca para que puedan ser utilizadas y/o integradas en otro software.

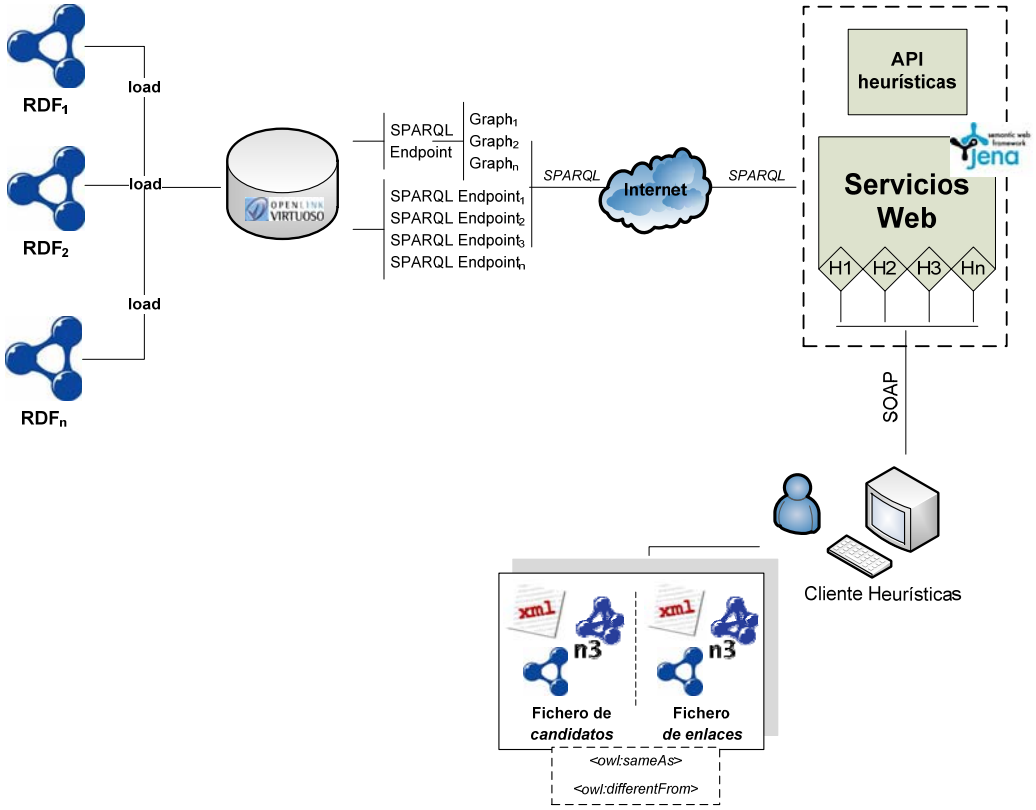


Figura 8.5. Arquitectura desplegada para la detección de duplicados en múltiples conjuntos de datos

El despliegue de los diferentes servicios recogidos en la plataforma sobre diversos conjuntos de datos genera como resultado dos ficheros en formato N3, RDF o XML.

- ***Enlaces.*** En esta salida se obtiene como resultado un fichero con las correspondencias establecidas entre los fenómenos geográficos duplicados en los conjuntos de datos. Estas correspondencias quedan formalizadas a través de la relación `<owl:sameAs>` y `<owl:differentFrom>`.
- ***Candidatos.*** En esta salida se ofrecen los resultados de posibles correspondencias adicionales entre fenómenos geográficos. La validación de estos resultados debe ser realizada por un usuario experto del dominio.

Heurísticas

En primer lugar se procede a consultar los tipos de recursos existentes en los diferentes SPARQL Endpoints o grafos entre los que se pretende realizar la comparación de recursos. Tras este paso, se recuperan los recursos de un determinado tipo. Esta limitación se introduce en el proceso consecuencia de las restricciones actuales de consulta de los SPARQL Endpoint. Estos

[5] *SOAP*. Es un protocolo ligero destinado al intercambio de información estructurada en un entorno descentralizado y distribuido. Utiliza tecnologías XML para definir un framework de comunicación extensible proporcionando unas construcciones (mensajes) que pueden ser intercambiados a través de diversos protocolos subyacentes. Para más información ver http://www.w3.org/TR/soap12-part1/

servicios no permiten recuperar grandes cantidades de información debido a problemas de memoria tanto en el servidor como en el cliente. Para solventar este problema se utilizan diversos parámetros en las consultas SPARQL (`limit` y `offset`), que limitan la extracción de recursos a un determinado número y permiten realizar el proceso de consulta y recuperación de los recursos de forma incremental. A continuación se describen los detalles de las diferentes heurísticas desarrolladas:

- **URI.** Esta heurística se basa en la comparación del elemento *localname* de las URI (*Qname*) para distintos SPARQL Endpoint o grafos de un SPARQL Endpoint determinado.

$$d\ (a_1, b_2) \geq\ \delta \Rightarrow a_1 = b_2$$

Para realizar dicha comparación, en primer lugar, se procede a recuperar los recursos (URI) de los diferentes SPARQL Endpoint o grafos.

```
Select distinct ?recurso
Where { GRAPH " + graphn + " { ?recurso a <" + t + ">} }
order by ?recurso limit Xn offset " + offset
```

Donde `?recurso` corresponde a la URI del recurso a comparar.
`graph`$_n$ indica el grafo donde se almacenan los conjuntos de datos a comparar.
`t` indica el tipo de recurso (fenómeno geográfico) a comparar.
`limit` corresponde al límite del número de resultados devueltos por la consulta.
X_n define el número de resultados delimitados por la operación `limit`.
`offset` indica la posición dentro del conjunto de datos en la que se continúa la consulta tras un número de soluciones recuperado.

Una vez recuperados los recursos de los diversos conjuntos de datos se procede a utilizar las métricas de similitud descritas con anterioridad. Los resultados de esta heurística dependen de los umbrales establecidos para las métricas, ya que se van a considerar posibles recursos similares aquellos que contemplen un valor de similitud igual o superior al umbral establecido. De este proceso se obtiene un fichero donde se listan los resultados obtenidos en formato N3, RDF o XML.

Esta métrica se clasifica entre los niveles léxico-sintáctico, debido a que las métricas consideradas permiten tratar con los diferentes problemas recogidos en estas capas (caracteres especiales, acentuación, mayúsculas/minúsculas, errores tipográficos, etc.). Además, dadas las características de esta heurística, se clasifica como independiente de dominio.

- ***Label* e idioma.** En esta heurística se compara el elemento *label* de los diferentes recursos almacenados en los distintos SPARQL Endpoint o grafos de un SPARQL Endpoint determinado.

$$d\ (a_1 label_1, b_2 label_2) \geq\ \delta \Rightarrow a_1 = b_2$$

Para realizar dicha comparación, se procede a recuperar los recursos y su elemento *label* en un idioma específico dentro los diferentes SPARQL Endpoint o grafos.

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
Select distinct ?recurso ?label
Where { GRAPH " + graph_n + " { ?recurso a <" + t + ">. ?recurso
rdfs:label ?label.
FILTER(langMatches(lang(?label),'" + idioma + "')).} }
order by ?recurso limit 100 offset " + offset
```

Donde `?label` corresponde al elemento *label* en el idioma especificado.
`idioma` indica el idioma definido para el elemento *label*.
`FILTER` es una operación que restringe los resultados a aquellos que cumplen la expresión definida en la condición de filtrado.

Una vez recuperados estos recursos se realiza la comparación entre aquellos elementos que presentan el mismo idioma. Asimismo, esta heurística utiliza las métricas de similitud descritas anteriormente. Por tanto, los resultados de esta heurística dependen de los umbrales establecidos para las métricas, ya que se van a considerar posibles recursos similares aquellos que contemplen un valor de similitud igual o superior al umbral establecido. De este proceso se obtiene un fichero (N3, RDF o XML) donde se listan los resultados obtenidos.

Esta métrica se clasifica entre los niveles léxico-sintáctico, consecuencia de que las métricas consideradas permiten tratar con los diferentes problemas recogidos en estas capas (caracteres especiales, acentuación, mayúsculas/minúsculas, errores tipográficos, etc.). Asimismo, dado que parte fundamental de esta heurística consiste en tratar con la información asociada al idioma esta heurística también trata problemas recogidos en el nivel pragmático (idioma y nombres oficiales y alternativos). Por otro lado, considerando las características de esta heurística, ésta se clasifica como independiente de dominio.

- ***URIlabel***. En esta heurística se comparan los elementos *localname* (*Qname*) y *label* de los diferentes recursos almacenados en los distintos SPARQL Endpoint o grafos de un SPARQL Endpoint determinado.

$$d\,(a_1, b_2 label_2) \geq \delta \Rightarrow a_1 = b_2$$

Para realizar dicha comparación, se procede a recuperar los recursos *localname* de un determinado SPARQL Endpoint o grafo.

```
Select distinct ?recurso
Where { GRAPH " + graph1 + " { ?recurso a <" + t + ">} }
order by ?recurso limit 100 offset " + offset
```

A continuación se listan todos los *labels* de cada recurso presentes en el SPARQL Endpoint o grafo contra el que pretende realizar la comparación.

```
Select ?label
Where { GRAPH " + graph2 + " { <" + r + ">
<http://www.w3.org/2000/01/rdf-schema#label> ?label } }
```

Donde `r` se refiere a las URI del SPARQL Endpoint o grafo de un SPARQL Endpoint determinado.
`?label` corresponde a todos los elemento *label* del recurso `r`.

Tras la recuperación de estos recursos la heurística desarrollada realiza la comparación de cada recurso del SPARQL Enpoint o `graph1` (*Qname* de cada URI) con los *labels* del SPARQL Enpoint o `graph2`. Si durante el proceso de comparación se encuentra un determinado grado de similitud, determinado por el umbral establecido, de forma inmediata se procede a comparar un nuevo recurso. Al igual que en las heurísticas anteriores, este proceso utiliza las métricas de similitud mencionadas con anterioridad, por tanto, los resultados van a depender del umbral establecido. Finalmente, de este proceso se obtiene un fichero donde se listan los resultados obtenidos en formato N3, RDF o XML.

Al igual que en los casos anteriores, esta métrica trata con problemas recogidos en los niveles léxico-sintáctico debido a las métricas consideradas. Asimismo, consecuencia de que esta heurística trata con nombres oficiales y alternativos asociados a una instancia también se encuentra dentro del nivel pragmático. Por otro lado, considerando las características de esta heurística, ésta se clasifica como independiente de dominio.

- **Latitud / Longitud** (**hWGS84**). Esta heurística trata la comparación de las coordenadas (latitud/longitud) de un fenómeno geográfico (puntual).

$$d\ (a_1\ lat/long_1, b_2\ lat/long_2) \geq\ distance_n \Rightarrow a_1 = b_2$$

Previamente al proceso de comparación se procede a la recuperación de la información de las coordenadas de los recursos recogidos en los distintos SPARQL Endpoint o grafos. Resulta habitual que esta información sea conforme con WGS84 *vocabulary*, ya que este vocabulario es un estándar del W3C para la representación de fenómenos geográficos de tipo puntual en el contexto de la Web Semántica. En la Tabla 8.14 se muestra un ejemplo de consulta para la recuperación de esta información.

Tabla 8.14. Recuperación de latitud y longitud

```
Select distinct ?recurso ?lat ?long ?tipo
Where { GRAPH " + graph1 + " { ?recurso a <" + t + ">. ?recurso
<http://www.w3.org/2003/01/geo/wgs84_pos#point> ?geo. ?geo a
?tipo. ?geo <http://www.w3.org/2003/01/geo/wgs84_pos#long>
?long. ?geo <http://www.w3.org/2003/01/geo/wgs84_pos#lat> ?lat}
}
order by ?recurso limit 100 offset " + offset;
```

Una vez recuperada esta información de los diferentes SPARQL Endpoint o grafos, esta heurística utiliza la librería GeoTools, que proporciona herramientas para el tratamiento de la información geoespacial, para la comparación de los diferentes valores de coordenadas. Esta comparación se realiza a través de la siguiente función:

$$p1(lat, long).isWithinDistance\ (p2(lat, long), distance)$$

Donde `p1(lat,long)` corresponde a los valores de latitud/longitud asociados a un fenómeno en el SPARQL Enpoint o grafo 1.
`p2(lat,long)` representa los valores de latitud/longitud de un fenómeno en el SPARQL Enpoint o grafo 2.
`Distance` corresponde a la longitud del radio desde el punto de interés (fenómeno).

Finalmente, de este proceso se obtiene un fichero (N3, RDF o XML) donde se listan los resultados obtenidos.

Dadas las características de esta heurística, ésta se clasifica en el nivel sintáctico (por el tratamiento de problemas de proximidad entre coordenadas y/o la presencia de errores) y como dependiente del dominio.

- **Geometría**. Según Kieler *et al.* (2009), la dificultad en la detección de similitudes para la geometría reside en el tratamiento de diferencias que son debidas a la generalización de la cartografía. Esta cuestión se complica si la diferencia de escala aumenta, ya que los conjuntos de datos con escala más pequeña contienen formas simplificadas geométricamente. Además, los objetos pueden ser eliminados o agregados a través de la generalización. Por tanto, los algoritmos de emparejamiento (*matching*) que tratan con comparaciones de geometría tienen importantes dificultades. Asimismo, las discrepancias geométricas en los límites de los objetos conlleva un incremento del número de posibles candidatos emparejados, consecuencia de que los objetos adyacentes pueden parcialmente superponerse.

  Esta heurística trata la comparación de la información geométrica recogida en fuentes de información geoespacial con escala de representación similar.

$$d\ (a_1\ geom_1, b_2\ geom_2) \geq\ distance_n \Rightarrow a_1 = b_2$$

  De forma previa al proceso de comparación se procede a la recuperación de la información geométrica de los distintos SPARQL Endpoint o grafos. Esta información puede ser conforme con el formato WKT o estar codificada a través de los elementos latitud/longitud. Esta diferencia requiere la especificación del tipo de formato existente en cada SPARQL Endpoint o grafo. En definitiva, las consultas a los conjuntos de datos se realizan utilizando el WKT (ver Tabla 8.15) o WGS84 (ver Tabla 8.14) cuando la información utiliza latitud/longitud.

  Una vez recuperada esta información de los diferentes SPARQL Endpoint o grafos, esta heurística utiliza la librería GeoTools para la comparación de geometrías y el tratamiento de dicha información conforme con las especificaciones del *Open Geospatial Consortium*.

Tabla 8.15. Recuperación de la geometría para el formato WKT

```
Select distinct ?recurso ?wkt
Where { GRAPH " + graph1 + " { ?recurso a <" + t + ">. ?recurso
<http://www.w3.org/2003/01/geo/wgs84_pos#geometry> ?geo. ?geo
<http://loki.cae.drexel.edu/~wbs/ontology/2004/09/ogc-gml#gid>
?wkt} }
order by ?recurso limit 100 offset " + offset
```

Para la comparación de la información geométrica sencilla (*puntual*) se utiliza la siguiente función:

$$p1.isWithinDistance\ (p2, distance)$$

Donde `p1` corresponde a la geometría recogida en el SPARQL Endpoint o grafo 1.
`P2` representa a la geometría almacenada en el SPARQL Endpoint o grafo 2.
`Distance` corresponde a la longitud del radio desde el punto de interés.

Para la comparación de geometrías más complejas (por ejemplo, *LineString* (curva) o

polígono) se utiliza la función espacial *buffer*, es decir, un polígono que enmarca el área de influencia al tomar una determinada distancia (umbral) desde un punto, línea u otro polígono.

$$g_3 = g_1.buffer\ (distance)$$

$$g_2.isWithin(g_3)$$

Donde g_1 corresponde a una geometría compleja, asociada a un fenómeno, recogida en un SPARQL Endpoint o grafo 1.
g_2 representa a una geometría compleja, asociada a un fenómeno, almacenada en un SPARQL Endpoint o grafo 2.
`Buffer(Distance)` corresponde a la operación que permite ampliar la geometría inicial con una distancia determinada, generando g_3 Esta operación es la base de la función utilizada para comparar las geometrías y verificar si g_1 está dentro de g_2 y viceversa.

Finalmente, de este proceso se obtiene un fichero (N3, RDF o XML) donde se listan los resultados obtenidos.

Esta métrica se clasifica en el nivel pragmático, debido a que trata con los problemas asociados a la geometría. Además, la información tratada en esta heurística la clasifica como dependiente del dominio.

- **Perímetro**. Esta heurística trata la identificación de instancias duplicadas a través del perímetro asociado a la geometría de un fenómeno abstraído en fuentes de información con escala de representación similar.

$$d\ (a_1\ perimeter_1, b_2\ perimeter_2) \geq\ distance_n \Rightarrow a_1 = b_2$$

Así, se procede a la recuperación de la información geométrica de los distintos fenómenos a través de los diferentes SPARQL Endpoint o grafos. Posteriomente, se utiliza la librería GeoTools para la extracción *on the fly* del valor perimetral asociado a una geometría.

```
public double perimeter(Geometry g) {
        AreaFunction af = new AreaFunction();
        Return af.getPerimeter(g);
    }
```

Esta métrica se clasifica en el nivel pragmático, por su consideración de los problemas asociados a los valores de atributos, y como dependiente del dominio.

- **Área**. Esta heurística trata la comparación de la información del área de una geometría de un fenómeno abstraído en fuentes de información con escala de representación similar.

$$d\ (a_1\ area_1, b_2\ area_2) \geq\ distance_n \Rightarrow a_1 = b_2$$

El procedimiento de esta heurística es el siguiente: En primer lugar se recupera la información geométrica de los distintos fenómenos a través de los diferentes SPARQL Endpoint o grafos. Seguidamente, utilizando la librería GeoTools se extrae *on the fly* el valor del área asociado a una geometría de tipo poligonal.

```
public double area(Geometry g) {
    AreaFunction af = new AreaFunction();
    return af.getArea(g);
}
```

Esta métrica se clasifica en el nivel pragmático, por su consideración de los problemas asociados a los valores de atributos, y como dependiente del dominio.

- **Combinación de heurísticas**. En general, como se afirma en (Hernández *et al.*, 1998) ningún atributo por sí mismo resulta suficiente para detectar las duplicidades de instancias. Por esta razón, el sistema desarrollado permite la combinación de las distintas heurísticas descritas (ver Figura 8.6). El objetivo de esta combinación es lograr un proceso de detección de duplicados que permita obtener unos resultados de mayor calidad y precisión. De este proceso se obtienen dos salidas: Por un lado, un fichero de duplicados (fichero de enlaces) donde se recogen las correspondencias establecidas entre recursos y, por otro lado, se genera un fichero de posibles duplicados (fichero de candidatos), cuya validación requiere un usuario experto.

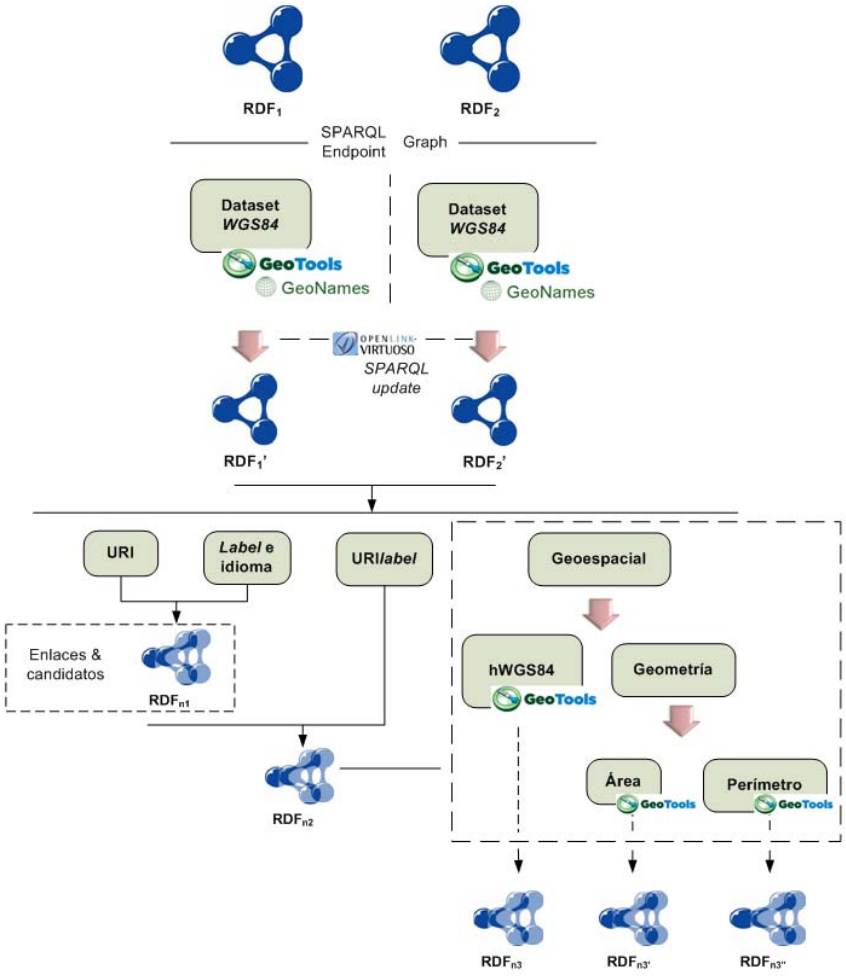


Figura 8.6. Visión general de la combinación de heurísticas

- **URI – *Label* e idioma**. Se recorren los resultados obtenidos de la aplicación de la heurística de comparación de URI y se comparan con cada uno de los resultados obtenidos de la heurística que trata con el elemento *label* y su idioma.

  Si se encuentra una coincidencia total entre los recursos estos pasan a considerarse duplicados y se anotan en el fichero de duplicados. Por el contrario,

si los recursos comparados no resultan iguales, estos quedan registrados en el fichero de posibles duplicados.

- **URI – URI*label***. Se recorren los resultados obtenidos de la heurística URI y se comparan con los resultados obtenidos de la heurística URI*label*. Si se encuentra una coincidencia total entre los recursos estos pasan a considerarse duplicados y se anotan en el fichero de duplicados. Si los recursos comparados no resultan iguales, estos quedan anotan en el fichero de posibles duplicados.

- **URI*label* – *Label* e idioma**. Se recorren los resultados obtenidos de la heurística URI*label* y se comparan con los resultados de *Label* e idioma. Si se encuentra una coincidencia total entre los recursos estos pasan a considerarse duplicados y se anotan en el fichero de duplicados. Por el contrario, si los recursos comparados no resultan iguales, estos se quedan registrados en el fichero de posibles duplicados.

- **URI – *Label* e idioma – URI*label* – hWGS84**. En este proceso se recorren y comparan todos los resultados obtenidos por la combinación de heurísticas URI, *Label* e idioma y URI*label* con los resultados obtenidos de la heurística hWGS84. Además, hWGS84 recorre los posbiles duplicados (resultados descartados) de anteriores heurísticas para hacer más exhaustivo el proceso de detección de duplicados. Al igual que en anteriores casos, si se encuentra una coincidencia total entre los recursos estos pasan a considerarse duplicados y se anotan en el fichero de duplicados. Si los recursos comparados no son iguales, estos se quedan registrados en el fichero de posibles duplicados.

- **URI – *Label* e idioma – URI*label* – geometría**. En este proceso se recorren y comparan todos los resultados obtenidos por las heurísticas URI, *Label* e idioma, URI*label* con los distintos resultados obtenidos de la heurística que trata con la geometría. Al igual que en anteriores casos de combinación de heurísticas, si se encuentra una coincidencia total entre los recursos estos pasan a considerarse duplicados y se anotan en el fichero de duplicados. Si los recursos comparados no resultan iguales, estos se quedan registrados en el fichero de posibles duplicados.

Una evaluación detallada de los diferentes enfoques heurísticos propuestos en este trabajo aparece en el Capítulo 9 de este trabajo.

8.4. Conclusiones

En este capítulo se ha descrito una visión general de la **duplicidad de instancias** en la información geoespacial y se han presentado los **problemas de las instancias** en el contexto de las fuentes de información utilizadas en este trabajo. Estos problemas han sido clasificados conforme a un enfoque basado en niveles o capas (léxica, sintáctica, semántica y pragmática) utilizado en distintas propuestas relacionadas con la interoperabilidad semántica. También, se han propuesto diversos **enfoques heurísticos** destinados a detectar **instancias duplicadas** en un mismo y entre múltiples conjuntos de datos. Estas heurísticas combinan información espacial y no-espacial. Además, se clasifican conforme a su dependencia e independencia del dominio y a los niveles o capas en los que se clasifican los problemas resueltos en las instancias.

El trabajo descrito en este capítulo contribuye a cubrir la carencia existente de una aproximación detallada de los **problemas en el ámbito de las instancias** durante el proceso de integración de información de múltiples fuentes de datos geoespaciales. Asimismo, la propuesta de una serie de técnicas para la detección de instancias duplicadas en conjuntos de datos geoespaciales supone una aproximación centrada en las características propias de la información geográfica.

Estas técnicas contribuyen al enriquecimiento de las escasas y limitadas propuestas para la detección de duplicados en el contexto de la información geoespacial.

Capítulo 9

Evaluación

Este capítulo describe la evaluación realizada para demostrar la pertinencia y validez de las propuestas presentadas en este trabajo. Los elementos sobre los que se lleva a cabo la evaluación son los siguientes: En primer lugar, se realiza una comparativa de los recursos ontológicos desarrollados en este trabajo con las ontologías existentes en el dominio hidrográfico (sección 9.1). Se continúa, en la sección 9.2, con la evaluación de la metodología propuesta para la integración de datos geoespaciales. Además, se evalúa la calidad de la transformación de las bases de datos originales a conjuntos de datos RDF y se recogen algunos datos resultantes de la publicación de *Linked Data* (sección 9.3). Por otro lado, se evalúan los diversos enfoques heurísticos desarrollados para la detección de instancias duplicadas en conjuntos de datos RDF (sección 9.4). La evaluación realizada sobre las distintas aportaciones de este trabajo se utiliza para la verificación de las hipótesis de partida, descrita en la sección 9.5. Finalmente, en la sección 9.6 se recogen unas conclusiones sobre los distintos aspectos tratados en este capítulo.

9.1. Recursos ontológicos

Como se refleja en el Capítulo 2 son diversas las ontologías existentes relacionadas con el dominio hidrográfico. Consecuencia de esta situación, en esta sección se procede a comparar y evaluar la cobertura de *hydrOntology* con otras ontologías del mencionado dominio. La comparación de estas ontologías se realiza conforme a la propuesta de comparación descrita en (Vilches-Blázquez *et al.*, 2009a). Esta propuesta está destinada a la evaluación del recubrimiento del dominio en ontologías del mismo área, tomando como base la definición de ontología recogida en (Gruber, 1993). Además, en esta comparativa se consideran aspectos adicionales. La Tabla 9.1 presenta los aspectos analizados de las distintas ontologías, donde se asocia el elemento correspondiente de la definición mencionada. A continuación se analizan los principales aspectos considerados en la evaluación de las ontologías tratadas:

Conceptualización. En primer lugar, en la Tabla 9.1 se refleja que únicamente los recursos ontológicos desarrollados en este trabajo se basan propiamente en metodologías extendidas para el desarrollo de ontologías. De esta manera, *hydrOntology* y la red de ontologías desarrollada utilizan METHONTOLOGY y NeOn *methodology*. Las restantes ontologías se desarrollan a partir de distintas técnicas de construcción de ontologías (por ejemplo: PhenomenOntology o Hydro) o en propuestas metodológicas propias (por ejemplo: la ontología de hidrología desarrollada por el Ordnance Survey).

Con respecto a su riqueza, los distintos valores (clases, propiedades (*object properties*) y atributos (*data properties*)) recogidos en las métricas de las ontologías (ver Tabla 9.1) son producto de la consideración de distintas fuentes de información con diferentes características de calidad y granularidad. Esta divergencia crea la diferencia de riqueza en los componentes de las ontologías. Así, entre los datos destaca el número de componentes de SWEET, aunque próximos a estos datos se sitúa la red de ontologías desarrollada en el contexto de este trabajo. En el otro extremo, aparecen las ontologías PhenomenOntology (delimitada al dominio hidrográfico) e hydrology con 21 clases. Con respecto a las propiedades, también destacan los valores ofrecidos por SWEET (541), así como los recogidos en PhenomenOntology e Hydro (no contemplan ninguna propiedad). Sin embargo, el número de atributos destaca en la red de *hydrOntology* (447 atributos), mientras *hydrology* y la ontología del *Ordnance Survey* presentan 5 y 4 atributos, respectivamente. Además, se destaca la ausencia de valores en la ontología de Tripathi *et al*. (2008), consecuencia de su no disponibilidad.

Compartición. Entre las ontologías analizadas predomina una visión del conocimiento hidrográfico centrada en aspectos globales, nacionales o en determinadas escalas de captura de la información. En este sentido, en la Tabla 9.1, se observa que las ontologías existentes contemplan una única fuente de información en su desarrollo (por ejemplo: Nomenclátor Geográfico Nacional, *Ordnance Survey Master Map Real World Objects Catalogue*, etc.). No obstante, se aprecia entre estas ontologías que no existe ningún recurso que trate el conocimiento hidrográfico desde una perspectiva transversal (de lo global a lo regional). Así, el desarrollo de la red de ontologías en torno a *hydrOntology* se presenta como el único recurso del dominio con estas características. En este sentido, la consideración de 20 fuentes de información distintas y 24 recursos ontológicos en la construcción de la red de ontologías hacen posible recoger esta visión transescalar del dominio hidrográfico.

Formalización y explicitación. Las diferentes ontologías se formalizan conforme a OWL (Dean *et al.*, 2004). Este lenguaje es una recomendación técnica del W3C Consortium y de este lenguaje se derivan 3 sublenguajes con expresividad incremental (OWL Lite, OWL DL y OWL Full) (McGuinness *et al.*, 2004). En la Tabla 9.1 se observa que únicamente Hydro e *hydrOntology* poseen una expresividad OWL Full y las restantes ontologías son OWL DL. Asimismo, se concreta la expresividad alcanzada por cada una de estas ontologías recogiendo en la tabla la codificación asociada a los operadores lógicos utilizados.

Entre los aspectos adicionales considerados en la evaluación del recubrimiento del dominio, se trata el análisis de los módulos incorporados en las distintas ontologías y el idioma en el que aparece el conocimiento descrito.

Módulos. Los módulos (ontologías) incorporados en las distintas ontologías corresponden a conocimiento relacionado con el dominio. De forma general, entre las ontologías analizadas se observa la utilización de diferentes módulos de la ontología SWEET (por ejemplo: hydrology). Asimismo, destaca la ontología del Ordnance Survey por incluir unos módulos distintos a los considerados en SWEET. No obstante, la red de ontologías desarrollada destaca por la adopción y adaptación de diferentes módulos provenientes de las distintas ontologías analizadas (por ejemplo: relaciones espaciales, mereología, unidades de medida, etc.), así como por recoger otros módulos propios (por ejemplo: Unidades administrativas, representación espacial, multilingüismo, etc.).

Idioma. Los recursos ontológicos desarrollados en el marco de este trabajo se caracterizan por contener un componente multilingüe del que carecen el resto de ontologías del dominio hidrográfico. Así, las ontologías desarrolladas contienen información en español, inglés, francés y catalán, mientras las restantes ontologías presentan información monolingüe en inglés o español.

Tabla 9.1. Comparativa de las ontologías hidrográficas existentes con los recursos ontológicos desarrollados

Ontologías	**Métricas**	**Fuentes de información**	**Metodología**	**Granularidad**	**Formalización**	**Módulos**	**Idioma**
(Gruber, 1993)	**Conceptualización**	**Compartición**	**Conceptualización**	**Compartición**	**Formalización y Explicitación**		
SWEET	4363 clases 541 propiedades 39 atributos	Ciencia del planeta Tierra (NASA)	-	Global	OWL DL (SHOIN(D))	200 ontologías modulares	@en
hydrology (Piasecki *et al*., 2005)	21 clases 28 propiedades 5 atributos	Ciencia e ingeniería hidráulica	-	-	OWL DL (ALOIF+(D))	Módulo de SWEET	@en
Ordnance Survey	239 clases 74 propiedades 4 atributos	Ordnance Survey Master Map Real World Objects Catalogue	Ordnance Survey's Conceptual ontology authoring method and Logical ontology authoring method	Nacional	OWL DL (SHOIQ(D))	*Spatial relations*, *Network relations*, *Language relations*, *Political geography* y *Mereology*	@en
Hydro	51 clases	Nomenclátor Geográfico Nacional	FCA	Nacional (1:25K - 1:5K)	OWL Full	-	@es
Phenomen Ontology	21 (117) clases - propiedades 114 atributos	BCN25	Ad hoc wrapper	Nacional (1:25.000)	OWL DL (AL (D))	-	@es
Tripathi *et al.*, 2008	-	SWEET	-	Global	OWL DL	Diversas ontologías de SWEET	@en
hydrOntology	**154 clases 47 propiedades 64 atributos**	**20 fuentes diferentes**	**METHONTOLOGY**	**Nacional-Regional-Local**	**OWL Full (SHOIN(D))**	**-**	**@es, en, fr, ca**
Red hydrOntology	**4143 clases 776 propiedades 447 atributos**	**20 fuentes distintas + 24 ontologías relacionadas**	**NeOn methodology**	**Desde global-hasta nivel regional-local**	**OWL DL**	**Relaciones espaciales, Mereología, División admin., Metadatos, Unidades de medida, Multilingüismo, Repres. espacial**	**@es, en, fr, ca**

9.2. Metodología de integración de datos geoespaciales

En esta sección se recoge la evaluación de la metodología para la integración de datos geoespaciales propuesta en este trabajo (Capítulo 5). La evaluación de la metodología propuesta está basada en dos aspectos. Por un lado, se evalúan aquellos principios de diseño considerados (condiciones necesarias y suficientes (sección 4.2)), capaces de ser comprobados de forma concreta y, por otro, se evalúa la aplicabilidad de la misma.

Las **condiciones necesarias**, definidas en (Paradela, 2001), son independientes del dominio de aplicación y comunes a cualquier metodología. A continuación se recoge la verificación de dichas condiciones por parte de la metodología de integración propuesta:

- *Eficiencia*. La metodología está descrita de forma detallada para que pueda ser utilizada de manera independiente por cualquier persona, sin posibilidad de ambigüedad en su proceso de desarrollo. Esta descripción utiliza diversas figuras y tablas que ayudan a clarificar su entendimiento. Asimismo, la definición del personal y su implicación en las diferentes tareas contribuye a conseguir esta condición de eficiencia.
- *Efectividad*. La metodología resuelve de manera efectiva los diferentes casos que se presentan en la instanciación de la misma para el dominio de la información hidrográfica. Prueba evidente de esta efectividad es el trabajo descrito en los Capítulos 4, 5, 6 y 7 de este trabajo.
- *Cuestiones sobre la esencia*. La metodología considera el "qué", "quién", "por qué", "cuándo" y "cómo" de cada tarea. A continuación, se describen estas cuestiones:

    - La metodología recoge, de forma general, *qué* se realiza en cada una de sus diferentes actividades. Esta descripción se concreta en las distintas tareas asociadas a cada actividad.
    - La descripción del personal implicado y la definición de sus roles, recogida en la sección 4.4, responde a la consideración de *quién*. Además, la especificación de los responsables en cada una de las tareas definidas precisa *cuándo* se requiere de su participación.
    - El *porqué* se concreta en la descripción del objetivo de cada una de las diversas actividades y tareas descritas.
    - Finalmente, la inclusión de diversas técnicas en la descripción de las diferentes tareas y la propuesta de múltiples herramientas para la realización de las mismas responde a *cómo* debe llevarse a cabo cada tarea.

- *Atendencia o conjunto de reglas metodológicas generales*. La metodología es conforme con las reglas metodológicas generales del Método Cartesiano de Descartes. Estas quedan reflejadas en:

    - *Evidencia*. Esta regla queda probada con las diferentes tareas que componen el estudio de viabilidad y el posterior análisis de fuentes de información realizado en la Actividad 2 de la metodología propuesta.
    - *Análisis*. La conformidad con esta regla queda reflejada en el hecho de que las diferentes actividades propuestas en la metodología recogen diversos problemas a solventar durante el proceso de integración de información. Dichas actividades se dividen en tantas tareas como resultan necesarias para la solución de dichos problemas. Así, por ejemplo, la Actividad 2 (*Análisis de fuentes*) contiene dos tareas destinadas al análisis y comparación de las fuentes de información, dos tareas para el análisis y comparación de los datos contenidos en las fuentes y, finalmente, una tarea designada a la revisión de los resultados con la participación de los expertos.
    - *Síntesis*. Como se evidencia en la regla anterior, las tareas que componen esta metodología dividen los problemas en pequeñas partes. Estos problemas se

analizan de forma particular y, con posterioridad, se sube el nivel de análisis mediante la comparación de resultados y/o un proceso de revisión-evaluación.

- *Finitud*. El proceso de desarrollo aparece concretado y detallado en el capítulo 4 de este trabajo. Con respecto al consumo de tiempo de cada tarea, éste es directamente dependiente de los miembros implicados en el proceso de desarrollo y de la pericia de los mismos.
- *Discernimiento*. La metodología de integración de datos geográficos está compuesta de un número concreto de actividades y tareas, por lo que esta condición está probada por el propio formato de la metodología.
- *Entorno*. Las consideraciones técnicas y tecnológicas que se realizan en esta metodología provoca que ésta sea clasificada dentro del grupo tecnológico.
- *Transparencia*. La metodología propuesta permite conocer, de forma rápida e intuitiva, cuál es la tarea que está siendo procesada, qué está siendo llevado a cabo, quién lo está llevando a cabo, etc., a través de las distintas figuras (recogen la visión general de la metodológica y/o la descomposición de tareas de las actividades propuestas) y tablas (detallan los productos de tareas, técnicas y participantes de cada actividad) incluidas en el Capítulo 4 de este trabajo. Asimismo, es posible profundizar en el proceso de desarrollo de la metodología con la descripción presente en el mencionado capítulo.

Las **condiciones suficientes**, como se refleja en el Capítulo 4, son específicas del dominio al que se aplica la metodología. En este caso, la evaluación de estas condiciones está fundamentada en la aplicación al dominio hidrográfico.

- *Depuración o curación de bases de datos*. La metodología está orientada a mejorar las bases de datos implicadas en el proceso de integración, así como a obtener pautas y lecciones de buenas prácticas tanto en el proceso de construcción de bases de datos como en la integración de fuentes de datos geoespaciales heterogéneas. De esta manera, este trabajo contribuye a la detección de diversos problemas en las bases de datos estudiadas. Entre estos problemas destaca la presencia de errores léxicos, ausencia de identificación del idioma asociada a las instancias, ausencia del tipo de fenómeno, presencia de fenómenos múltiples, presencia de múltiples instancias en un mismo campo, etc. Diversos ejemplos relativos a las distintas problemáticas expuestas, así como diferentes propuestas de cambios recomendadas se recogen en la Tarea 7.1 (sección 4.5.7). Además, este trabajo produce una serie de recomendaciones o buenas prácticas destinadas a la depuración de las bases de datos utilizadas en este trabajo (ver Anexo). A continuación se detallan las principales recomendaciones resultantes de este trabajo:

    - **Claves primarias**. Aunque este punto puede parecer trivial, es importante disponer de una clave primaria para identificar unívocamente un registro dentro de las tablas. Por ejemplo, NOMGEO y NGCE no disponen de claves primarias, por tanto, se recomienda añadir una nueva columna donde se introduzca la clave primaria las tablas de dichas bases de datos.
    - **Sistema de referencia**. Para el tratamiento de la información geoespacial resulta necesario conocer el sistema de referencia utilizado en la captura de los datos. Esta información, con frecuencia, no se recoge en la base de datos ni en su metadatos asociados. Así, se recomienda la incorporación de esta información en las tablas y/o en los metadatos asociadas a las bases de datos.
    - **Transformaciones de sistemas de referencia**. En ocasiones, las bases de datos que forman parte del proceso de integración poseen distintos sistemas de referencia. La integración de esta información conlleva la transformación y armonización de los sistemas de referencia. En este sentido, se recomienda la transformación de las bases de datos a WGS84 (EPSG:4326), consecuencia de

que *WGS84 vocabulary* es el único modelo/ontología existente para la representación gráfica de los datos en el contexto de la Web de los Datos.

- **Transformación de coordenadas**. Conforme a la armonización anterior, en algunos casos también resulta necesario la transformación de las coordenadas de formato Grados-Minutos-Segundos a Grados. Se recomienda realizar este cambio porque los campos latitud y longitud de *WGS84 vocabulary* requieren este formato.
- **Múltiples instancias**. En ocasiones, en un mismo registro de la base de datos aparecen múltiples instancias. Esta multiplicidad de información se debe a que en un mismo registro se recoge el nombre oficial y alternativo asociado a un fenómeno. Este hecho produce una mala identificación de datos. De esta manera, se recomienda dividir la información asociada a la instancia en tantas columnas como resulte necesario. Este hecho contribuirá a la publicación de datos más legibles por el usuario final.
- **Tipo de fenómeno**. En algunas de las bases de datos estudiadas no se define con claridad el tipo de fenómeno al que están asociadas las instancias. Así se pueden encontrar fenómenos de tipo "Puerto" sin distinción entre "puerto de mar" o "puerto de montaña". En este trabajo se recomienda realizar una clara y concisa definición de los fenómenos recogidos en una base de datos.
- **Idioma**. A pesar de que gran parte de las bases de datos estudiadas contemplan información multilingüe (castellano, catalán, euskera y gallego), resulta frecuente la ausencia de la información asociada al idioma de cada instancia. Por tanto, se recomienda que siempre que sea posible, se identifique el idioma de la instancia generada.

- *Dominio*. En este caso, la metodología propuesta se aplica al dominio de la información hidrográfica. La aplicación a este dominio produce como resultado el desarrollo de dos recursos ontológicos: una ontología individual (*hydrOntology*) y una red de ontologías en torno a *hydrOntology*. Además, se ha generado siete conjuntos de datos RDF con información hidrográfica (y sus características geométricas asociadas) del territorio nacional español. Estos conjuntos de datos han sido publicados conforme a los requisitos de *Linked Data*. Finalmente, se han desarrollados diferentes enfoques heurísticos (dependientes e independientes del dominio geográfico) para la detección de instancias duplicadas y el establecimiento de relaciones entre un mismo y diferentes conjuntos de datos.
- *Inspirada en prácticas existentes*. La metodología propuesta está fundamentada en diferentes propuestas metodológicas existentes en las áreas de ingeniería del software (por ejemplo, MÉTRICA v.3; Kruchten, 2000), ingeniería de conocimientos (por ejemplo, Gómez-Pérez *et al*., 1997; Schreiber *et al*., 1999), ingeniería ontológica (por ejemplo, Fernández-López *et al.*, 1997; Staab *et al.*, 2001; Gómez-Pérez *et al.*, 2003; etc.), *ontology matching* (Euzenat *et al.*, 2007) e ingeniería de datos. Esta consideración se realiza con el objetivo de aprovechar las buenas prácticas, facilitar su entendimiento y extender su utilización en diferentes organizaciones.
- *Independiente de herramientas*. De forma general, como se expone con anterioridad (sección 4.2), una metodología debe ser independiente de las tecnologías existentes. Por esta razón, esta metodología se desarrolla con independencia de las herramientas utilizadas por los diferentes actores implicados, limitándose, en ocasiones, a realizar un listado, no exhaustivo, de posibles herramientas a utilizar. Así, por ejemplo, para la tarea de descripción de esquemas (*Tarea 2.1*) se propone la utilización de herramientas como TOAD o SQuirreL SQL y para la tarea de creación de *mappings* (*Tarea 4.1*) se propone la utilización de herramientas como ODEMapster, D2R Server o Triplify.
- *Gran escala*. Esta metodología trata con diferentes tipos de recursos de información (catálogos, tesauros, diccionarios, ontologías, bases de datos, etc.) con objeto de llevar el proceso de integración a diferentes niveles. El tratamiento de esta diversidad de recursos de información se recoge en diversas secciones de este trabajo. Así, por

ejemplo, en el capítulo 5, se describe la utilización de diferentes recursos ontológicos y no ontológicos para el desarrollo de *hydrOntology* y de su red de ontologías y/o, en el Capítulo 6, se listan las múltiples bases de datos consideradas en este trabajo.

- *Flexibilidad*. La flexibilidad de esta metodología queda demostrada con el hecho de no limitarse a la utilización de ciertas herramientas para realizar su proceso de desarrollo, como se describe en la evaluación de la condición "independiente de herramientas". Asimismo, esta metodología no se limita a la consideración de aspectos técnicos, sino que en su proceso de desarrollo considera y se retroalimenta del conocimiento de expertos del dominio. La implicación del experto está presente, con diferente participación, en las diversas actividades propuestas, con excepción de la *Actividad 6*.
- *Almacenamiento, disponibilidad y mantenimiento de datos espaciales*. La metodología es consecuente con los principios INSPIRE en lo que respecta al almacenamiento y mantenimiento de las fuentes de información. Así, esta metodología promueve el almacenamiento y mantenimiento de las fuentes por sus propios productores y no realiza modificaciones directas sobre las fuentes, efectuando únicamente propuestas de depuración.
- *Interoperabilidad*. Esta metodología contribuye a alcanzar la interoperabilidad estructural –a través del desarrollo de una red de ontologías destinada a la armonización y explicitación semántica del conocimiento hidrográfico– y semántica –a través del establecimiento de *mappings* entre ontología-bases de datos y la detección de duplicados de instancia–. Asimismo, la utilización de los lenguajes RDF y OWL, recomendaciones del W3C, aportan el añadido de la interoperabilidad sintáctica. Estos lenguajes poseen mayor expresividad que XML, ampliamente utilizado en el contexto de la interoperabilidad geoespacial.

Con respecto a la aplicabilidad de la metodología, esta ha sido utilizada en el contexto del proyecto CENIT España Virtual. Asimismo, su aplicabilidad queda demostrada a través del caso de aplicación recogido en este trabajo (información del dominio hidrográfico), descrito en el Capítulo 4 (Estudio de viabilidad del proceso de integración (sección 4.5.1)), 5 (Desarrollo de recursos ontológicos de dominio en el contexto de la información hidrográfica), 6 (Generación de instancias y publicación de *Linked Data* en el dominio hidrográfico) y 7 (Detección de instancias duplicadas en conjuntos de datos heterogéneos).

9.3. Conjuntos de datos RDF y publicación de Linked Data

El proceso de generación de instancias, descrito en el Capítulo 7, se aplica sobre las siguientes fuentes de datos: NOMGEO, NGCE, BCN200, BTN25, EGM, Atlas e ICA. Una descripción detallada de estas fuentes se recoge en el Anexo.

Tabla 9.2. Número de recursos y tripletas generadas en cada fuente

Fuente	Número de recursos	Número de tripletas
NOMGEO	72	378.904
NGCE	26	23.022
BTN25	36	21.057.655
BCN200	48	7.364.990
EGM	29	7247
Atlas	23	4180
ICA	38	21832
Total	**272**	**28.857.830**

La generación de RDF se caracteriza por cubrir toda la información vinculada con el dominio hidrográfico presente en las mencionadas fuentes. La información presente en dichas fuentes tiene carácter multilingüe. Por este motivo se genera RDF en castellano, catalán, gallego y

euskera (idiomas presentes en las fuentes de información). En la Tabla 9.2 se recoge el número de tripletas RDF generado y el número de recursos hidrográficos en cada una de las fuentes. En los datos recogidos destaca el número total de tripletas generado (más de 28,8 millones de tripletas). Este número de tripletas está vinculado a un total de 272 recursos hidrográficos. También destaca el número de tripletas generado de las fuentes BTN25 (más de 21 millones de tripletas) y BCN200 (más de 7 millones de tripletas). Y, con respecto a los recursos, sobresalen por su número los pertenecientes a NOMGEO (72 recursos distintos).

Asimismo, la realización de este trabajo ha permitido la **publicación de información hidrográfica del territorio nacional español conforme a los principios de *Linked Data.*** Esto ha propiciado la inclusión de esta información en el contexto de *Linking Open Data* (ver Figura 9.1).

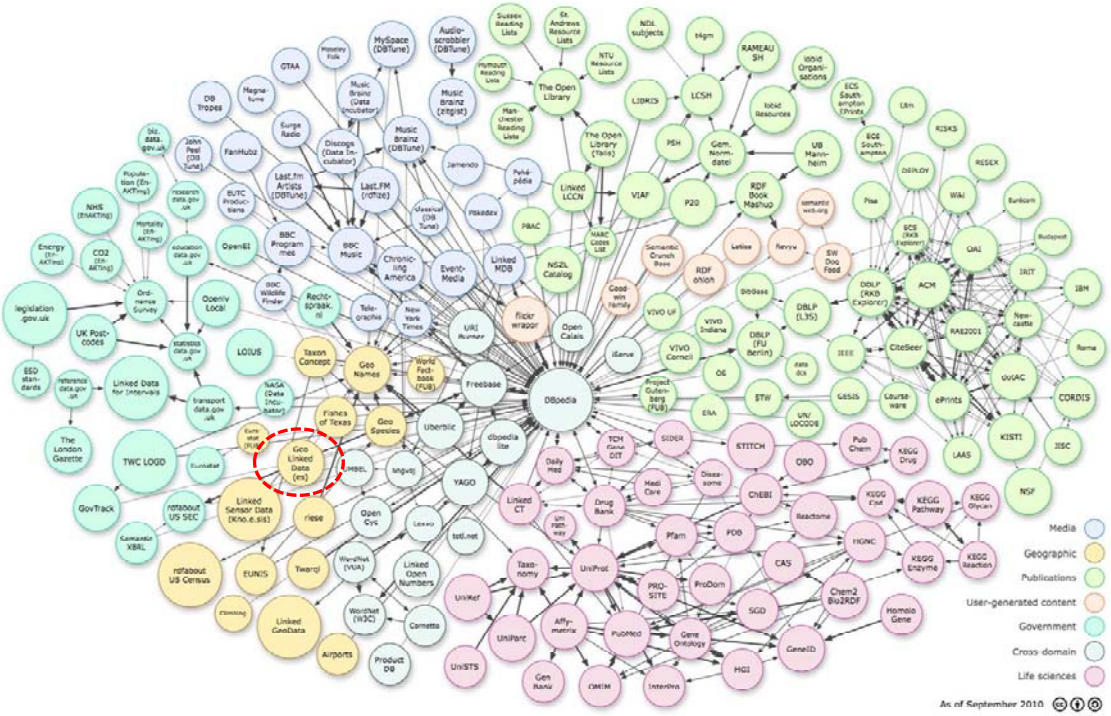


Figura 9.1. GeoLinked Data en la nube de *Linking Open Data*

La diferentes propuestas de *Linked Data* geoespacial (ver Tabla 9.4) demuestran que, incluso entre los principales proveedores (las iniciativas de Ordnance Survey y OpenStreetMap), únicamente se utiliza el elemento puntual (representados por un par de coordenadas (latitud, longitud) para la representación de la información geográfica. No obstante, a diferencia de las propuestas descritas, el desarrollo de este trabajo ha permitido tratar y representar este tipo de representación geométrica (punto) y geometrías más complejas (*LineString*).

Asimismo, se observa que uno de los *triple store* más utilizados es Virtuoso y que se usan diversas herramientas para la generación de RDF (Triplify, D2R, ODEMaspter y GEOMETRY2RDF). Con respecto a las ontologías utilizadas, se puede comprobar que éstas son diversas, aunque como denominador común todas las propuestas utilizan el vocabulario de WGS84 Geo Positioning. Los dominios de la información publicada también son diferentes (puntos de interés, topónimos e información general), aunque Ordnance Survey y GeoLinked Data tienen como punto en común la consideración de información de unidades administrativas. Finalmente, se observa que el idioma inglés es predominante, aunque GeoLinked Data, DBpedia y GeoNames proporcionan diferentes idiomas, por ejemplo: español, italiano, francés, etc. En la Tabla 9.4 los idiomas aparecen codificados utilizando ISO 639-1 (ISO 639-1, 2002).

Asimismo, la información publicada contribuye al **enriquecimiento de la Web de los Datos con información geoespacial en español**. Este enriquecimiento se demuestra con las cifras recogidas en la Tabla 9.3. Así, en Julio de 2010 el número de tripletas (recursos) RDF en español ascendía a un total de 79.461, esto suponía un 0,02% del total. Con la publicación del *Linked Data* geoespacial generado en el contexto de este trabajo el porcentaje de recursos a finales de agosto pasó a ser un 5,04 % (más de 21,5 millones de recursos). Este porcentaje de información sitúa al idioma español, en las fechas citadas, como el segundo idioma de la Web de los Datos, por encima de idiomas como el japonés (0,44%), francés (0,05%) o alemán (0,03%), aunque aún muy alejado de la información en inglés (más del 94% de los recursos publicados).

Tabla 9.3. Impacto de la iniciativa GeoLinked Data (Julio-Agosto 2010)

Idioma	Nº tripletas	Julio (%)
en	402.368.717	99,17
ja	1.881.981	0,46
fr	221.011	0,05
de	138.862	0,03
pl	102.769	0,03
it	89.190	0,02
es	**79.461**	**0,02**

Idioma	Nº tripletas	Agosto (%)
en	402.368.717	94,19
es	**21.548.329**	**5,04**
ja	1.881.981	0,44
fr	221.011	0,05
de	138.862	0,03
pl	102.769	0,02
it	89.190	0,02

Los idiomas de las tablas están codificados utilizando ISO 639 (ISO-639-1, 2002)

9.4. Detección de instancias duplicadas

Esta sección recoge los detalles del proceso de experimentación sobre las heurísticas desarrolladas para la detección de instancias duplicadas. Así, la sección se organiza de la siguiente manera: En la sección 9.4.1 se describen las principales características de los conjuntos de pruebas. Las medidas utilizadas en la evaluación se recogen en la sección 9.4.2. Finalmente, en la sección 9.4.3 se analizan los resultados obtenidos de la experimentación y se comparan con los derivados de la herramienta *Silk*.

9.4.1. Conjunto de pruebas

Para realizar la evaluación de los diferentes enfoques heurísticos se han tomado como referencia dos conjuntos de datos RDF (NOMGEO y NGCE). Una descripción de las características de las fuentes originales se recoge en el Anexo de este trabajo.

La experimentación realizada sobre los referidos conjuntos de datos se ciñe a las instancias de tipo "Río" presentes en ambos conjuntos. El volumen de instancias relacionadas con este tipo de fenómeno asciende a un total de 3.500 y 420 recursos, respectivamente. A pesar de la disparidad, estas fuentes capturan información del mismo ámbito y los resultados de la tarea 6.1 (Actividad 6) recogen la existencia de información duplicada entre ambas fuentes.

Junto a la existencia de información duplicada, la selección de estos conjuntos de datos viene determinada por las características de la información geométrica contenida (puntual - latitud/longitud). Estas características permitirán comparar los resultados obtenidos de las heurísticas desarrolladas en este trabajo con uno de los principales *framework* para la detección de duplicados en la Web de los Datos (*Silk*).

Tabla 9.4. Comparativa de los diferentes *Linked Data* geoespaciales con GeoLinked Data

Proveedor	Origen de los datos	*Triple store*	Información	Fuentes	generación RDF	Patrón URI	Vocabulario	Geometría	Idioma
Ordnance survey	Ordnance survey	Talis	Unidades Administrativas	Un nomenclátor (1:50,000) y un nomenclátor administrativo para Gran Bretaña	-	http://data.ordnancesurvey.co.uk/id/.+	Spatial Relations ontology, Administrative Geography ontology, WGS84 Geo Positioning, FOAF, OWL, y Gazetteer ontology	Punto	en
Linked GeoData	OpenStreetMap	Virtuoso	Puntos de interés	Base de datos OpenStreetMap (paradas de autobús, señales de tráfico, correos)	Triplify	http://linkedgeodata.org/triplify/+	LGD ontology, WGS84 Geo Positioning	Punto	en
GeoNames	GeoNames	RDF dump	Topónimos	Fuentes de información utilizadas por GeoNames	-	http://sws.geonames.org/id/	GeoNames ontology, WGS84 Geo Positioning	Punto	en, es, de, ca, nb, it, da, fr.
DBpedia	Wikipedia	Virtuoso	Información general	Wikipedia	D2R Server	http://dbpedia.org/resource/+	DBpedia ontology, WGS84 Geo Positioning	Punto	92 lang.
GeoLinked Data	**IGN-E**	**Virtuoso**	**Hidrografía, Unidades administrativas**	**Ver Anexo**	**ODEMapster y GEOMETRY2 RDF**	**http://geo.linkeddata.es/resource/(fuente)/FeatureType/+**	**Red hydrOntology**	**Punto & *LineString***	**es, ca, eu, & gl**

9.4.2. Medidas

Para la evaluación del proceso de detección de duplicados se han analizado los indicadores estándares para comprobar la eficacia de este proceso: precisión (*precision*) (Salton *et al.*, 1986), exhaustividad (*recall*) (Salton *et al.*, 1986) y las combinaciones de las anteriores conocidas como *Medida E* o Medida Armónica.

- **Precisión**: Medida que indica la proporción de correspondencias correctas del total de las descubiertas. A esta definición (Frakes *et al.,* 1992) añade que el resultado de esta operación está entre 0 y 1. Así, la recuperación perfecta es en la que únicamente se recuperan los documentos relevantes y, por lo tanto, tiene un valor de 1. La expresión matemática de la precisión en un escenario de detección de duplicados sería la siguiente:

$$P = \frac{n_{ddc}}{n_{dr}}$$

  Donde P corresponde al valor obtenido de la medida de precisión.
  n_{ddc} indica el número de duplicados detectados que son correctos.
  n_{dd} indica el número de duplicados recuperados.

- **Exhaustividad**: Medida que captura la proporción de correspondencias correctas descubiertas con respecto al total de las existentes. Si el resultado de este cálculo tiene como valor 1, se tendrá la exhaustividad máxima, ya que se habrá encontrado todo lo relevante que había en la fuente de información. La expresión matemática de la exhaustividad en un escenario de detección de duplicados sería la siguiente:

$$E = \frac{n_{ddc}}{n_{dc}}$$

  Donde E corresponde al valor obtenido de la medida de exhaustividad.
  n_{ddc} indica el número de duplicados detectados que son correctos.
  n_{dc} indica el número de duplicados correctos.

- **Medida E**: La precisión y la exhaustividad están compensadas, por lo que un sistema con una exhaustividad muy alta pero con baja precisión y viceversa no resultaría adecuado. Para evitar este tipo de situaciones se proponen funciones que combinan ambas métricas. La medida E es una función que combina las dos anteriores y toma valores en el intervalo [0,1]. Esta medida permite además ponderar la prioridad de la precisión sobre la exhaustividad o viceversa. La importancia relativa de una con respecto a otra se indica mediante un parámetro constante b. Su expresión matemática es la siguiente:

$$E = \frac{1 + b^2}{\frac{b^2}{E} + \frac{1}{P}}$$

  Donde valores de b > 1 dan prioridad a P sobre E y valores de b < 1 dan prioridad a E sobre P. En el caso en que b = 1, ambas métricas se ponderan por igual y la función toma el nombre de *medida f* o medida armónica.

9.4.3. Experimentación

Esta sección recoge los detalles del proceso de experimentación de las distintas heurísticas desarrolladas. Además, como elemento adicional se realiza una comparación de los resultados

con los ofrecidos por el *framework Silk*. Como se menciona en la sección 3.3, esta comparación con *Silk* se produce por ser la única propuesta en el contexto de la Web de los Datos que considera entre sus métricas un aspecto propio de la información geoespacial.

9.4.3.1. Consideraciones previas

En la evaluación de las heurísticas desarrolladas se consideran los siguientes aspectos:

Consideraciones generales

- El tratamiento de los *label* en esta experimentación se limita a aquellos recursos relacionados con el idioma español.
- En la evaluación de los resultados se tiene como referencia la detección de instancias duplicadas realizada por un experto de dominio (*gold standard*). Esta revisión entre las fuentes seleccionadas detecta un total de 401 instancias duplicadas (recogiendo instancias en castellano, catalán, gallego y eukera). Asimismo, al considerar únicamente instancias en castellano la cantidad de duplicados se reduce a 382 instancias.

Enfoques heurísticos

- En las heurísticas relacionadas con información independiente del dominio (URI, *Label* e idioma y URI*label*) se establecen como referencia los umbrales 0,99-0,96.
- Para las métricas utilizadas en las heurísticas independientes del dominio, concretamente aquellas métricas categorizadas como ***similitud*** (métricas que se caracterizan por ofrecer valores de similitud entre 0 y 1. Para más detalles ver la sección 8.3.2.2 - métricas de similitud), se han obtenido distintos valores en función de que los umbrales establecidos se alcancen en distinto número de métricas (≥2, ≥3 y ≥4).
- Para las heurísticas relacionadas con información dependiente del dominio, concretamente en este experimento se utiliza la heurística hWGS84, se aplican los umbrales 1, 3 y 5 kilómetros.
- El análisis de los datos resultantes producto de la experimentación se centra en los resultados recogidos en los ficheros de enlaces.

Silk

- Se utilizan las métricas Jaro, jaroWinkler, Levenshtein y qGrams para medir la similitud entre *labels*. En estas métricas se establecen como referencia los umbrales 0,99-0,96 y en la combinación de estas métricas se utiliza el operador de agregación “max”.
- En ocasiones, ante la ausencia de detección de instancias se establecen umbrales inferiores en las métricas que tratan con *labels*. En estos casos, los umbrales establecidos son: 0,90; 0,80 y 0,70.
- Se utiliza la métrica wgs84 para medir la distancia entre dos puntos. En esta métrica se aplican los umbrales 1, 3 y 5 km.
- Para la combinación de métricas de similitud entre cadenas y distancia geográfica entre puntos se utiliza el operador de agregación *average*.

9.4.3.2. Análisis de resultados

La detección de duplicados en un mismo conjunto de datos se realiza mediante la aplicación de un enfoque heurístico semiautomático (sección 8.3.2.1) y la supervisión de expertos de dominio. Este enfoque, aplicado sobre aquellas instancias que poseen más de un valor de latitud/longitud, conlleva:

- La distinción de instancias que poseen un nombre similar pero diferente localización (*owl:differentFrom*).
- El establecimiento de una relación de similitud entre aquellas instancias que presenten características comunes (*owl:sameAs*).

La aplicación de este enfoque heurístico sobre los conjuntos de datos originales detecta 10 y 394 instancias de tipo "Río" (en NGCE y NOMGEO, respectivamente) con dos o más valores de latitud/longitud asociados (ver Tabla 9.5).

Tabla 9.5. Detección de duplicados en un mismo conjunto de datos

	Datos originales	***≥2 lat/long***	***owl:sameAs***	***owl:differentFrom***	**Datos actualizados**
NGCE	420	10	4	6	426
NOMGEO	3500	394	309	85	3585

La supervisión realizada por el experto de dominio sobre los datos obtenidos identifica 4 relaciones de tipo *owl:sameAs* y 6 de tipo *owl:differentFrom* en los datos de NGCE. Mientras, en NOMGEO se identifican 309 relaciones de tipo *owl:sameAs* y 85 de tipo *owl:differentFrom*. El establecimiento de estas relaciones conlleva la aparición de nuevas instancias en los conjuntos de datos. De esta manera, en NGCE se pasa de 420 a 426 instancias, mientras el número de instancias en NOMGEO pasa de 3500 a 3585 instancias de tipo "Río". En la Tabla 9.6 se muestra un ejemplo con el resultado de la detección de duplicados en un mismo conjunto de datos.

Tabla 9.6. Ejemplo de diferenciación de instancias duplicadas

Alhama, Río at geo.linkeddata.es
http://geo.linkeddata.es/resource/R%C3%ADo/Alhama%2C%20R%C3%ADo_Cervera%20del%20R%C3%ADo%20Alhama

owl:differentFrom
* <http://geo.linkeddata.es/resource/R%C3%ADo/Alhama%2C%20R%C3%ADo_Corumbela>

owl:sameAs
* <http://geo.linkeddata.es/resource/R%C3%ADo/Alhama%2C%20R%C3%ADo_Cintru%C3%A 9nigo>
* <http://geo.linkeddata.es/resource/R%C3%ADo/Alhama%2C%20R%C3%ADo_Villarraso>

geo:geometry
* <http://geo.linkeddata.es/resource/wgs84/42.0144274433984_-1.94388130850726>

rdfs:label
* Alhama, Río (es)

rdf:type
* geoes:R%C3%ADo

Análisis individualizado

URI. De forma general, en los valores resultantes de la aplicación de esta heurística (Tabla 9.7) se observa que el mayor número de instancias duplicadas se obtiene al considerar ≥ 2 métricas (394, 397, 397 y 406 instancias para los umbrales 0,99 - 0,96, respectivamente). Esta diferencia también se observa al comprobar el número de instancias duplicadas correctamente detectadas conforme a los resultados ofrecidos por la evaluación del experto (339, 340, 340 y 341 respectivamente para los umbrales mencionados). No obstante, la consideración de ≥ 2 métricas conlleva la generación de un elevado número de errores (entre 55 y 65) en los resultados. Este hecho provoca que los valores de precisión se sitúen en un 0,86 (con la excepción del umbral 0,96, donde esta se sitúa en un 0,84) y la exhaustividad en un 0,85. Esta similitud de medidas sitúa a la medida *f* en un 0,85.

Por otro lado, la consideración de ≥ 3 o ≥ 4 métricas conlleva la obtención de valores uniformes, aunque se obtiene un número inferior de instancias duplicadas, comparadas con los valores obtenidos en ≥ 2. Sin embargo, los datos obtenidos arrojan un escaso número de errores (entre 5 y 12 dependiendo del umbral establecido). Así, la precisión se sitúa en un 0,98 y el valor de la

exhaustividad oscila entre 0,74 y 0,77. La combinación de los valores anteriores sitúa a la medida *f* entre 0,84 y 0,86.

En este caso, los resultados obtenidos por este enfoque heurístico no se comparan con *Silk* debido a que dicho *framework* no trata con URI.

Tabla 9.7. Resultados de la detección de duplicados a través de URI

Umbral	Métricas	Total	✓	✖	Precisión	Exhaustividad	Medida *f*
	≥ 2	394	339	55	0,86	0,85	0,85
0,99	**≥ 3**	302	297	5	0,98	0,74	0,84
	≥ 4	302	297	5	0,98	0,74	0,84
	≥ 2	397	340	57	0,86	0,85	0,85
0,98	**≥ 3**	302	297	5	0,98	0,74	0,84
	≥ 4	302	297	5	0,98	0,74	0,84
	≥ 2	397	340	57	0,86	0,85	0,85
0,97	**≥ 3**	311	305	6	0,98	0,76	0,86
	≥ 4	302	297	5	0,98	0,74	0,84
	≥ 2	406	341	65	0,84	0,85	0,85
0,96	**≥ 3**	320	308	12	0,96	0,77	0,85
	≥ 4	302	297	5	0,98	0,74	0,84

***Label* e idioma.** En este enfoque se procede a analizar los resultados de la detección de duplicados, en las fuentes mencionadas, partiendo de *labels* cuyo idioma sea español. Por tanto, esta delimitación de la experimentación conlleva una reducción en los resultados con respecto a los obtenidos al tratar con URI. Asimismo, los resultados obtenidos con esta heurística se comparan con los ofrecidos por *Silk* (ver Tabla 9.8 y Tabla 9.9).

Como se observa en la Tabla 9.8, la heurística *Label* e idioma recoge un mayor número de instancias detectadas que las obtenidas por *Silk* (ver Tabla 9.9). Asimismo, se comprueba que *Silk* no ofrece resultados para las métricas Levenshtein y qGrams en los umbrales 0,98-0,96.

Tabla 9.8. Resultados de la detección de duplicados utilizando *labels* y su idioma

Umbral	Métricas	Total	✓	✖	Precisión	Exhaustividad	Medida *f*
	≥ 2	376	302	74	0,80	0,79	0,80
0,99	**≥ 3**	311	284	27	0,91	0,74	0,82
	≥ 4	311	284	27	0,91	0,74	0,82
	≥ 2	378	304	74	0,80	0,80	0,80
0,98	**≥ 3**	311	284	27	0,91	0,74	0,82
	≥ 4	311	284	27	0,91	0,74	0,82
	≥ 2	378	304	74	0,80	0,80	0,80
0,97	**≥ 3**	322	290	32	0,90	0,76	0,82
	≥ 4	311	284	27	0,91	0,74	0,82
	≥ 2	386	305	81	0,79	0,80	0,79
0,96	**≥ 3**	328	290	38	0,88	0,76	0,82
	≥ 4	311	284	27	0,91	0,74	0,82

Los resultados obtenidos sitúan el valor de precisión y exhaustividad de esta heurística en torno al 0,80 al utilizar ≥ 2 métricas. Sin embargo, al utilizar ≥ 3 o ≥ 4 métricas la precisión alcanza un 0,91 y la exhaustividad se sitúa en un 0,76. En *Silk* la precisión oscila entre 0,96 y 0,93. En

cambio, los valores de exhaustividad resultan inferiores a los recogidos por el enfoque heurístico desarrollado, alcanzando un máximo de 0,71 (en el umbral 0,96 para el proceso combinado de métricas y jaroWickler).

Los valores de precisión y exhaustividad hacen que la medida *f* en el enfoque heurístico alcance valores de 0,80 al considerar ≥ 2 métricas y 0,82 al considerar ≥ 3 y ≥ 4 métricas, mientras en *Silk* oscila entre 0,74 (umbral 0,99) y 0,81 (combinación de métricas en los umbrales 0,97 y 0,96).

Tabla 9.9. Resultados de la detección de duplicados mediante las métricas de *Silk*

Umbral	Métricas	Total	✓	✖	Precisión	Exhaustividad	Medida *f*
	Jaro	241	231	10	0,96	0,60	0,74
	jaroWinkler	241	230	11	0,95	0,60	0,74
0,99	**Levenshtein**	241	231	10	0,96	0,60	0,74
	qGrams	241	231	10	0,96	0,60	0,74
	combinación	241	232	9	0,96	0,61	0,74
	Jaro	244	234	10	0,96	0,61	0,75
	jaroWinkler	272	260	12	0,96	0,68	0,80
0,98	**Levenshtein**	-	-	-	-	-	-
	qGrams	-	-	-	-	-	-
	combinación	272	262	10	0,96	0,69	0,80
	Jaro	264	254	10	0,96	0,66	0,79
	jaroWinkler	279	265	14	0,95	0,69	0,80
0,97	**Levenshtein**	-	-	-	-	-	-
	qGrams	-	-	-	-	-	-
	combinación	279	267	12	0,96	0,70	0,81
	Jaro	280	265	15	0,95	0,69	0,80
	jaroWinkler	291	270	21	0,93	0,71	0,80
0,96	**Levenshtein**	-	-	-	-	-	-
	qGrams	-	-	-	-	-	-
	combinación	291	272	19	0,93	0,71	0,81

En definitiva, los datos analizados demuestran que el enfoque heurístico *Label* e idioma es capaz de detectar un mayor número de instancias duplicadas, tanto en números totales como en duplicados correctamente detectados, que las ofrecidas por *Silk*. Asimismo, esto se refleja con los valores ofrecidos por la medida *f*, donde se observa que el mencionado enfoque heurístico ofrece un valor más elevado en los distintos umbrales fijados.

URI*label*. La heurística que combina URI (*Qname*) y *label* (ver Tabla 9.10) conlleva un incremento del número de instancias duplicadas detectadas, en relación a los valores obtenidos por la heurística que trata únicamente con los *labels* y su idioma. Además, los errores (entre 22 y 76 según la evaluación del experto) son inferiores a los detectados por la heurística *Label* e idioma.

En los datos derivados de esta heurística se observa que los valores de precisión resultan inferiores al considerar ≥2 métricas que los obtenidos al considerar ≥3 y ≥4 métricas. Sin embargo, los valores de exhaustividad son mayores al considerar ≥2 métricas, alcanzando valores de 1. En la compensación de estas medidas (precisión y exhaustividad) destaca el valor de ≥3 métricas para todos los umbrales (0,93). No obstante, los valores de la medida *f* oscilan entre 0,91-0,93.

Tabla 9.10. Resultados de la detección de duplicados mediante la comparación de URI y *labels*

Umbral	Métricas	Total	✓	✖	Precisión	Exhaustividad	Medida *f*
0,99	**≥ 2**	446	380	66	0,85	0,99	0,92
	≥ 3	370	348	22	0,94	0,91	0,93
	≥ 4	370	347	23	0,94	0,91	0,92
0,98	**≥ 2**	449	381	68	0,85	1,00	0,92
	≥ 3	370	348	22	0,94	0,91	0,93
	≥ 4	370	347	23	0,94	0,91	0,92
0,97	**≥ 2**	449	381	68	0,85	1,00	0,92
	≥ 3	383	356	27	0,93	0,93	0,93
	≥ 4	370	347	23	0,94	0,91	0,92
0,96	**≥ 2**	458	382	76	0,83	1,00	0,91
	≥ 3	392	359	33	0,92	0,94	0,93
	≥ 4	370	347	23	0,94	0,91	0,92

En este caso, los resultados obtenidos por este enfoque heurístico no se comparan con *Silk* debido a que dicho *framework* no permite la combinación entre URI y *labels*.

hWGS84. Esta heurística trata de detectar duplicados tomando como referencia una distancia sobre los valores de latitud/longitud asociados a las instancias. Como se mencionó con anterioridad, en esta experimentación se utilizan los siguientes umbrales: 1, 3 y 5 km.

En la Tabla 9.11 se comprueba que el número de instancias detectadas por hWGS84 resulta superior a las obtenidas por *Silk*. No obstante, los valores de precisión de *Silk* se sitúan por encima de los obtenidos con hWGS84 consecuencia del equilibrio entre escasos resultados e instancias correctamente detectadas (por ejemplo, en el umbral 1km *Silk* detecta 1 instancia, mientras en hWGS84 se obtienen 17). Sin embargo, la exhaustividad con hWGS84 alcanza un valor de 0,26, mientras en *Silk* el valor máximo se sitúa en 0,09. La disparidad de resultados entre los valores de precisión y exhaustividad obtenidos para estos enfoques sitúan los valores de la medida *f* de hWGS84 por encima de los valores de *Silk*. Así, el valor superior de la medida *f* es de 0,30 en hWGS84 y 0,16 en *Silk* (umbral 5 km), mientras el valor inferior se sitúa en 0,08 (hWGS84) y 0,00 (*Silk*) para el umbral de 1 kilómetro.

Tabla 9.11. Resultados de la detección de duplicados mediante hWGS84 (heurísticas) y wgs84 (*Silk*)

	***Heurísticas* (hWGS84)**						***Silk* (wgs84)**					
Umbral (km)	**Total**	✓	×	***P***	***E***	***f***	**Total**	✓	×	***P***	***E***	***f***
1	25	17	8	0,68	0,04	0,08	1	1	-	1,00	0,00	0,00
3	131	65	66	0,50	0,16	0,24	32	17	15	0,53	0,04	0,08
5	301	105	196	0,35	0,26	0,30	64	37	27	0,58	0,09	0,16

A modo de resumen, la aplicación individualizada de las distintas heurísticas sitúan los resultados más destacados, conforme a la compensación ofrecida por la medida *f*, en: 0,86 (≥ 3 métricas en el umbral 0,97) para la heurística que trata con las URI; 0,82 (≥ 3 y ≥ 4 métricas en los umbrales 0,99-0,96) en *Label* e idioma; 0,93 (≥ 3 métricas en el umbral 0,99-0,96) en URI*label* y 0,30 en hWGS84 (umbral 5 km). Asimismo, los resultados más destacados ofrecidos por *Silk* se sitúan en 0,81 en el tratamiento de *labels* (combinación de métricas en los umbrales 0,97 y 0,96), mientras en wgs84 alcanza un 0,16 (umbral 5 km).

El análisis individualizado de las heurísticas refleja que los resultados obtenidos son dependientes de las características consideradas en cada heurística. Así, por ejemplo, en la

heurística que trata con las URI (nombre de instancias) los resultados van a depender únicamente de la similitud de las cadenas (*Qname*). Por tanto, la consideración de URI en la detección de duplicados puede conllevar la detección de falsos duplicados. Asimismo, la consideración de la similitud de valores de latitud/longitud también puede derivar en falsos positivos por la proximidad de otros ríos (por ejemplo, la proximidad de afluentes al cauce un río principal). Estos hechos manifiestan que la detección de instancias duplicadas conlleva la combinación de aspectos dependientes e independientes del dominio (descrito en el Capítulo 8). Por tanto, a continuación se realiza un análisis de la combinación de las distintas heurísticas desarrolladas en este trabajo.

Análisis combinado

URI, *label* e idioma y URI*label*. La combinación de las heurísticas que tratan con la información independiente del dominio o no espacial, (URI, *label* e idioma y URI*label*) presenta los resultados recogidos en la Tabla 9.12. Esta combinación descubre correctamente un número de instancias duplicadas que oscila entre 290 y 256. Sin embargo, también destaca el número de errores que aparecen al considerar ≥2 métricas, respecto a los resultados obtenidos por ≥3 y ≥4 métricas.

Estos resultados provocan que los valores de precisión al considerar ≥2 métricas resulten inferiores a los obtenidos considerando ≥3 y 4 métricas. No obstante, la exhaustividad logra sus valores más elevados cuando se consideran ≥ 2 métricas (situándose en un valor promedio de 0,76). La medida *f* obtenida en esta combinación de heurísticas alcanza un promedio de 0,80; situándose su valor máximo en 0,81 (≥ 3 métricas en el umbral 0,97).

Tabla 9.12. Resultados de la detección de duplicados mediante la combinación de las heurísticas URI, *Label* e idioma y URI*label*

Umbral	Métricas	Total	✓	✖	Precisión	Exhaustividad	Medida *f*
	≥ 2	340	289	51	0,85	0,76	0,80
0,99	**≥ 3**	261	256	5	0,98	0,67	0,80
	≥ 4	261	256	5	0,98	0,67	0,80
	≥ 2	342	289	53	0,85	0,76	0,80
0,98	**≥ 3**	261	256	5	0,98	0,67	0,80
	≥ 4	261	256	5	0,98	0,67	0,80
	≥ 2	342	289	53	0,85	0,76	0,80
0,97	**≥ 3**	268	262	6	0,98	0,69	0,81
	≥ 4	261	256	5	0,98	0,67	0,80
	≥ 2	350	290	60	0,83	0,76	0,79
0,96	**≥ 3**	274	262	12	0,96	0,69	0,80
	≥ 4	261	256	5	0,98	0,67	0,80

Los resultados obtenidos por la combinación de estos enfoques heurísticos no se pueden comparar con *Silk* debido a que en dicho *framework* no es posible realizar una combinación similar de métricas.

URI, *label* e idioma, URI*label* y hWGS84. La aplicación de esta combinación de heurísticas y métricas en Silk permite comprobar que conforme se incrementa el umbral de distancia geográfica establecido (en hWGS84 y wgs84) se detectan un mayor número de instancias duplicadas (ver Tabla 9.13 y Tabla 9.14). A pesar de las excepciones en el umbral de 5 km, este incremento también está presente en el número de instancias correctamente identificadas. Con respecto al número de errores o falsos positivos, en la combinación de heurísticas este permanece relativamente estable en los umbrales de 1 y 3 km, mientras en el umbral de 5 km el

Tabla 9.13. Resultados de la detección de duplicados mediante la combinación de las heurísticas URI, *Label* e idioma, URI*label* y hWGS84

		URI, *Label* e idioma, URI*label* y hWGS84 (1km)						**URI, *Label* e idioma, URI*label* y hWGS84 (3km)**						**URI, *Label* e idioma, URI*label* y hWGS84 (5km)**					
Umbral métricas	**Nº métricas**	**Total**	**✓**	**×**	***P***	***E***	***f***	**Total**	**✓**	**×**	***P***	***E***	***f***	**Total**	**✓**	**×**	***P***	***E***	***F***
0,99	**≥ 2**	342	289	53	0,85	0,76	0,80	349	298	51	0,85	0,78	0,82	353	276	77	0,78	0,72	0,75
	≥ 3	264	258	6	0,98	0,68	0,80	271	266	5	0,98	0,70	0,81	276	270	6	0,98	0,71	0,82
	≥ 4	264	260	4	0,98	0,68	0,80	271	267	4	0,99	0,70	0,82	276	238	38	0,86	0,62	0,72
0,98	**≥ 2**	344	289	55	0,84	0,76	0,80	351	298	53	0,85	0,78	0,81	355	302	53	0,85	0,79	0,82
	≥ 3	264	258	6	0,98	0,68	0,80	271	266	5	0,98	0,70	0,81	276	270	6	0,98	0,71	0,82
	≥ 4	264	260	4	0,98	0,68	0,80	271	267	4	0,99	0,70	0,82	276	238	38	0,86	0,62	0,72
0,97	**≥ 2**	344	289	55	0,84	0,76	0,80	351	298	53	0,85	0,78	0,81	355	302	53	0,85	0,79	0,82
	≥ 3	271	264	7	0,97	0,69	0,81	278	272	6	0,98	0,71	0,82	284	277	7	0,98	0,73	0,83
	≥ 4	264	260	4	0,98	0,68	0,80	271	267	4	0,99	0,70	0,82	276	238	38	0,86	0,62	0,72
0,96	**≥ 2**	352	292	60	0,83	0,76	0,80	359	299	60	0,83	0,78	0,81	363	303	60	0,83	0,79	0,81
	≥ 3	277	264	13	0,95	0,69	0,80	284	271	13	0,95	0,71	0,81	290	277	13	0,96	0,73	0,82
	≥ 4	264	260	4	0,98	0,68	0,80	271	267	4	0,99	0,70	0,82	276	238	38	0,86	0,62	0,72

Tabla 9.14. Resultados de la detección de duplicados con *Silk* mediante la combinación de métricas de similitud de cadenas y distancia geográfica

	***Label* & wgs84 (1km)**						***Label* & wgs84 (3km)**						***Label* & wgs84 (5km)**					
Umbral métricas	**Total**	**✓**	**×**	***P***	***E***	***f***	**Total**	**✓**	**×**	***P***	***E***	***f***	**Total**	**✓**	**×**	***P***	***E***	***f***
0,99	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0,98	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0,97	-	-	-	-	-	-	-	-	-	-	-	-	1	1	-	1,00	0,00	0,01
0,96	-	-	-	-	-	-	-	-	-	-	-	-	3	3	-	1,00	0,01	0,02
0,9	-	-	-	-	-	-	5	5	-	1,00	0,01	0,03	13	13	-	1,00	0,03	0,07
0,8	3	3	-	1,00	0,01	0,02	20	18	2	0,90	0,05	0,09	43	38	5	0,88	0,10	0,18
0,7	7	7	-	1,00	0,02	0,04	38	32	6	0,84	0,08	0,15	67	55	12	0,82	0,14	0,24

número de errores se incrementa considerablemente.

En Silk, como se muestra en la Tabla 9.14, la combinación de las distintas métricas (Jaro, JaroWinkler, Levenshtein, qGrams y wgs84) únicamente ofrece resultados en los umbrales 0,97 y 0,96 (umbral 5km). Los datos obtenidos en dichos umbrales resultan escasos. Por este motivo, se amplian los umbrales establecidos en esta experimentación en un intento de conseguir un mayor número de resultados. En los datos obtenidos se identifica un comportamiento similar al identificado en la combinación de heurísticas, a pesar del escaso número de resultados obtenidos en los distintos umbrales establecidos.

Con respecto a la medida de precisión, la combinación heurística ofrece valores inferiores al considerar ≥2 métricas y superiores cuando se tratan ≥3 y ≥4 métricas. Sin embargo, los valores de exhaustividad ofrecen resultados superiores en ≥2 métricas, aunque en el umbral 0,99 (5km) el valor resulta equiparable al obtenido en ≥3 métricas. En *Silk* la precisión resulta elevada en aquellos umbrales donde resultan escasas las instancias detectadas. Sin embargo, en aquellos umbrales donde el número de instancias se incrementan los valores de precisión resultan inferiores a los obtenidos por la combinación de heurísticas. La exhaustividad obtenida en *Silk* es considerablemente inferior a los datos obtenidos por las heurísticas. Esta desigualdad entre valores también se recoge en los derivados de la medida *f*.

En definitiva, los datos recogidos en la Tabla 9.13 y Tabla 9.14 muestran los beneficios aportados en el proceso de detección de instancias duplicadas por parte de la combinación de heurísticas desarrolladas en el contexto de este trabajo frente a los resultados obtenidos por *Silk*. Incluso, ampliando los umbrales establecidos en esta experimentación, los resultados obtenidos por *Silk* resultan muy inferiores a los obtenidos por las heurísticas propuestas.

***Análisis de la relación con recursos de la Web de los Datos* (*DBpedia*)**

En esta sección se procede comparar ambas propuestas (heurísticas *versus Silk*) adoptando como contexto de experimentación las fuentes tratadas con anterioridad (NGCE - NOMGEO) y DBpedia (una base de conocimiento impulsada por la comunidad y fuente de referencia en el contexto de la Web de los Datos). La experimentación se lleva a cabo considerando la combinación de heurísticas (URI, *Label* e idioma, URI*label* y hWGS84) y la agregación de métricas *Label* & wgs84 de *Silk*.

Tabla 9.15. Relación entre DBpedia y NGCE

		Proceso combinado (1km)				**Proceso combinado (3km)**				**Proceso combinado (5km)**			
Umbral métricas	**Nº métricas**	**Total**	**✓**	**×**	***P***	**Total**	**✓**	**×**	***P***	**Total**	**✓**	**×**	***P***
0,90	**≥ 2**	17	17	-	1,00	21	21	-	1,00	21	21	-	1,00
	Silk	-	-	-	-	-	-	-	-	-	-	-	-
0,80	**≥ 2**	27	18	9	0,67	32	23	9	0,72	32	23	9	0,72
	Silk	-	-	-	-	1	1	-	1,00	4	4	-	1,00
0,70	**≥ 2**	81	20	61	0,25	87	27	60	0,31	87	27	60	0,31
	≥ 3	1	1	-	1,00	1	1	-	1,00	1	1	-	1,00
	≥ 4	1	1	-	1,00	1	1	-	1,00	1	1	-	1,00
	Silk	-	-	-	-	5	5	-	1,00	6	6	-	1,00

Esta experimentación servirá para evaluar la viabilidad de aplicación de los enfoques heurísticos desarrollados en la detección de instancias duplicadas entre fuentes de información del dominio geoespacial y fuentes independientes del dominio (por ejemplo, DBpedia) con cierto contenido de información geográfica.

DBpedia - NGCE. En la relación entre estas fuentes (ver Tabla 9.15) se observa que el número de duplicados se incrementa conforme se aumentan los umbrales de métricas y distancias consideradas. En la mencionada tabla se recoge un mayor número de resultados por parte de la combinación de heurísticas respecto a los ofrecidos por *Silk*. Junto a los escasos resultados, también destaca la ausencia de datos en los umbrales 0,90; 0,80 y 0,70 (en el umbral de 1 km para los dos últimos casos).

Entre los datos recogidos destaca la totalidad de instancias correctamente detectadas (17, 21 y 21) en el umbral 0,90 y ≥2 métricas (para todos los umbrales de distancia fijados en hWGS84). En *Silk* las instancias correctas coinciden con el número total, aunque la cantidad resulta inferior a las detectadas por el enfoque heurístico (entre 6 y 1 instancia). Así, la medida de precisión de los datos toma valor 1 en los datos obtenidos con *Silk* y en aquellos umbrales del enfoque heurístico donde no se detectan errores (0,90). En los restantes umbrales la combinación de heurísticas sitúa el valor de precisión entre 0,25 y 0,72.

DBpedia – NOMGEO. En los datos obtenidos de la relación entre estas fuentes de información (ver Tabla 9.16) se observa, de forma generalizada, una completa similitud en los resultados obtenidos por la combinación de heurísticas. En *Silk* los datos reflejan leves variaciones entre umbrales y destaca la ausencia de valores en algunos de estos umbrales.

Con respecto al número de instancias detectadas, la combinación de heurísticas descubre 27 instancias, mientras *Silk* únicamente consigue relacionar 1 instancia. Así, los datos sitúan el valor de la medida de precisión en 1 (en aquellos umbrales donde el número total de instancias coincide con instancias correctamente detectadas). Mientras, en los umbrales relacionados con 0,80 y 0,70 la precisión se sitúa en 0,37 y 0,08 respectivamente. En *Silk* alcanza un valor de 0,50 (umbrales 0,80 y 0,70), producto de obtener una instancia correcta y un falso positivo.

Tabla 9.16**. Relación entre DBpedia y NOMGEO**

		Proceso combinado (1km)				**Proceso combinado (3km)**				**Proceso combinado (5km)**			
Umbral métricas	**Nº métricas**	**Total**	**✓**	**×**	***P***	**Total**	**✓**	**×**	***P***	**Total**	**✓**	**×**	***P***
0,90	**≥ 2**	22	22		1,00	22	22		1,00	23	23	-	1,00
	Silk	-	-	-	-	1	1		1,00	1	1	-	1,00
0,80	**≥ 2**	68	25	43	0,37	68	25	43	0,37	68	25	43	0,37
	Silk	-	-	-	-	2	1	1	0,50	-	-	-	-
0,70	**≥ 2**	336	27	309	0,08	336	27	309	0,08	336	27	309	0,08
	Silk	1	1	-	1,00	-	-	-	-	2	1	1	0,50

En definitiva, los datos analizados producto de la relación de DBpedia - NGCE y NOMGEO demuestran la viabilidad de la aplicación del enfoque heurístico para el establecimiento de relaciones entre fuentes de información geoespacial y fuentes independientes del dominio con cierto contenido de información geográfica. En este sentido, se puede afirmar que la escasa presencia de este tipo de información entre los recursos de información de DBpedia hace que los resultados obtenidos no sean aún mejores. Además, estos resultados han demostrado que el enfoque heurístico desarrollado supone una importante mejora en la detección de instancias duplicadas con respecto a una de las herramientas más extendidas en el contexto de la Web de los Datos (*Silk*).

9.5. Verificación de hipótesis

En esta sección se procede a la verificación de las hipótesis de partida planteadas en el Capítulo 3. Esta verificación se realiza conforme a los resultados obtenidos en la evaluación de las propuestas presentadas en este trabajo.

En este trabajo se ha propuesto una metodología, basada en ontologías, para la integración de datos geográficos de bases de datos heterogéneas (hipótesis **H1**). Esta metodología es conforme con los principios de la Directiva Europea INSPIRE (referidos al almacenamiento, disponibilidad y mantenimiento de los datos espaciales) y posee influencias de las recomendaciones y buenas prácticas propuestas en ingeniería del software, ingeniería del conocimiento, *ontology matching*, *Linking Open Data*, etc. Además, la aplicabilidad de esta metodología queda demostrada a través del caso de aplicación recogido en este trabajo (información del dominio hidrográfico), descrito en el Capítulo 4, 5, 6, y 7. Finalmente, la independencia de herramientas queda reflejada por la diversidad de herramientas propuestas en las diferentes tareas que componen las actividades descritas.

La comparación realiza en la sección 9.1 pone de manifiesto que los recursos ontológicos desarrollados (*hydrOntology*) constituyen un marco de referencia común y compartido para la comunidad geoespacial, consecuencia de la consideración de múltiples fuentes de información, diversos recursos ontológicos, diferentes criterios de estructuración y metodologías extendidas. Asimismo, su formalización en lógica de descripciones y la implementación en OWL ofrecen a este recurso un nivel de formalización y expresividad ausente en la mayoría de fuentes de información geoespacial. Estas características permiten la verificación de la **hipótesis H2** planteada en el Capítulo 4.

Por otro lado, se ha descrito como la generación de siete conjuntos de datos RDF (**hipótesis H3**) conlleva la solución de diversos problemas de las bases de datos originales (diferencias léxico-sintácticas, ausencia de tipología de fenómenos y presencia de nombres oficiales y alternativos). Una descripción detallada de estos problemas y de las soluciones propuestas se recoge en el Capítulo 6 (sección 6.4) de este trabajo. Además, la generación del RDF permite la agrupación de los diferentes registros (asociados a un mismo fenómeno) de la base de datos en un recurso único. Esto conlleva una primera depuración de instancias duplicadas dentro de una fuente de información.

La **hipótesis H4** queda verificada con el trabajo de generación de instancias de la información geométrica descrito en el Capítulo 7. Así, se ha demostrado que es posible tratar la información de la geometría almacenada en bases de datos, obteniendo como resultado geometría en un formato estándar (GML). Además, a través de la librería GEOMETRY2RDF, se incorpora esta característica de la información geoespacial en la generación de RDF.

Adicionalmente, la publicación de GeoLinked Data en la Web de los Datos y los datos ofrecidos sobre el impacto de los conjuntos de datos RDF publicados (información hidrográfica del territorio nacional español), así como el incremento de la presencia del idioma español, permiten la verificación de la **hipótesis H5**.

Finalmente, los resultados obtenidos del proceso de detección de instancias duplicadas han proporcionado evidencias positivas sobre la detección de instancias duplicadas en conjuntos de datos RDF geoespaciales heterogéneos (**hipótesis H6**). Asimismo, se puede afirmar que los resultados de la evaluación son positivos y motivadores, y que se intuye que el funcionamiento de las diversas heurísticas desarrolladas puede ser generalizable a las instancias de otros tipos de fenómenos geográficos.

9.6. Conclusiones

Las principales conclusiones que se pueden extraer de la evaluación descrita en este capítulo son las siguientes:

La evaluación del recubrimiento del dominio en las ontologías hidrográficas analizadas ha demostrado que las ontologías desarrolladas (*hydrOntology* y la red en torno a *hydrOntology*) constituyen un marco de referencia común y compartido para la comunidad geoespacial. Este

hecho es consecuencia de la consideración de múltiples fuentes de información, diversos recursos ontológicos, diferentes criterios de estructuración y metodologías extendidas en el área de la Ingeniería Ontológica. Además, se ha reflejado la diferencia de granularidad de las ontologías desarrolladas (desde una perspectiva global hasta información local) y la inclusión de información multilingüe (castellano, inglés, francés y catalán) frente a las ontologías existentes.

En lo que respecta a la metodología de integración de datos geoespaciales, se ha realizado una evaluación centrada en dos aspectos: Por un lado, se han considerado los principios de diseño (condiciones necesarias y suficientes descritas en el Capítulo 4 (sección 4.2). Por otro lado, se ha tratado la aplicabilidad de la metodología propuesta.

El análisis de los principios de diseño ha puesto de manifiesto la utilidad de dicha metodología para la depuración o curación de bases de datos, como se ha comprobado en las distintas recomendaciones plasmadas. Asimismo, se ha evidenciado que esta metodología sigue los principios de la Directiva Europea INSPIRE referidos al almacenamiento, disponibilidad y mantenimiento de los datos espaciales y su posesión de influencias procedentes de las recomendaciones y buenas prácticas propuestas en ingeniería del software, ingeniería del conocimiento, *ontology matching*, *Linking Open Data*, etc. La independencia de herramientas, la consideración de diversas fuentes de información en el proceso de integración y su contribución a alcanzar la interoperabilidad estructural y semántica también son aspectos a destacar.

Con respecto a la aplicabilidad, esta ha quedado demostrada a través del caso de aplicación recogido en este trabajo (información del dominio hidrográfico), descrito en el Capítulo 4, 5, 6, y 7.

Los datos resultantes del proceso de generación de RDF (más de 28,8 millones de tripletas RDF pertenecientes a siete conjuntos de datos diferentes) y su publicación en la nube de *Linking Open Data* han demostrado el enriquecimiento de la Web de los Datos con información hidrográfica. Además, la publicación de esta información, como se refleja en los datos ofrecidos, ha supuesto un incremento de la información en español en la referida Web.

En la detección de instancias duplicadas, los resultados obtenidos en la evualuación de los distintos enfoques heurísticos han demostrado la viabilidad de las propuestas desarrolladas para la detección de instancias duplicadas en conjuntos de datos RDF heterogéneos. Además, las diferentes comparaciones realizadas han evidenciado que el enfoque heurístico desarrollado supone una mejora para la detección de instancias duplicadas con respecto a una de las herramientas más extendidas en el contexto de la Web de los Datos (*Silk*). Asimismo, considerando los resultados obtenidos por los diversos enfoques heurísticos desarrollados, se intuye que su funcionamiento puede ser generalizable a las instancias de otros tipos de fenómenos geográficos.

Capítulo 10

Conclusiones y líneas futuras de investigación

Este capítulo realiza un recorrido por los principales elementos descritos en este trabajo, tratando los problemas de de investigación abiertos, los objetivos formulados y las contribuciones realizadas a cada problema (sección 10.1). Finalmente, la sección 10.2 proporciona algunas líneas de investigación futuras.

10.1. Conclusiones

El objetivo general de este trabajo es integrar información de bases de datos heterogéneas en el contexto de la información geográfica. Como resultado de la investigación realizada en el marco de este trabajo se ha desarrollado una metodología, basada en ontologías, para la integración de datos geoespaciales heterogéneos, se han desarrollado recursos ontológicos de dominio y multilingües en el contexto de la información hidrográfica, se han transformado e integrado diversos conjuntos de datos, se ha publicado información geoespacial (concretamente información hidrográfica) conforme a los principios de *Linked Data* y se han propuesto diversas heurísticas para la detección de instancias duplicadas. A continuación, se recogen para cada uno de los principales temas de investigación tratados en este trabajo los problemas de investigación abiertos, los objetivos formulados y las contribuciones aportadas.

Heterogeneidad en la información geoespacial

Problema de investigación abierto

En la literatura relacionada con el dominio geoespacial son múltiples los trabajos que recogen diversas categorizaciones y características de los distintos tipos de heterogeneidad presentes en la información geográfica. Los diversos enfoques existentes se centran, fundamentalmente, en los problemas presentes en las bases de datos. Sin embargo, en la literatura referida no existe una recopilación exhaustiva de los tipos de disparidades existentes en diferentes ámbitos de la información geográfica. Asimismo, no existe una propuesta de interrelación entre los principales problemas de esta información y los diferentes tipos de heterogeneidad.

Por otro lado, no existe en la literatura del dominio una caracterización detallada de los problemas en el ámbito de las instancias durante el proceso de integración de información de múltiples bases de datos geográficas.

Objetivos

Para solventar las limitaciones anteriores, se formulan los siguientes objetivos científicos.

O2. Identificar y armonizar los diferentes tipos de heterogeneidad existentes en diferentes contextos de la información geográfica e interrelacionarlos con los principales problemas de esta información.

O3. Refinar los problemas existentes en el contexto de la integración de instancias procedentes de bases de datos geoespaciales heterogéneas.

Contribuciones

Con respecto al área de la heterogeneidad de la información geoespacial se presentan las siguientes contribuciones:

Se recoge una recopilación exhaustiva de las diferentes clasificaciones de tipos de heterogeneidad vinculadas con la información geográfica. Esta clasificación recoge los problemas presentes en el ámbito de las bases de datos, las visiones sobre la realidad geoespacial y los modelos de conocimiento. Asimismo, estos tipos de heterogeneidad (categorizados en niveles, siguiendo la analogía de las propuestas presentadas) se vinculan con algunos de los principales problemas de la información geográfica.

Además, este trabajo aporta un **refinamiento de los problemas de integración** que se pueden detectar en bases de datos geoespaciales heterogéneas en el contexto de la información hidrográfica. Estos problemas son abstraídos de forma general para el dominio de la información geográfica y, en ciertos casos, resultan manifiestamente dependientes del dominio hidrográfico. La clasificación de estos problemas sigue un enfoque basado en capas (léxica, sintáctica, semántica y pragmática) que aparece en diferentes trabajos sobre interoperabilidad semántica (Euzenat, 2001; Corcho, 2005; Barrasa, 2007).

Integración de información

Problema de investigación abierto

A pesar de la diversidad de enfoques de integración de información basados en ontologías existentes, tanto en el plano general como en el dominio de la información geográfica (descritos en el Capítulo 3, sección 3.2.6 y 3.2.7), hasta el momento no se ha desarrollado ninguna metodología que se pueda aplicar en este área de integración de información. Asimismo, de los diferentes enfoques tratados se extraen limitaciones para su replicabilidad, conocimiento, testeo y aplicación.

Objetivos

O1. Proponer una metodología para la integración basada en ontologías de información de bases de datos heterogéneas en el dominio geográfico.

Contribuciones

En este trabajo se han **identificado las fases, actividades y tareas** a realizar en la **propuesta metodológica desarrollada** para la integración de información, basada en ontologías, de bases de datos geoespaciales heterogéneas. Asimismo, se han identificado los **participantes**, **técnicas a utilizar** y las **entradas y salidas** asociadas a cada tarea. Finalmente, se aporta un ejemplo práctico de cada una de las tareas propuestas en las diversas actividades con la finalidad de demostrar la aplicabilidad de la misma.

Problema de investigación abierto

Una de las principales limitaciones para llevar a cabo una propuesta de integración de información basada en ontologías en el dominio geoespacial es la ausencia de recursos

ontológicos. Esta limitación resulta aún mayor si se requieren conceptualizaciones globales de la realidad de los fenómenos geográficos. Asimismo, en el subdominio hidrográfico no existen recursos ontológicos de dominio y multilingües que recojan determinados fenómenos propios del contexto nacional español.

Por otro lado, el proceso de integración de información en el dominio geográfico posee como aspecto característico y diferenciador la presencia de la información geométrica. Sin embargo, hasta el momento, no existe ninguna propuesta que permita tratar y transformar, de manera específica, la información geométrica recogida en las bases de datos geoespaciales a RDF.

Objetivos

O4. Integrar información de bases de datos heterogéneas para proporcionar una visión armonizada de la realidad recogida en los datos geoespaciales.

Contribuciones

Para la consecución del objetivo vinculado con la integración de información de bases de datos heterogéneas se han realizado las siguientes contribuciones:

- La creación **de recursos ontológicos de dominio y multilingües en el área de la información hidrográfica**. Estos recursos están centrados en fenómenos representables cartográficamente, con objeto de unificar las diversas fuentes y modelos existentes en la comunidad geográfica a diferentes niveles.

- La **generación de siete conjuntos de datos RDF**, a través de la creación de *wrappers* entre los recursos ontológicos creados y las diversas fuentes de información (bases de datos) estudiadas.

Adicionalmente se ha realizado la siguiente contribución tecnológica:

- El desarrollo de la librería GEOMETRY2RDF Las contribuciones en este área también incluyen la implementación de una librería, descrita en el Capítulo 7, que permite el **tratamiento e inclusión de las características geométricas de la información geoespacial** a través de la incorporación de *Geography Markup Language* (GML) en el RDF generado. Asimismo, el tratamiento realizado sobre la información geométrica ha permitido unificar las múltiples entradas de un topónimo presentes en una misma base de datos. De esta manera, se pasa de tener diferentes registros (filas) de una instancia en una tabla de una base de datos, donde se almacenan las diferentes geometrías (segmentos) asociadas a un fenómeno, a conseguir una **visión más unificada de la geométrica en torno a un único recurso** (instancia), incluyéndose dicha información en el RDF generado.

Publicación de datos integrados (*Linked Data*)

Problemas de investigación abiertos

La cantidad de datos enlazados publicados en la Web de los Datos ha experimentado un enorme crecimiento en los últimos años. La lista de recursos ya disponibles en *Linked Data* crece día a día. El mayor auge hasta ahora se ha producido en el contexto de la publicación de datos del sector público, siendo referencia los gobiernos del Reino Unido y de Estados Unidos. Asimismo, el fenómeno de *Linked Data* se está extendiendo a otros sectores, entre los que destacan los medios de comunicación, infraestructuras y logística, el ámbito universitario y científico. Sin embargo, hasta el momento, en el ámbito de la información geográfica los datos publicados no resultan abundantes, ya que no se han iniciado suficientes esfuerzos en esta dirección. Asimismo, a pesar de las diferentes iniciativas destinadas a la publicación de datos

geoespaciales en la Web de los Datos, no existe hasta el momento una iniciativa abierta destinada al enriquecimiento de la Web de Datos con datos geoespaciales del territorio nacional español. Además, esta situación evidencia una pobre presencia de los datos en español.

Objetivos

O5. Publicar información geoespacial del territorio nacional español conforme a los requerimientos de *Linked Data*.

Contribuciones

La principal contribución se centra en la **publicación de información hidrográfica del territorio nacional español conforme a los principios de *Linked Data.*** Esta información permite el **enriquecimiento de la Web de los Datos con información geoespacial en español**. Este enriquecimiento se demuestra con las cifras recogidas en el Capítulo 9, donde se comprueba que la publicación de los datos generados en el contexto de este trabajo hace que la información en español ascienda del 0,02% al 5,04%[1]. Este porcentaje de información sitúa al idioma español, en las fechas citadas, como el segundo idioma de la Web de los Datos.

Además, las contribuciones en este área incluyen la implantación de la arquitectura descrita en el Capítulo 7 (sección 7.5). Esta contribución permite la publicación de los diversos conjuntos de datos a través de SPARQL Endpoints, así como su visualización a través de un *FrontEnd* de *Linked Data* para los SPARQL Endpoints desplegados y la visualización gráfica de la información geográfica (aplicación web y *faceted browser*).

Detección de duplicados

Problemas de investigación abiertos

Con respecto a la detección de duplicados, se ha detectado que no existen en la bibliografía del área trabajos que recopilen, de forma exhaustiva, los métodos de detección de duplicados considerando las características de la información geoespacial. Asimismo, entre los diferentes enfoques existentes resultan realmente escasas y limitadas las propuestas de detección de duplicados entre conjuntos de datos RDF que consideren las características propias de la información geográfica.

Objetivos

O6. Proponer métodos para la detección de instancias duplicadas en conjuntos de datos RDF heterogéneos.

Contribuciones

En primer lugar se realiza un estudio detallado de la detección de duplicados en las diversas ramas de la ciencia de la información donde este tema resulta objeto de estudio. Así, se mencionan diferentes propuestas para la detección de duplicados en el contexto de la información estadística, las bases de datos y la inteligencia artificial. Finalmente, consecuencia del contexto de este trabajo se presentan los enfoques existentes en la Web de los Datos y los aplicados a la información geoespacial.

Por otro lado, **se proponen diversos enfoques heurísticos para la detección de instancias duplicadas en conjuntos de datos RDF heterogéneos y se formalizan las relaciones** (*<owl:sameAs>* y *<owl:differentFrom>*) **entre instancias** pertenecientes a un mismo y/o diferentes conjuntos de datos. Las heurísticas desarrolladas, de forma individual, están

[1] Estos datos corresponden a la información publicada en la iniciativa *Open Linking Data* entre los meses de Julio-Agosto de 2010.

clasificadas como independientes y dependientes del dominio geográfico y se dividen en distintas capas (léxica, sintáctica, semántica y pragmática). Las heurísticas desarrolladas, a diferencia de otras propuestas existentes, tratan con las características de la información geoespacial para la detección de instancias duplicadas.

Adicionalmente, las contribuciones en este área también incluyen la implantación de una arquitectura, descrita en el Capítulo 8, que permite desplegar los enfoques heurísticos desarrollados para la detección de duplicados e interconectar los diferentes conjuntos de datos RDF publicados conforme a los principios de *Linked Data*.

10.2. Líneas futuras de investigación

A pesar del tratamiento de diversos problemas de investigación abiertos en el contexto de la integración de información geoespacial, este trabajo abre una serie de problemas que pueden ser resueltos en futuras investigaciones. Así, a continuación se mencionan los temas de mayor interés desde nuestra perspectiva:

Metodología.

- La metodología propuesta se centra en la integración de información de bases de datos relacionales. Sin embargo, la información geográfica está contenida en diferentes formatos. Un aspecto interesante que puede extender el trabajo realizado es la consideración de otros tipos de fuentes de información (hojas de cálculo, shapefiles, servicios web, servicios RESTful, etc.) en el proceso de integración de información. Este trabajo contribuiría a generalizar la metodología propuesta y al enriquecimiento del proceso de integración.

- Actualmente, la metodología desarrollada se está aplicando bajo el contexto del proyecto España Virtual, más concretamente, en el desarrollo de la iniciativa GeoLinked Data. Este hecho ha puesto de manifiesto la aplicabilidad de esta metodología. No obstante, se pretende extender la aplicación de esta metodología a otros dominios de conocimiento donde el componente espacial de la información sea un elemento fundamental (por ejemplo: información estadística, biodiversidad, etc.)

- Finalmente, otro trabajo futuro de interés sería analizar cómo afecta el origen (*provenance*) y la diferencia de calidad de las fuentes de información en el proceso de integración. Asimismo, aunque en la metodología propuesta se consideran diversos aspectos relacionados con la evolución de las fuentes de información, resultaría interesante analizar de manera detallada los problemas y consecuencias relacionadas con la evolución dinámica de las fuentes de información en el proceso de integración.

Recursos ontológicos y multilingüismo.

- La realidad geoespacial está directamente influenciada por el contexto que la rodea. Asimismo, la perspectiva de la globalización y el componente multilingüe comienzan a adquirir la importancia que le confiere la explotación actual de la información. Por tanto, un aspecto de interés sería tratar la localización de ontologías y sus peculiaridades en el dominio geoespacial.

Detección de duplicados

- Debido a que la Web de los Datos resulta cada vez más un contexto multilingüe, la tarea de detección de duplicados se presenta ante un nuevo escenario. Por tanto, otra futura

investigación de interés consistirá en la generación de métodos destinados a la detección de instancias similares en diferentes idiomas dentro de la Web de los Datos.

- Otro aspecto interesante en el ámbito de la detección de duplicados sería la determinación de la calidad de las fuentes de información participantes en el proceso de integración a través de fuentes publicadas en la Web de los Datos. Esto contribuiría a determinar pesos en el proceso de integración, así como a la depuración o curación de las propias fuentes de información.

RDF* y *Linked Data.

- Otro aspecto de interés tecnológico consiste en completar la librería desarrollada para el tratamiento y generación de RDF de la información geométrica almacenada en diferentes formatos, por ejemplo, *shapefiles*. Asimismo, unido al tratamiento de la geometría, sería interesante tratar la unificación automática de los segmentos que componen la representación (geometrías complejas) de un fenómeno. De esta manera, se daría continuidad geométrica y representacional a los fenómenos geográfico en tripletas RDF.

- Otro trabajo futuro de interés consiste en la aplicación de operaciones espaciales sobre conjuntos de datos RDF y cómo las herramientas tradicionales del área (Sistemas de Información Geográfica) y/o su combinación con los nuevos enfoques de la Web de los Datos (*Triple Store* con extensiones espaciales, Geo-SPARQL, etc.) pueden ayudar en el almacenamiento y tratamiento de la información geográfica.

- En este trabajo se ha tratado en detalle la iniciativa *Linked Data* (descrita en el Capítulo 3, sección 3.4). Además, uno de las principales contribuciones de este trabajo es la publicación de información hidrográfica del territorio nacional español conforme a los principios de *Linked Data*. Un interesante trabajo futuro sería interrelacionar la iniciativa *Linked Data* y las Infraestructuras de Datos Espaciales (IDE) a través de los diferentes servicios ofrecidos por el *Open Geospatial Consortium*.

Bibliografía

Abraham T., Roddick J.F (1999) Survey of Spatio-Temporal Databases. GeoInformatica, Volume 3, Number 1, pp. 61-99(39)

Adali, S., Candan, K., Papakonstantinoy, Y., Subrahmanian, V. (1996) Query caching and optimization in distributed mediator systems. In Proc. Of ACM SIGMOD Conf on Management of Data. Montreal, Canada.

Adali, S. Emery, R. (1995) A uniform framework for integrating knowledge in heterogeneous knowledge systems. In Proc of the Eleventh IEEE International Conference of Data Engineering.

AdV (1998) Amtliches Topographisch-Kartographisches Informationssystem ATKIS. Landesvermessungsamt NRW, Bonn.

AGENT, CARTACOM and GAEL. In Headway in Spatial Data Handling. 13th SDH'08. Lecture Notes in Geoinformation and Cartography, Ruas, A., Gold, C. (Eds.) pp. 277-296. ISBN: 978-3-540-68565-4. Montpellier, France.

Aguado de Cea, G., Montiel-Ponsoda, E., Ramos, J. A. (2007) Multilingualidad en una aplicación basada en el conocimiento TIMM, Monográfico para la revista SEPLN

Albrecht, J. (1999) Geospatial information standards a comparative study of approaches in the standardisation of geospatial information. Computers & Geosciences, 25:9–24.

Alonso G, Abbadi AE (1994) "Cooperative modelling in applied geographic research". International Journal of Intelligent and Cooperative Information Systems, 3(1): 83-102

Ambite, J. L., Knoblock, C. A. (1998) "Flexible and Scalable Query Planning in Distributed and Heterogeneous Environments," in Proceedings of the Fourth International Conference on Artificial Intelligence Planning Systems, Pittsburgh Pennsylvania, USA, pp. 3-10.

Ananthakrishna, R., Chaudhuri, S., Ganti, V. (2002) Eliminating fuzzy duplicates in data warehouses. In: Proceedings of the 28th International Conference on Very Large Data Bases, VLDB Endowment, 586-597

Arens, Y., Knoblock, C. A., Shen, W.-M. (1996) "Query reformulation for dynamic information integration," J. Intell. Inf. Syst., vol. 6, no. 2-3, pp. 99–130.

Arpinar, B., Sheth, A., Ramakrishnan, C., Usery, L., Azami, M., & Kwan, M.-P. (2006) Geospatial ontology development and semantic analytics. In J. P. Wilson & A. S. Fotheringham (Eds.), Handbook of Geographic Information Science: Blackwell Publishing

Arpírez, J.C., Gómez-Pérez, A., Lozano, A., Pinto, H.S. (1998) (ONTO)2Agent: An ontology-based WWW broker to select ontologies. In: Gómez-Pérez A, Benjamins RV (eds) ECAI'98 Workshop on Applications of Ontologies and Problem-Solving Methods. Brighton, United Kingdom, pp 16–24

Arroyo, S., de Bruijn J., Ding, Y. (2003) Ontology Mapping and Aligning. Deliverable D1.4, Esperonto. http://www.sti-innsbruck.at/fileadmin/documents/deliverables/Esperonto/Del14-v01.1.doc

Attaluri, G. K., Bradshaw, D. P., Coburn, N., Larson, P.-A., Martin, P., Silberschatz, A., Slonim, J., Zhu, Q. (1995) The CORDS multidatabase Project. IBM SYSTEMS JOURNAL, Vol. 34, No. 1. http://www.research.ibm.com/journal/sj/341/attaluri.pdf

Auer, S., Lehmann, J., Hellmann, S. (2009) LinkedGeoData – Adding a spatial Dimension to the Web of Data, ISWC 2009.

Ayers, D., Vlkel, M. (2008) Cool uris for the semantic web. Interest Group Note 20080331, W3C, 2008. http://www.w3.org/TR/2008/NOTE-cooluris-20080331/. W3C Interest Group Note 31 March 2008.

Bakillah, M., Mostafavi, M.A., Bédard, Y. (2006) A semantic similarity model for mapping between evolving geospatial data cubes. In: LNCS. Volume 4278/2006.

Bakillah, M., M. Mostafavi, Y. Bédard, (2006) A semantic similarity model for mapping between evolving geospatial data cubes, International Workshop on Semantic-based Geographical Information Systems (SeBGIS'06). Springer LCNS 4277/4278.

Baksai Elespuru, A., Gutierrez Valenzuela, M. (2007) Software para la manipulación de bases de datos espaciales PostGIS. I Jornadas de SIG Libre. Girona: Universitat. Disponible en: http://hdl.handle.net/10256/1225

Barish, G., DiPasquo, D., Knoblock, C. A., Minton, S. (2000) "Dataflow plan execution for software agents," in Proceedings of the Fourth International Conference on Autonomous Agents, C. Sierra, M. Gini, and J. S. Rosenschein, Eds. Barcelona, Spain: ACM Press, pp. 138–139.

Barrasa, J. (2006) Semantic upgrade and publication of legacy data. In Ontologies for Software Engineering and Software Technology.

Barrasa, J., Corcho, O., Gómez-Pérez, A. (2003) Fund finder: A case study of database-to-ontology mapping. In Natalya F. Noy, editor, Proceedings of the Semantic Integration Workshop (ISWC 2003), Sanibel Island, Florida. USA.

Barrasa, J., Corcho, O., Gómez-Pérez, A. (2004) R2O, an extensible and semantically based database-to-ontology mapping language. In: Second Workshop on Semantic Web and Databases (SWDB'04), Chris Bussler and Val Tannen, (eds.). Toronto (Canada).

Barrasa Rodríguez, J. (2007) Modelo para la definición automática de correspondencias semánticas entre ontologías y modelos relacionales. Tesis doctoral, Universidad Politécnica de Madrid. Madrid.

Barrasa, J., Gómez-Pérez, A. (2006) Upgrading relational legacy data to the semantic web. In Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland). WWW '06. ACM Press, New York, NY, 1069-1070.

Baru, C. K., Gupta, A., Ludäscher, B., Marciano, R. Papakonstantinou, Y., Velikhov, P., Chu, V. (1999) Xml-based information mediation with mix. In Alex Delis, Christos Faloutsos, and Shahram Ghandeharizadeh, editors, SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA, pages 597-599. ACM Press.

Batini, C. and Scannapieco, M., (2006) Data Quality: Concepts, Methodologies and Techniques, Data-Centric Systems and Applications. Springer.

Batini, C., Lenzerini, M., Navathe, S. (1986) "A comparative analysis of methodologies for database schema integration," ACM Computer Survey, vol. 18, no. 4.

Baumgartner, N., Gottesheim, W., Mitsch, S., Retschitzegger, W., Schwinger, W. (2009) Same, Same but Different - A Survey on Duplicate Detection Methods for Situation Awareness, Submitted to: 8th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE), Vilamoura, Portugal, Springer.

Bayardo, R. J., Bohrer, Jr., W., Brice, R., Cichocki, A., Fowler, J., Helal, A., Kashyap, V., Ksiezyk, T., Martin, G., Nodine, M., Rashid, M., Rusinkiewicz, M., Shea, R., Unnikrishnan, C., Unruh, A., Woelk, D. (1997) "InfoSleuth: Agent-based semantic integration of information in open and dynamic environments," in Proc.of the ACM SIGMOD International Conference on Management of Data, vol. 26,2. New York: ACM Press, 13–15, pp. 195–206.

Bechhofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., Mcguiness, D.L., Schneider, P.F. and Stein, L.A. (2004) OWL Web Ontology Language Reference, Technical Report, W3C, http://www.w3.org/TR/owl-ref/

Beeri, C., Kanza, Y., Safra, E., Sagiv, Y. (2004) Object fusion in geographic information systems. Proceedings of the Thirtieth international conference on Very large data bases – Vol. 30, 816 – 827. Toronto, Canada.

Beneventano, D., Bergamaschi, S. (2004) The MOMIS methodology for integrating heterogeneous data sources, in: IFIP World Computer Congress, Toulouse, France.

Bennett, B., Cristani, M. (Eds.). (2004) Spatial Cognition and Computation: special issue on spatial vagueness, uncertainty and granularity. Spatial Cognition and Computation, Springer.

Bennett, B., Mallenby, D., Third, A. (2008) An Ontology for Grounding Vague Geographic Terms. Proceedings of the 5th International Conference on Formal Ontology in Information Systems (FOIS-08), edited by C. Eschenbach and M. Gruninger, IOS Press.

Bergamaschi, S., Castano, S., di Vimercati, S. D. C., Montanari, S., Vincini, M. (1998) An intelligent approach to information integration. In Guarino, N., editor, Formal Ontology in Information Systems, pages 253–267. IOS Press.

Bergamaschi, S., Castano, S., Vincini, M. (1999) Semantic integration of semistructured and structured data sources. SIGMOD Rec., 28(1):54–59.

Belin, T. R., Rubin, D. B. (1997) A method for calibrating false-match rates in record linkage. In Record Linkage – 1997: Proceedings of an International Workshop and Exposition, 81– 94. U.S. Office of Management and Budget (Washington).

Bernard L., Craglia, M., Gould, M., & Kuhn, W. (2005) Towards an SDI research agenda. Proceedings of EC & GIS Workshop, Sardinia.

Berners-Lee, T. (2006) Linked data. World Wide Web design issues. http://www.w3.org/DesignIssues/LinkedData.html.

Bernstein, P. A., Giunchiglia, F., Kementsietsidis, A., Mylopoulos, J., Serafini, L., Zaihrayeu, I. (2002) Data management for peer-to-peer computing: A vision. In Proc. 5th Int. Workshop on the World Wide Web and Databases (WebDB 2002).

Berrueta, D., Fernández, S., Frade, I. (2008) Cooking HTTP content negotiation with Vapour. ESWC2008 workshop on Scripting for the Semantic Web (SFSW2008), Tenerife (Spain).

Bilenko, M., Mooney, R. (2002) Learning to combine trained distance metrics for duplicate detection in databases. Technical Report Technical Report AI 02-296, Artificial Intelligence Lab, University of Texas at Austin.

Bilenko, M., Mooney, R. J., Cohen, W. W., Ravikumar, P., Fienberg, S. E. (2003) Adaptive name matching in information integration. IEEE Intelligent Systems, 18(5):16-23.

Bishr, Y. (1997). Semantic Aspects of Interoperable GIS. ITC Publication no 56, PhD thesis, Wageningen Agricultural University and International Institute for Aerospace Survey and Earth Sciences (ITC), Enschede, the Netherlands. ISBN 90 6164 1411.

Bishr Y (1998) "Overcoming the semantic and other barriers to GIS interoperability". International Journal of Geographical Information Science, 12(4): 299–314.

Bishr, Y., Pundt, H., Kuhn, W., Radwan, M. (1999) Probing the concept of information communities - a first step towards a road of semantic interoperability. In: Goodchild, M.F., Egenhofer, M.J., Fegeas, R., Kottman, C.A. (Eds.), Interoperating Geographic Information Systems. Kluwer Academic Publishers, Boston, MA, pp. 55-71.

Bishr, Y. A., Pundt, H., Ruther, C. (1999a), Proceeding on the Road of Semantic Interoperability - Design of a Semantic Mapper Based on a Case Study from Transportation. In: Vckovski, A.

et al. (eds): Interoperating Geographic Information Systems. Proceedings of the 2nd International Conference INTEROP'99. Zurich, Switzerland. Springer-Verlag, 203–215.

Bizer, C., Cyganiak, R., Heath, T. (2007) How to publish linked data on the web. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

Bizer, C., Heath, T., Berners-Lee, T. (2009) Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS) 5(3):1-22.

Bizer, C., Heath, T., Idehen, K., Berners-Lee, T. (2008) Linked data on the web (LDOW2008). In Proceeding of the 17th International Conference on World Wide Web, pages 1265-1266, Beijing, China.

Bizer, C. (2003) D2r map - a database to rdf mapping language. In 12th Intl World Wide Web Conference Budapest.

Bizer, C., Cyganiak, R., Heath, T. (2007) How to publish Linked Data on the Web. http://www4.wiwiss.fuberlin.de/bizer/pub/LinkedDataTutorial/

Bizer, C., Heath, T. and Berners-Lee, T. (2009) Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS). Vol. 5(3), Pages 1-22.

Blanco, J.M., Goni, A., Illarramendi, A. (1999) Mapping among knowledge bases and data repositories: Precise definition of its syntax and semantics. Information Systems, 24(4):275–301.

Blanco, J. M., Illarramendi, A., Goñi, A. (1994) "Building a federated database system: An approach using a knowledge based system", International Journal on Intelligent and Cooperative Information Systems, vol. 3, no. 4, pp. 415-455.

Bleiholder, J., Naumann, F. (2008) Data fusion. ACM Computing Surveys 41(1).

Boletín Oficial del Estado (BOE) (1986) Ley 7/1986, de 24 de enero, de Ordenación de la Cartografía. http://www.catastro.meh.es/pdf/ley_000786_2401.pdf

Boletín Oficial del Estado (BOE) (2007) Real Decreto 1545/2007, de 23 de noviembre, por el que se regula el Sistema Cartográfico Nacional (BOE de 30 de noviembre de 2007).

Borgida, A., Mylopoulos, J. (2004) Data semantics revisited. In Semantic Web and Databases. 2nd International Workshop, SWDB 2004. Toronto Canada. Springer Verlag.

Boulcema, O., Essid, E., Lacroix, Z. (2002) A WFS-Based Mediation System for GIS Interoperability. Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems, 23–28.

Brodeur, J., Bédard, Y. (2001) Geosemantic Proximity, a Component of Spatial Data Interoperability. The International Workshop on "Semantics of Enterprise Integration", ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA 2001), Tampa Bay, Florida.

Brodeur, J., Bédard, Y., Edwards, G., Moulin, B. (2003) Revisiting the Concept of Geospatial Data Interoperability within the Scope of a Human Communication Process. Transactions in GIS, 7, pp. 243-265.

Bowers, S., Ludäscher, B. (2004) An Ontology-Driven Framework for Data Transformation in Scientific Workflows. International Workshop on Data Integration in the Life Sciences (DILS'04), Leipzig (Germany).

Brugger, B. P., Barrera, R., Frank, A. U., Beard, K., Ehlers, M. (1989) Research Topic on MultipleRepresentations in NCGIA Initiative 3 Workshop on Multiple Representations, pages 53-67.

Bright, M.W., Hurson, A.R. Pakzad, S. (1994). Automated resolution of semantic heterogeneity in multidatabases. ACM Transactions on Database Systems, 19(2), 212-253.

Buccella, A., Cechich, A. (2007) Towards Integration of Geographic Information Systems. Electronic Notes in Theoretical Computer Science 168, pp.: 45-59. Elsevier Science.

Buccella, A., Cechich, A., Brisaboa, N. R. (2006) A Federated Layer to Integrate Heterogeneous Knowledge. Electronic Notes in Theoretical Computer Science 142, p.79–97. Elsevier.

Buehler, K. (2003) OpenGIS Reference Model. Open Geospatial Consortium (OGC).

Buneman, P., Raschid, L., Ullman J.D. (1997) Mediator languages - a proposal for a standard. SIGMOD Record, 26(1):39–44.

Butenuth, M., Gösseln, G., Tiedge, M., Heipke, C., Lipeck, U., Sester, M. (2007) Integration of heterogeneous geospatial data in a federated database. ISPRS Journal of Photogrammetry & Remote Sensing 62, p. 328–346. Elsevier.

Calvanese, D., De Giacomo, G., Lenzerini, M. (2001) A Framework for Ontology Integration. SWWS 2001: 303-316.

Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D., Rosati, R. (1998) Information integration: Conceptual modeling and reasoning support. In Proc. of the 6th Int. Conf. on Cooperative Information Systems (CoopIS'98), pages 280–291.

Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R. (2004) Logical foundations of peer-to-peer data integration. In Proc. Of the 23rd ACM SIGACT SIGMOD SIGART Sym. on Principles of Database Systems (PODS-2004).

Camara, G., Thome, R., Freitas, U., Monteiro, A.M.V. (1999) Interoperability In Practice: Problems in Semantic Conversion from Current Technology to OpenGIS. In Proceedings of INTEROP '99, Switzerland, pp. 129-138.

Castano, S., De Antonelli, V. (1997). Semantic dictionary design for database interoperability. In. Proceedings of the International Conference on Data Engineering (ICDE97), IEEE Computer Society, Birmingham, UK.

Castano, S., Ferrara, A., Lorusso, D., Montanelli, S. (2008) On the ontology instance matching problem. In: Proceedings of the 19th International Conference on Database and Expert Systems Applications, Turin, Italy, IEEE, 180-184

Castellanos, M., Saltor, F., García-Solaco, M. (1992) A canonical model for interoperability among object-oriented and relational databases. In Özsu, M. T., Dayal, U., and Valduriez, P., (eds), Distributed Object Management: Papers from the International Workshop on Distributed Management (IWDOM), p. 309–314, Edmonton, Alberta, Canada. Morgan Kaufmann.

Chamberlin, D., Florescu, D., Robie, J., Simeon, J. Stefanascu, M. (2001) XQuery: A query language for XML. Technical report, World Wide Web Consortium.

Chang, Y. S., Park, H. D. (2006) XML Web-Service-based development model for GIS applications. International Journal of Geographical Information Science, 20(4):371-399.

Chatterjee A., Segev A., (1995) "Rule Based Joins in Heterogeneous Databases," Decision Support Systems, vol. 13, no. 1, pp. 313-333.

Chaudhuri, S., Dayal, U. (1997) An overview of data warehousing and olap technology. ACM SIGMOD Record, 26(1):65–74.

Chaudhri V K, Farquhar A, Fikes R, Karp PD, Rice JP (1998) Open Knowledge Base Connectivity 2.0.3. Technical Report KSL-98-06, Knowledge Systems Laboratory, Stanford, CA, http://www.ai.sri.com/ okbc/okbc-2-0-3.pdf.

Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J. D., Widom, J. (1994) "The TSIMMIS project: Integration of heterogeneous information sources," in 16th Meeting of the Information Processing Society of Japan, Tokyo, Japan, pp. 7–18.

Chen, H., Wu, Z., Wang, H., Mao, Y. (2006a) Rdf/rdfs-based relational database integration. In Ling Liu, Andreas Reuter, Kyu-Young Whang, and Jianjun Zhang, editors, Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006. IEEE Computer Society.

Chen, H., Wu, Z., Wang, Y., Cui, M., Yin, A. (2006b) From legacy relational databases to the semantic web:an in-use application for traditional chinese medicine. In Proceedings of the fifth International Semantic Web Conference. ISWC2006, Athens, Georgia, USA.

Chen, P. (1975) The entity - relationship model. Towards a unified view of data. In International Conference on Very Large Databases, VLDB'75.

Choi, N., Song, I.Y., Han, H. (2006) A survey on ontology mapping. ACM SIGMOD Record 35(3), 34-41

Chou, H. H. (2005) BioDig: architecture for integrating heterogeneous biological data repositories using ontologies. Master Thesis. Massachusetts Institute of Technology (MIT), Department of Electrical Engineering and Computer Science. Massachusetts. http://dspace.mit.edu/bitstream/handle/1721.1/33116/62233391.pdf?sequence=1

Christaki, A., Tzouvaras, V., Fresa, A., Caffo, R., Sola, P. G., Kollias, S. (2007) Achieving Interoperability in the MichaelPlus Project. Workshop on Ontology-Driven Interoperability for Cultural Heritage Objects. Tirrenia, Pisa (Italy).

Codd EF (1970) A relational model of data for large shared databanks. Communications of the ACM, 13 (6), pp: 377-387.

Cohen, T. L., Baitty, J. R. (2005) Health Resources and Services Administration Geospatial Data Warehouse: Integrating Spatial and Tabular Extract, Transform, and Load Processes. ESRI 2005 International Users Conference.

Cohen, W. W. (2000) Data integration using similarity joins and a word-based information representation language. ACM Transactions on Information Systems 18(3):288–321.

Cohen, W. W., Richman, J. (2002) Learning to match and cluster large high-dimensional data sets for data integration. In Proceedings of The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002).

Commission of the European Communities (2007) Directiva 2007/2/CE del Parlamento Europeo y del Consejo, de 14 de marzo de 2007, por la que se establece una infraestructura de información espacial en la Comunidad Europea (INSPIRE).

Conrad, S. (1997) Federated Database Systems-Concepts of Data Integration. Springer, Heidelberg, 331 p.

Contreras, J., Benjamins, V. R., Blázquez, M., Losada, S., Salla, R., Sevilla, J., Navarro, D., Casillas, J., Mompó, A., Patón, D., Corcho, O., Tena, P., Martos, I. (2004) "A semantic portal for the international affairs sector." in EKAW, pp. 203–215.

Corcho O. (2005) A layered declarative approach to ontology translation with knowledge preservation. Frontiers in AI and its Applications. Dissertations in AI. IOS Press.

Corcho, O., Fernández-López, M., Gómez-Pérez, A., López-Cima, A., (2005) Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications. Springer-Verlag, Ch. Building legal ontologies with METHONTOLOGY and WebODE, pp. 142–157

Portele, C. (Ed.) (2007) OpenGIS Geography Markup Language (GML) Encoding Standard. Version 3.2.1. Open Geospatial Consortium. Document available at http://www.opengeospatial.org/ standards/gml

Cruz, I. F., Rajendran, A. (2003) Semantic Data Integration in Hierarchical Domains. IEEE Intelligent Systems March-April: 66-73.

Cui, Z., Jones, D. and O'Brien, P. (2001) Issues in Ontology-based Information Integration. IJCAI 2001– Seattle, USA.

Cui, Z., O'Brien, P. (2000) Domain Ontology Management Environment. In Proceedings of the 33rd Hawaii International Conference on System Sciences.

Cyrille, A., Ngomo, N., Auer, S. (2010) Limes - a timeefficient approach for large-scale link discovery on the web of data. http://svn.aksw.org/papers/2011/WWW LIMES/public.pdf

Davidson, P. (2009) Designing URI Sets for the UK Public Sector, UK Chief Technology Officer Council. http://www.cabinetoffice.gov.uk/media/301253/puiblic_sector_uri.pdf

Davidson, S., Buneman, P., Kosky, A. (1998) Semantics of database transformations. Lecture Notes in Computer Science, 1358:55–60

De Hoog, R. (1998) Methodologies for Building Knowledge Based Systems: Achievements andProspects. In: Liebowitz J (ed.) Handbook of Expert Systems. CRC Press Chapter 1, Boca Raton, Florida. 1998.

De Leon, A., Saquicela, V., Vilches-Blázquez, L.M., Villazón-Terrazas, B., Priyatna, F., Corcho, O. (2010) Geographical Linked Data: a Spanish Use Case. In proceeding I-SEMANTICS Triplification Challenge'10 Proceedings of the 6th International Conference on Semantic Systems. ACM 978-1-4503-0014-8/10/09, ISBN: 978-1-4503-0014-8, Graz (Austria).

De Miguel, A., Piattini, M. (1993) Concepción y diseño de bases de datos. Ed. RA-MA.

Dean, M., Schreiber, G. (2004) OWL Web Ontology Language Reference. W3C Recommendation. Disponible en: http://www.w3.org/TR/owl-ref/

Deng, F., Rafiei, D. (2006) Approximately detecting duplicates for streaming data using stable bloom filters. In: Proceedings of the 2006 ACM SIGMOD Intl. Conference on Management of Data, New York, NY, USA, ACM Press, 25-36

Deutsch, A., Fernandez, M., Florescu, D., Levy, A., Suciu, D. (1999) A query language for XML. In Proceedings of the World-Wide Web 8 Conference, pages 1155-1169.

Devogele, T., Parent, C., Spaccapietra, S. (1998) On spatial database integration. International Journal of Geographical Information Science, vol 12 - Issue 4, pp. 335-352.

Devogele, T., Trevisan J., Raynal, L. (1996) Building a Multi-Scale Database with Scale-Transition Relationships. Proceedings of the 7th Spatial Data Handling Symposium. Delft, The Netherlands, pp. 337-351.

Díaz, L., Granell, C., Gould, M. (2008) Spatial Data Integration over the Web. Ferraggine, V. E., Doorn, J.H., Rivero, L.C. (Eds.) Encyclopedia of Database Technologies and Applications, 2nd edition. Information Science Reference (in press, due 2008)

Dittrich, K. R., Jonscher, D. (1999) All Together Now — Towards Integrating the World's Information Systems. In Masunaga, Yoshifumi and Spaccapietra, Stefano, (eds.), Advances in Multimedia and Databases for the New Century, p. 109–123, Kyoto, Japan, World Scientific Press.

Doan, A., Halevy, A.Y. (2005) Semantic-integration research in the database community. AI Magazine, 26 83–94

Dolbear, C., Goodwin, J., Mizen, H., Ritchie, J. (2007) 'Semantic interoperability between topographic data and flood defence ontology'. Technical Report. Ordnance Survey. Available at http://www.ordnancesurvey.co.uk/oswebsite/partnerships/research/publications/docs/2005/Semantic%20Interoperability_CDolbear_sem.pdf

Domenig, R., Dittrich, K. R. (2000) A query based approach for integrating heterogeneous data sources. In Proc. of 9th Int'l Conf. on Information and Knowledge Management.

Dong, X., Halevy, A.Y., Madhavan, J. (2005) Reference Reconciliation in Complex Information Spaces. In: SIGMOD Conference. 85–96

Downs, E., Clare, P., Coe, I. (1998) Structured Analysis and Design Method (SSADM). Prentice Hall, New Jersey

Drafting Team "Data Specifications" (2007) Drafting Team "Data Specifications" – deliverable D2.3: Definition of Annex Themes and Scope. D2.3_v2.0. Disponible en: http://www.eurogeographics.org/eng/documents/inspireDataspecD2_3v2.0.pdf

Duchêne, C., Gaffuri, J. (2008) Combining Three Multi-agent Based Generalisation Models: AGENT, CARTACOM and GAEL. In Headway in Spatial Data Handling. 13th SDH'08. Lecture Notes in Geoinformation and Cartography, Ruas, A., Gold, C. (Eds.) pp. 277-296. ISBN: 978-3-540-68565-4. Montpellier, France.

Duschka, O. M., Genesereth, M. (1997a) Answering recursive queries using views. In Proc. of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS). Tucson, Arizona.

Duschka, O. M., Genesereth, M. (1997b) Query planning in info-master. In Proceedings of the ACM Symposium on Applied Computing, San Jose, CA.

Egenhofer, M. J. (2002) "Toward the Semantic Geospatial Web", 10th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS-2002). McLean, Virginia, USA, ACM Press, 1–4.

Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S. (2007) Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1), 1-16.

Elmagarmid, A., Rusinkiewicz, M., Sheth, A. (eds.) (1999) Management of Heterogeneous and Autonomous Database Systems. Morgan Kaufmann.

Elmasri, R., Navathe, S.B. (2000) Fundamentals of Database Systems. Addison Wesley Longman, 3 edition.

Espinoza, M., Gómez-Pérez, A., Mena, E. (2008) Enriching an ontology with multilingual information. In Proc. of 5th European Semantic Web Conference (ESWC'08), Tenerife, (Spain).

ESRI (1998) ESRI Shapefile Technical Description. An ESRI White Paper. Disponible en: http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

ESRI (2003) Spatial Data Standards and GIS Interoperability. ESRI White Papers, ww.isotc211.org/Outreach/Newsletter/Newsletter_04_2004/Appendix_4.pdf

Essid, M. O. Boucelma, O., Colonna, F. M., Lassoued, Y. (2004) Query processing in a geographic mediation system. In Proc of ACM-GIS, pages 101–108.

European Commission, (2000a) Adopting a multiannual community programme to stimulate the development and use of European digital content on the global networks and to promote the linguistic diversity in the information society: A proposal for council decision. http://europa.eu.int/ISPO/docs/econtent/COM2000_323_en.pdf

European Commission, (2000b) Commercial exploitation of Europe's public sector information: Executive summary. Technical report, Pira International Ltd., University of East Anglia and Knowledge Ltd., ftp://ftp.cordis.lu/pub/econtent/docs/2000_1558_en.pdf

European Commission (2004) European Interoperability Framework for Pan-European eGovernment Services, version 1.0. ISBN 92-894-8389-X. Belgium. Disponible en http://ec.europa.eu/idabc/servlets/Doc?id=19529

European Environment Agency (1997-1999) CORINE Land Cover. http://reports.eea.europa.eu/COR0-landcover/en

European Parliament and the Council of the European Union (2000) Directive 2000/60/EC of the European Parliament and of the Council for establishing a framework for Community action in the field of water policy. Brussels, p 72.

European Parliament and the Council of the European Union (2003) Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive/psi_directive_en.pdf (último acceso Enero 2011).

Euzenat, J., Shvaiko, P. (2007) Ontology Matching. Springer-Verlag, Berlin Heidelberg. ISBN: 3-540-49611-4

Euzenat, J. (2001) Towards a principled approach to semantic interoperability. In: Gómez-Pérez A., Grüninger M, Stuckenschmidt H, Uschold M. (eds.) IJCAI 2001 Workshop on Ontology and Information Sharing, Seattle, Washington.

Fagin, R., Kolaitis, P. G., Miller, R. J., Popa, L. (2003) Data exchange: Semantics and query answering. In Proceedings of the 9th International Conference on Database Theory (ICDT'03), pages 207–224. Springer-Verlag.

Fagin, R., Kolaitis, P. G., Popa, L. (2005) Data exchange: Getting to the core. ACM Transactions on Database Systems 30(1): 174-210.

Federal Geographic Data Committee (FGDC) (2001) "Content Standard for Digital Geospatial Metadata", vol. 2001, FGDC-STD-001-1998, ed. National Spatial Data Infrastructure.

Fellegi, I. P., and Sunter, A. B. (1969) "A Theory for Record Linkage," Journal of the American Statistical Association, 40, 1183-1210.

Fernández-López M, Gómez-Pérez A, Pazos A, Pazos J (1999) Building a Chemical Ontology Using Methontology and the Ontology Design Environment. IEEE Intelligent Systems & their applications 4(1):37–46

Fielding, J.M., Simon, J., Ceusters, W., Smith, B. (2004) 'Ontological Theory for Ontological Engineering: Biomedical Systems Information Integration', Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning (KR2004), Whistler, BC.

Fileto, R. (2001) Issues on Interoperability and Integration of Heterogeneous Geographical Data. Proceedings of 3rd Brazilian Symposium on GeoInformatics.

Fonseca, F., Câmara, G., Monteiro, A. M. (2006) 'A Framework for Measuring the Interoperability of Geo-Ontologies', Spatial Cognition and Computation, vol. 6, no. 4, pp. 307-329.

Fonseca, F., Davis, C., Câmara, G. (2003) Bridging Ontologies and Conceptual Schemas in Geographic Information Integration. GeoInformatica, Volume 7, Number 4, pp. 355-378(24). Springer

Fonseca, F., Egenhofer, M. (1999) Ontology-Driven Geographic Information Systems. In Medeiros, C. B. (ed.) 7th ACM Symposium on Advances in Geographic Information Systems, Kansas City, MO, pp.14-19.

Fonseca, F., Egenhofer, M., Agouris, P. Câmara, C. (2002) Using ontologies for integrated geographic information systems. Transactions in GIS 6(3), p.231-257.

Fonseca, F., Egenhofer, M., Davis, C. (2000) Ontology-Driven Information Integration. In The AAAI—2000 Workshop on Spatial and Temporal Granularity.

Fonseca, F., Egenhofer, M., Davis, C., Câmara, G. (2002) Semantic Granularity in Ontology-Driven Geographic Information Systems. AMAI Annals of Mathematics and Artificial Intelligence - Special Issue on Spatial and Temporal Granularity 36(1-2): pp. 121-151.

Fontanillo, E. (1986) Diccionario de Geografia. Ed. Anaya. Madrid.

Frakes, W. B., Baeza Yates, R. (ed.) (1992) Information Retrieval: data structures and Algorithms. Mexico: Prentice-Hall.

Friedman, M., Levy, A., Millstein, T. (1999) Navigational plans for data integration. In Proceedings of the National Conference on Artificial Intelligence. (AAAI'99), pages 67–73. AAAI Press/The MIT Press, 1999.

Friedman, M., Weld, D. (1997) Efficient execution of information gathering plans. In Proceedings of the International Joint Conference on Artificial Intelligence. Nagoya, Japan.

Friis-Christensen, A., Schade, S., Peedell, S. (2005) Approaches to solve schema heterogeneity at the European level. In K. Fullerton, editor, Abstract book of 11th EC-GI&GIS Workshop -- ESDI: Setting the Framework, pages 14-16, Alghero, Sardinia (Italy).

Galhardas, H., Florescu, D., Shasha, D., Simon, E. (2000) An extensible framework for data cleaning. In ICDE, 312.

Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J. (2006) Modelling ontology evaluation and validation. Proceedings of the 3rd European Semantic Web Conference (ESWC2006), LNCS Volume 4011/2006, 140-154.

Ganter B, Wille R (1997) Formal Concept Analysis: Mathematical Foundations. Springer-Verlag New York, Inc., Secaucus, NJ, USA Translator-C. Franzke.

Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., Widom, J. (1997) The TSIMMIS Project: Integration of Heterogeneous Information Sources. Journal of Intelligent Information Systems, 8 (2): 117-132.

Ghawi, R., Cullot, N. (2007) Database-to-Ontology Mapping Generation for Semantic Interoperability", Third International Workshop on Database Interoperability (InterDB 2007), held in conjunction with VLDB 2007, Vienna, Austria.

Glaser, H., Jaffri, A. and Millard, I. (2009) Managing Co-reference on the Semantic Web. In: WWW2009 Workshop: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

Goasdoue, F., Lattes, V., Rousset, M.C. (2000) The use of CARIN language and algorithms for information integration: The PICSEL system. International Journal of Cooperative Information Systems, 9(4):383–401

Gobierno de La Rioja (2005) Catálogo de topónimos 1:5.000, Gobierno de La Rioja.

Gobierno de Navarra (2000) Revisión de la metodología de producción y actualización de la cartografía a escala 1:5.000', Gobierno de Navarra. Departamento de Obras Públicas, Transportes y Comunicaciones.

Gobierno Vasco (2005) Relación de formas gráficas a escala 1:10.000 y 1:25.000', Gobierno vasco. Departamento de Ordenación del Territorio y Medio Ambiente.

Goble, C., Stevens, R., Bechhofer, G. Ng, S., Paton, N., Baker, P., Peim, M., Brass, A. (2001) "Transparent Access to Multiple Bioinformatics Information Sources," IBM Systems Journal, vol. 40, no. 2, pp. 534–551.

Goh, C. H. (1997) "Representing and reasoning about semantic conflicts in heterogeneous information sources," Ph.D. dissertation, Massachusetts Institute of Technology (MIT).

Goh, C.H., Bressan, S., Siegel, M. and Madnick, S. E. (1999) Context Interchange: New Features and Formalisms for the Intelligent Integration of Information. ACM Transactions on Information Systems, Vol. 17(3), 270–293.

Goh, C. H., Madnick, S. E., Siegel, M. (1994) Context Interchange: Overcoming the Challenges of Large-Scale Interoperable Database Systems in a Dynamic Environment. In Third International Conference on Information and Knowledge Management (CIKM 1994), pages 337–346, Gaithersburg, USA, November 29 – December 2. ACM.

Gómez-Pérez, A. (2004) Ontology Evaluation. Handbook on Ontologies. International Handbooks on Information Systems. Springer.

Gómez-Pérez, A., Fernández-López, M., Corcho, O. (2003) Ontological Engineering. Springer Verlag. Advanced Information and Knowledge Processing series. ISBN 1-85233-551-3.

Gómez-Pérez, A., Juristo, N., Montes, C., Pazos, J. (1997) Ingeniería del Conocimiento. Editorial Centro de Estudios Ramón Areces, S.A. ISBN: 84-8004-269-9.

Gómez-Pérez, A., Ramos, J.A., Rodríguez-Pascual, A., Vilches-Blázquez, L.M. (2008) The IGN-E Case: Integrating through a hidden ontology. In Headway in Spatial Data Handling. 13th SDH'08. Lecture Notes in Geoinformation and Cartography, Ruas, A., Gold, C. (Eds.) pp. 417-435. ISBN: 978-3-540-68565-4. Montpellier, France.

Goodchild, M.F. (1991) Issue of quality and uncertainty, in Advances in Cartography, Müller Ed., Barking, Essex/ Elsevier pp 113-139.

Goodchild, M., Egenhofer, M., Fegeas, R. (1997) Interoperating GISs: Report of a Specialist Meeting Held under the Auspices of the Varenius Project, Panel on Computational Implementations of Geographic Concepts. National Center for Geographic Information and Analysis, Santa Barbara. California.

Goodchild, M., Egenhofer, M. J., Fegeas, R. Kottman, C. Eds. (1999) Interoperating Geographic Information Systems. The International Series in Engineering and Computer Science, Kluwer.

Goodwin, J., Dolbear, C., Hart, G. (2008) Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web. Transactions in GIS, Volume 12, Supplement 1, pp. 19-30(12) Blackwell Publishing.

Govorov, M.O. (1995) Representation of the generalized data structures for Multi-Scale GIS, in 17 th ICA/ACI Barcelona pp 2491-2495.

Granell, C., Gould, M., Manso, M.A., & Bernabé, M.A. (2007). Spatial Data Infrastructures. In: Karimi, H. (ed.), Handbook of Research on Geoinformatics. Hershey: Information Science Reference.

Green, J., Dolbear, C., Hart, G., Engelbrecht, P., Goodwin. J. (2008) Creating a semantic integration system using spatial data. ISWC 2008, Karlsruhe, Germany.

Greenwood, E. (1973) Metodología de la investigación social. Paidós, Buenos Aires, Argentina.

Greenwood J, Hart G. (2003) "Sharing Feature Based Geographic Information - A Data Model Perspective". En 7th Int'l Conference on GeoComputation. United Kingdom.

Grenon, P., Smith, B. (2004) SNAP and SPAN: Towards Dynamic Spatial Ontology. Spatial Cognition and Computation, 4(1), 69-104.

Gruber, T. R. (1993) A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition. #5: 199-220.

Gruber, T.R. (1995) Toward principles for the design of ontologies used for knowledge sharing, International Journal of Human-Computer Studies, v.43 n.5-6.

Grüninger, M., Fox, M. S. (1994) The role of competency questions in enterprise engineering. In Proceedings of the IFIP WG5.7 Workshop on Benchmarking - Theory and Practice, Trondheim, Norway.

Grüninger, M., Uschold, M. (2007) Ontologies and Semantic Integration, in Software Agents for the Warfighter, Information Technology Assessment Consortium (ITAC). Edited by Jeff Bradshaw, Institute for Human and Machine Cognition (IHMC), University of West Florida.

Gu, L., Baxter, R., Vickers, D., Rainsford, C. (2003) Record linkage: Current practice and future directions. Technical Report 03/83, CSIRO Mathematical and Information Sciences.

Guan, J. H., S. G. Zhou, J. P. Chen, X. L. Chen, Y. An, W. Yu, R. Wang, X. J. Liu (2003) Ontology-Based GML Schema Matching For Spatial Information Integration. Proceedings of the Second International Conference on Machine Learning and Cybernetics, Volume 4, Issue, 2-5 Page(s): 2240 – 2245. Xian, China.

Guarino, N. (1998) Formal Ontology in Information Systems. 1st International Conference on Formal Ontology in Information Systems (FOIS'98). IOS Press. p.3-15.

Gupta, A., Marciano, R., Zaslavsky, I., Baru, C. (1999) Integrating GIS and imagery through xml-based information mediation. In Proc of NSF International Workshop on Integrated Spatial Databases: Digital Images and GIS.

Gutierrez, C., Hurtado, C.A., Vaisman A. A. (2007) Introducing Time into RDF. IEEE Trans. Knowl. Data Eng. 19(2): 207-218.

Güting, R.H. (1994) An Introduction to Spatial Database Systems. The International Journal on Very Large Databases (VLDB) 3(4): pp.: 357-399.

Haase, P., Brockmans, S., Palma, R., Euzenat, J., d'Aquin. M. (2007) NeOn D1.1.2: Updated Version of the Networked Ontology Model. NeOn project. http://www.neon-project.org.

Haase, P., Rudolph, S., Wang, Y., Palma, R., Euzenat, J., d'Aquin, M. (2006) NeOn D1.1.1: Networked Ontology Model. NeOn project. http://www.neon-project.org.

Hakimpour F (2003) "Using Ontologies to Resolve Semantic Heterogeneity for Integrating Spatial Database Schemata". Ph.D. thesis, Zurich University. Switzerland.

Hakimpour, F., Geppert, A. (2001) Resolving semantic heterogeneity in schema integration: an Ontology based Approach. In: Proc. of Int. Conf. on Formal Ontologies in Information Systems FOIS'01 (Eds. Ch. Welty and B. Smith), New York, ACM Press, pp. 297-308.

Halevy, A. Y., Ives, Z., Suciu, D., Tatarinov, I. (2003) Schema mediation in peer data management systems. In Proc. of the 19th IEEE Int. Conf. on Data Engineering (ICDE 2003), pages 505-516.

Halevy, A. Y., Rajaraman, A., Ordille, J. J. (2006) Data Integration: The Teenage Years. In International Conference on Very Large Data Bases (VLDB), pages 9–16.

Hammer, J., McLeod, D. (1993) An approach to resolving semantic heterogeneity in a federation of autonomous, heterogeneous database systems. Journal for Intelligent and Cooperative Information Systems, 2(1):51–83.

Harpring, P. (1997) 'Proper words in proper places: The Thesaurus of Geographic Names', Museum Document Association Information, vol. 2, no. 3, pp. 5-12.

Harpring, P., Beecroft, A., Johnson, R., Ward, J. (2006) 'TGN Editorial Guidelines: About TGN'.http://www.getty.edu/research/conducting_research/vocabularies/guidelines/tgn_1_contents_intro.pdf

Hart, G., Dolbear, C. (2007) What's So Special about Spatial? In the Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 Are Shaping the Network Society, ed. Arno Scharl and Klaus Tochtermann. Advanced Information and Knowledge Processing Series. London: Springer.

Hart, G., Dolbear, C., Goodwin, J., Kovacs, K. (2007) 'Domain Ontology Development'. Technical Report I-0003

Hasselbring, W. (2000) Information system integration. Communications of the ACM, 43(6):33–38.

Haunert, J.-H., Sester, M. (2008) Area collapse and road centerlines based on straight skeletons, GeoInformatica 12 (2), pp. 169–191

Heath, McArthur, Vetter, (2005) Metadata lessons from the iLumina digital library, Communications of the ACM 48 (7) p. 68–74.

Hernández, M.A., Stolfo, S.J. (1998) Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery 2(1) 9-37

Hernandez, T., Kambhampati, S. (2004) Integration of Biological Sources: Current Systems and Challenges Ahead. SIGMOD Record, 33.

Herzog, T.N., Scheuren, F.J., Winkler, W.E. (2007) Data Quality and Record Linkage Techniques. Springer.

Hey, T., Tansley, S., Tolle, K. (2009) The Fourth Paradigm. Data-Intensive Scientific Discovery. Microsoft Research.

Hiramatsu, K., Reitsma, F. (2004) GeoReferencing the Semantic Web: ontology based markup of geographically referenced information. Paper presented at the Joint EuroSDR/ Euro-Geographics workshop on Ontologies and Schema Translation Services, Paris, France.

Hodge, G. (2000) Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub91/contents.html.

Hodnett, E. (1955) The Art of Problem Solving. New York. Harper & Row Publishers.

Hu, Y.H, Ge, L. (2007) "A Supervised Machine Learning Approach to Toponym Disambiguation." In The Geospatial Web: How Geobrowsers, Social Software, and the Web 2.0 Are Shaping the Network Society, ed. Scharl, A., Tochtermann, K. 117-128. Springer.

Hull, R. (1997) Managing semantic heterogeneity in databases: a theoretical prospective Symposium on Principles of Database Systems. Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. Tucson, Arizona, United States, pp.: 51 – 61. ISBN:0-89791-910-6

Hultgren, T., Dai, Q. (2006) Rapid Deployment of EPA Enterprise Geospatial Data Warehouse in Katrina Aftermath. ESRI International Users Group Conference.

Hurson, A. R., Bright, M. W. (1991) Multidatabase Systems: An Advanced Concept in Handling Distributed Data. Advances in Computers, 32:149–200.

IEEE (1990) Standard Glossary of Data Management Terminology. IEEE Std 610.5.

IEEE (1990b) Standard Glossary of Software Engineering Terminology. IEEE Std 610.12 (Revision and redesignation of IEEE Std 792-1983).

IEEE (1995) Guide for Developing Software Life Cycle Processes. IEEE Std 1074.1.

IEEE (1996) Standard for Developing Software Life Cycle Processes. IEEE Computer Society. New York. IEEE Std 1074-1995

IEEE (1997) Standard for Developing Software Life Cycle Processes. IEEE Std 1074. (Revision of IEEE Std 1074-1995; Replaces IEEE Std 1074.1-1995).

IEEE (1999) Standard for Information Technology. Software Life Cycle Processes. Reuse Processes. IEEE Std 1517-1999 (R2004).

Inmon, W. H. (1996) Building the Data Warehouse. John Wiley and Sons, New York.

Institut Cartogràfic de Catalunya (2001) Diccionario de datos de la Base topográfica 1:5.000 v2.0 (BT-5M)'.

Institute Geographique National France International, Institut Géographique National France, Centro Nacional de Información Geográfica Spain and University of Zaragoza. (2005a) Common Model Activity - Final report. SDIGER: A cross-border interadministration SDI to

support WFD information access for Adour-Garonne and Ebro River Basins. http:/www.idee.es/sdiger/public_docs/CommonModelActivity_FinalReport_v0.3.pdf

Institute Géographique National France International, Institut Géographique National France, Centro Nacional de Información Geográfica and University of Zaragoza. (2005b). SDIGER UML models. http://www.idee.es/sdiger/public_docs/ThemWaterUMLModels.pdf

Institute of Electrical and Electronics Engineers (IEEE) (1990) "Interoperability" in IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries.

Isaac, A. van der Meij, L. Schlobach, S. Wang, S. (2007) An Empirical Study of Instance-Based Ontology Matching. Proceedings of the 6th ISWC, LNCS 4825, 253-266.

ISO 639-1 (2002) ISO 639-1:2002 Codes for the representation of names of languages -- Part 1: Alpha-2 code

ISO 639-2 (1998) International Organization for Standardization ISO 639-2:1998 Codes for the representation of names of languages -- Part 2: Alpha-3 code

ISO 19103 (2005) International Organization for Standardization 19103:2005 Geographic information -- Conceptual schema language.

ISO 19112 (2003b) International Organization for Standardization 19112:2003 'Geographic information -- Spatial referencing by geographic identifiers'.

ISO 19115 (2003) International Organization for Standardization 19115:2003. Geographic information - Metadata.

ISO 19119 (2002) International Organization for Standardization/DIS 19119. Geographic Information Services.

ISO 19136 (2007) International Organization for Standardization ISO 19136 Geographic information - Geography Markup Language (GML)

ISO 19139 (2007) International Organization for Standardization 19139:2007 Geographic information -- Metadata -- XML schema implementation

ISO (1985) International Organization for Standardization 5964:1985. Documentation - Guidelines for the establishment and development of multilingual thesauri.

ISO (1986) International Organization for Standardization 2788:1986. Guidelines for the establishment and development of monolingual thesauri.

ISO 2382-1 (1993) International Organization for Standardization 2382-1:1993 Information technology – Vocabulary – Part 1: Fundamental terms.

ISO (2005), International Organization for Standardization 19110:2005 'Geographic Information – Methodology for feature cataloguing'.

ISO (2005a) International Organization for Standardization 19110. "Geographic Information" – Methodology for feature cataloguing.

ISO (2005b) International Organization for Standardization 19128: Geographic information -Web map server interface.

ISO (2005c) International Organization for Standardization 19118: Geographic information – Encoding.

ISO (2007a) International Organization for Standardization ISO/CD 19142 Geographic information - Web Feature Service. Edition 2.

Ives, Z., Florescu, D., Friedman, M., Levy, A., Weld, D. (1999) An adaptive query execution engine for data integration. In Proc. of ACM SIGMOD Conf. on Management of Data, pages 299-310.

Ives, Z.G., Halevy, A. Y., Weld, D. S. (2001) Integrating Network-Bound XML Data. IEEE Data Eng. Bull., vol. 24 (2), p. 20-26.

Ives, Z., Halevy, A., Weld, D. (2003) An xml query engine for network-bound data. VLDB Journal, Special Issue on XML Query Processing.

Jaffri, A., Glaser, H., and Millard, I. (2007) URI Identity Management for Semantic Web Data Integration and Linkage. In Proceedings of the Workshop on Scalable Semantic Web Systems (Vilamoura, Portugal) Springer.

Jaffri, A., Glaser, H. and Millard, I. (2008) URI Disambiguation in the Context of Linked Data. In : Linked Data on the Web (LDOW2008), Beijing, China.

Jang, S.-G., Kim, T. J. (2006) Modeling an Interoperable Multimodal Travel Guide System using the ISO 19100 Series of International Standards. Proceedings of the 14th ACM International Symposium on Advances in Geographic Information Systems, Arlington, Va., Silvia Nittel and Rolf A. de By (eds.), ACM Press, 115-122.

Jang, S. G., Kim, T. J. (2007) Semantically Interoperable Geospatial Information Processing: A Review and A Canonical Model Approach. In proceedings of the 10th AGILE International Conference on Geographic Information Science. Aalborg, Denmark.

Jaro, M. A. (1989) Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association 84:414–420.

Jaro, M. A. (1995) Probabilistic linkage of large public health data files (disc: P687-689). Statistics in Medicine 14:491–498.

Jefferey, S.R., Alonso, G., Franklin, M.J., Hong, W., Widom, J. (2006) Declarative support for sensor data cleaning. In: Proceedings of the 4th International Conference on Pervasive Computing, Dublin, Ireland, Springer, 83-100

Jeusfeld, M. A., Quix, C., Jarke, M. (1998) "Design and analysis of quality information for data warehouses," in International Conference on Conceptual Modeling / the Entity Relationship Approach, pp. 349–362.

Junta de Castilla y León (2005) Diccionario de datos cartografía 1:5.000', versión 1.1.

Junta de Extremadura (2005) 'Codificación de la cartografía Digital', Junta de Extremadura. Dirección General de Urbanismo, Arquitectura y Ordenación del Territorio.

Kalfoglou, Y., Schorlemmer, M. (2003) Ontology Mapping: The State of the Art. The Knowledge Engineering Review 18(1), 1-31

Kashyap, V., Sheth, A. (1996) Semantic and schematic similarities between database objects: a context-based approach, The VLDB Journal — The International Journal on Very Large Data Bases, v.5 n.4, p.276-304.

Keßler, C., Raubal, M., Wosniok, C. (2009) "Semantic Rules for Context-Aware Geographical Information Retrieval" 4th European Conference on Smart Sensing and Context. Berlin

Kieler, B., Huang, W., Haunert J.-H., Jiang, J. (2009) Matching River Datasets of Different Scales , Advances in GIScience: Proceedings of 12th AGILE Conference on GIScience, Lecture Notes in Geoinformation and Cartography , p. 135-154 , Springer, Berlin.

Kim, W., Choi, I., Gala, S., Scheevel, M. (1995) Modern Database: The Object Model, Interoperability, and Beyond, chapter On Resolving Schematic Heterogeneity in Multidatabase Systems, pages 521–550. ACM Press / Addison-Wesley Publishing Company.

Kim, W., Seo, J. (1991) Classifying schematic and data heterogeneity in multidatabases systems. IEEE Computer, 24(2), 12-18.

Kimball, R., Ross, M. (2002) The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. John Wiley.

King, R.L., Durbha, S., Younan, N.H. (2005) Interoperability in Costal Zone Monitoring Systems: Resolving Semantic Heterogeneities Through Ontology Driven Middleware, In Proc. of 31st International Symposium on Remote Sensing of Environment.

Kirk, T., Levy, A., Sagiv, Y., Srivastava, D. (1995) The information manifold. In Proceedings of the AAAI Spring Symposium: Information Gathering from Heterogeneous Distributed Environments, pages 85-91.

Klien, E., Lutz, M., Einspanier, U., Hübner, S. (2004) An Architecture for Ontology-Based Discovery and Retrieval of Geographic Information. 7th AGILE, Heraklion (Greece).

Klien, E., Probst, F. (2005) Requirements for geospatial ontology engineering. Conference on Geographic Information Science (AGILE 2005) Estoril (Portugal).

KnowledgeWeb Consortium (2005). Specification of a common framework for characterizing alignment. Deliverable 2.2.1. KWEB/2004/D2.2.1/v2.0. Disponible en: http://www.inrialpes.fr/exmo/cooperation/kweb/heterogeneity/deli/kweb-221.pdf

Köpcke, H., Rahm, E. (2009) Frameworks for entity matching: A comparison, Data & Knowledge Engineering. Volume 69 (2), pages 197-210

Kruchten, P., (2000) The Rational Unified Process: An Introduction, Addison Wesley.

Kuhn, W. (2003). Semantic Reference Systems. International Journal of Geographical Information Science, 17(5), 405-409.

Kuhn, W. (2005) Geospatial Semantics: Why, of What, and How? In Spaccapietra, S. Zimányi, E. (Eds.): Journal on Data Semantics III. Lecture Notes in Computer Science, 3534 (3), 1-24.

Kwok, C. T., Weld, D. S. (1996) Planning to gather information. In Proceedings of the AAAI Thirteenth National Conference on Artificial Intelligence.

Lacasta, J., Muro-Medrano, P.R., Nogueras-Iso, J., Zarazaga Soria, F.J. (2005) Web Ontology Service, a key component of a Spatial Data Infrastructure. Proceedings of the 11th EC GI&GIS Workshop: ESDI: ESDI Setting the Framework.

Lacasta, J., Nogueras-Iso, J., Béjar, R., Muro-Medrano, P.R., Zarazaga-Soria, F.J. (2007) A Web Ontology Service to facilitate interoperability within a Spatial Data Infrastructure: Applicability to discovery. Data & Knowledge Engineering, vol. 63(3): p. 947-971.

Lagrange, J.P., Ruas, A. (1994) Geographic information modelling: GIS and generalisation, in Spatial Data Handling, Waugh and Healey Eds., Taylor&Francis, pp 1099-1117.

Landers, T., Rosenberg, R. L. (1982) An Overview of MULTIBASE. In Schneider, Hans-Jochen, editor, Second International Symposium on Distributed Data Bases (DDB 1982), pages 153–184, Berlin, Germany, September 1–3. North-Holland.

Larsen, M. (1999) Multiple imputation analysis of records linked using mixture models. In Statistical Society of Canada Proceedings of the Survey Methods Section, 65–71. Statistical Society of Canada (McGill University, Montreal).

Lassila O, McGuiness D. (2001) The Role of Frame-Based Representation on the Semantic Web. Technical Report. Knowledge Systems Laboratory. Stanford University. KSL-01-02.

Latre, M.A., Zarazaga-Soria, F.J., Nogueras-Iso, J., Béjar, R. Muro-Medrano, P.R. (2005) SDIGER: A cross-border inter-administration SDI to support WFD information access for Adour-Garonne and Ebro River Basins. Proceedings of the 11th EC GI & GIS Workshop, ESDI Setting the Framework. pag 5-7. Alghero, Sardinia (Italia).

Lattes, F. G. V., Rousset, M.-C. (2000) "The Use of CARIN Language and Algorithms for Information Integration: The PICSEL System," International Journal of Cooperative Information Systems, vol. 9, no. 4, pp. 383–401.

Laurini, R. (1998) Spatial multi-database topological continuity and indexing: a step towards seamless GIS data interoperability. International Journal of Geographical Information Science. Vol. 12, No. 4, 373–402.

Leeuwen, J. (1990) Formal Models and Semantics. Handbook of Theoretical Computer Science, XIV (Cambridge, USA: MIT Press).

Lemmens, L., Wytzisk, A., de By, R., Granell, C., Gould, M., van Oosterom, P (2006) Integrating semantic and syntactic descriptions to chain geographic services. IEEE Internet Computing 10(5), 42-52.

Lemmens, R., Vries, M. (2004) Semantic Description of Location Based Services using an Extensible Location Ontology. Paper presented at the Muenster GI-Days. Germany.

Lenzerini, M. (2002) Data Integration: A Theoretical Perspective. In Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Pp.: 233-246. ISBN: 1-58113-507-6. Madison, Winsconsin, USA.

Lesk, M. (2004) Online Data and Scientific Progress: Content in Cyberinfrastructure. Invited talk at the University of Edinburgh. http://archiv.twoday.net/stories/337419/

Levy, A. Y. (2000) Answering Queries Using Views: A Survey. Very Large Database Journal. 10(4):270-294.

Levy, A. Y. (2000) Logic-based techniques in data integration. In Jack Minker, editor, Logic-Based Artificial Intelligence, pages 575-595. Kluwer Academic Publishers, Dordrecht.

Levy, A. Y., Rajaraman, A., Ordille, J. J. (1996) Querying heterogeneous information sources using source descriptions. In Proc. of the Int. Conf. on Very Large Data Bases (VLDB), Bombay, India.

Levy A.Y., Rousset, M.C. (1998) Combining horn rules and description logics in CARIN. Artif. Intell., 104(1-2):165–209.

Lieberman, J. (2006) Geospatial Semantic Web Interoperability Experiment Report. Discussion paper. OGC 06-002r1 Version: 0.5.

Lieberman, J., Singh, R., Goad, R. (2007) W3C Geospatial Ontologies W3C Incubator Group Report 23 October 2007

Lin, K., Ludaescher, B. (2004) GEON: Ontology-Enabled Map Integration. Twenty-fourth Annual ESRI International User Conference, San Diego, CA.

Litwin, W., Abdellatif, A. (1986) Multidatabase interoperability. IEEE Computer, 19(12):10–18.

Lohfink, A., Carnduff, T., Thomas, N. and Ware, J.M., (2007) "An Object-Oriented Approach to the Representation of Spatiotemporal Geographic Features", Proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems (ACM GIS 2007). ISBN: 978-1-59593-914-2.

López-Pellicer, F.J.; Vilches-Blázquez, L.M.; Nogueras-Iso, J.; Corcho, O.; Bernabé, M.A.; Rodríguez, A.F. (2007) Using a hybrid approach for the development of an ontology in the hydrographical domain. In 2nd Workshop Ontologies for urban development. Conceptual Models for Urban Practitioners. J. Teller; C. Tweed; G. Rabino (Eds.) pp. 43 – 56. Bologna.

Luo, Y. W., Liu, X. P., Wang, W. J., Wang, X. L., Xu, Z. Q. (2004) QoS Analysis on Web Service Based Spatial Integration. Lecture Note on Computer Science, 3251: 42-49.

Lutz, M., Klien, E. (2006) Ontology-based retrieval of geographic information Export Find Similar. International Journal of Geographical Information Science, Vol. 20, No. 3. 233-260.

Lutz, M., Sprado, J., Klien, E., Schubert, C., Christ, I. (2009) Overcoming semantic heterogeneity in spatial data infrastructures. Computers & Geosciences, Vol. 35 (4) 739-752.

Madhavan, J., Bernstein, P. A., Rahm, E. (2001) Generic schema matching with cupid. In Apers, P. M. G., Atzeni, P., Ceri, S., Paraboschi, S., Ramamohanarao, K., and Snodgrass, R. T., editors, VLDB 2001, Proceedings of 27th International Conference on Very Large Databases, September 11-14. Roma, Italy, pages 49–58. Morgan Kaufmann.

Manolescu, I., Florescu, D., Kossmann, D. (2001) Answering xml queries on heterogeneous data sources. In Proc. of VLDB, pages 241¬-250.

Manso, M.A., Wachowicz, M. (2009) GIS Design: A Review of Current Issues in Interoperability. Geography Compass, Vol. 3 (3), 1105 – 1124.

Mark (1989) Multiple views of multiple representations, in NCGIA Initiative 3 Workshop on Multiple Representations, pp 68-71.

Mark, D.M., Turk, A.G. (2003a) Ethnophysiography. Workshop on Fundamental Issues in Spatial and Geographic Ontologies, 23 September 2003 (in conjunction with COSIT03), http://www.geog.buffalo.edu/ncgia/ethnophysiography/PreCOSIT-Ethnophysiography.pdf

Mark, D.M., Turk, A.G. (2003b) Landscape Categories in Yindjibarndi: Ontology, Environment, and Language. in Spatial Information Theory - Foundations of Geographic Information Science. COSIT 2003, Kartause Ittingen, Switzerland. 2003: Springer, Lecture Notes in Computer Science 2825: pp. 31-49.

McGuinness, D., Fox, P., Brodaric, B., and Kendall, E. (2009) The Emerging Field of Semantic Scientific Knowledge Integration. IEEE Intelligent Systems 24(1):25-26.

McGuinness, D.L., van Harmelen, F. (2004) OWL Web Ontology Language Overview. W3C Recommendation. http://www.w3.org/TR/owl-features/

McKee (2000) Who wants a GDI?, In Geospatial Data Infrastructure – Concepts, cases and good practice, New York, Oxford University Press, pag. 13-24.

Meersman, Zahir Tari (Eds.): On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS, OTM Confederated International Conferences CoopIS, DOA, ODBASE, GADA, and IS 2007, Vilamoura, Portugal, November 25-30, 2007, Proceedings, Part II. Lecture Notes in Computer Science 4804 Springer 2007, ISBN 978-3-540-76835-7

Mena, E., Illarramendi, A., Kashyap, V., Sheth, A. (2000) "OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies", International Journal on Distributed And Parallel Databases (DAPD), 8(2):223-272, Kluwer Academic Publishers, ISSN 0926-8782.

Menestrina, D., Benjelloun, O., Garcia-Molina, H. (2006) Generic Entity Resolution with Data Confidences. First International VLDB Workshop on Clean Databases, Seoul (Korea).

MÉTRICA v.3. MÉTRICA versión 3. Metodología de planificación, desarrollo y mantenimiento de sistemas de información. Informe Técnico, Ministerio para las Administraciones Publicas, España. http://www.map.es/csi/metrica3.

Michalowski, M., Ambite, J.L., Knoblock, C. A., Minton, S., Thakkar, S., Tuchinda, R. (2004) "Retrieving and Semantically Integrating Heterogeneous Data from the Web," IEEE Intelligent Systems, vol. 19, no. 3, pp. 72-79.

Mizoguchi, R., Vanwelkenhuysen, J., Ikeda, M., (1995) Task Ontology for reuse of problem solving knowledge. Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing (KBKS'95). IOS Press. p.46-57.

Moliner, M. (1984) Diccionario de uso del español. Madrid: Editorial Gredos. p. 1590. ISBN: 84-249-1344-2

Monge, A., Elkan, C. (1997) An efficient domain-independent algorithm for detecting approximately duplicate database records. In The proceedings of the SIGMOD 1997 workshop on data mining and knowledge discovery.

Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-P érez, A., Peters, W. (2008) Modelling multilinguality in ontologies. In Proceedings of the 22nd International Conference on Computational Linguistics, Coling 2008, Companion volume - Posters and Demonstrations, Manchester, UK, pp. 67-70.

Morocho, V., Saltor, F., Pérez Vidal, L. (2001) Hacia la Integración de Bases de Datos Espaciales. In proceedings of Jornadas de Sistemas de Información Geográfica, p. 69-84, Almagro. Spain.

Morocho, V., Saltor, F., Pérez-Vidal, L (2003) Ontologies: Solving semantic heterogeneity in federated spatial database system. In Proceedings of 5th International Conference on Enterprise Information System (ICEIS). Angers, France.

Morris, A., Velegraki, Y., and Bouquet, P. (2008) Entity Identification on the Semantic Web. Proceedings of the 5th Workshop on Semantic Web Applications and Perspectives (SWAP2008), Roma (Italy).

Motta, E., Sabou, M. (2006) "Next Generation Semantic Web Applications," Proc. 1st Asian Semantic Web Conf., LNCS 4185, Springer, pp. 24–29.

Müller, J.C., Lagrange, J.P., Weibel, R., Salgé, F. (1995) Generalisation: state of the art and issues, in GIS and GENERALISATION, Müller, Lagrange and Weibel Eds., Taylor & Francis pp 3-17.

MurMur project (2002) "Multi-representations and multiple resolutions in geographic databases" IST Programme of the European Community - Project 10723. http://lbdwww.epfl.ch/e/MurMur

Murray, D. (2004) GML – User Perspectives. Map Asia 2004 Conference, Beijing, China.

Murray, K. Shiell, D. (2004) A Framework for Geographic Information in Great Britain. The Cartographic Journal, 41(2): 123-129.

Naughton, J., DeWitt, D., Maier, D., Aboulnaga, A., Chen, J., Galanis, L., Kang, J., Krishnamurthy, R., Luo, Q., Prakash, N., Ramamurthy, R., Shanmugasundaram, J., Tian, F., Tufte, K., Viglas, S., Wang, Y., Zhang, C., Jackson, B., Gupta, A., Chen, R. (2001) The Niagara Internet query system. IEEE Data Engineering Bulletin.

Navarrete, A. (2006) Semantic integration of thematic geographic information in a multimedia context. PhD Thesis, Universidad Pompeu Fabra. Barcelona (Spain)

Naiman, C. F., Ouksel, A. M. (1995) A classification of semantic conflicts in heterogeneous database systems. Journal of Organizational Computing, Volume 5, Issue 2, pp. 167 - 193.

Nerbert (2001) Developing Spatial Data Infrastuctures: The SDI Cookbook, Version 1.1, Global Spatial Data Infrastructure, Technical Committee.Date, C. J. (1993) Introducción a los sistemas de bases de datos, vol., 1. Addison-Wesley Iberoamericana, 5 edition.

Newcombe, H. B. (1967) Record linking: The design of efficient systems for linking records into individual and family histories. American Journal of Human Genetics, 19(3):335-359.

Newcombe, H. B., Kennedy, J. M., Axford, S.J., James, A.P. (1959) Automatic linkage of vital records. Science, 130(3381):954-959.

Nikolov, A., Uren, V. S., Motta, E., De Roeck, A. N. (2008) Refining Instance Coreferencing Results Using Belief Propagation. In Proceedings of ASWC'2008. pp.405~419

Nilsson, M., Baker, T., Johnston, P. (2008) Interoperability levels for Dublin Core metadata. Disponible en: http://dublincore.org/architecturewiki/InteroperabilityLevels

Nowak, J., Nogueras-Iso, J., Peedell, S. (2005) Issues of multilinguality in creating a European SDI – The perspective for spatial data interoperability, in: Proceedings of the 11th EC GI & GIS Workshop, ESDI Setting the Framework. Alghero, Italy.

Noy, N.F. (2004) Semantic integration: A survey of ontology-based approaches. SIGMOD Rec. 33(4) 65-70

OGC (2003) 'OpenGIS Reference Model Version 0.1.2', Open Geospatial Consortium, OGC Inc. Wayland, MA, USA.

OGC (2004) Geolinked Data Access Service (GDAS), Version: 0.9.1, OGC Inc. Wayland, MA, USA.

Open Archives Initiative http://www.openarchives.org/ (último acceso Enero 2011)

OpenGIS Consortium (1999a) Semantics and Information Communities, The OpenGIS Abstract Specification Topic 14, version 4. Wayland. http://www.opengeospatial.org/docs/99-114.pdf.

OpenGIS Consortium (1999b) Features. The OpenGIS Abstract Specification Topic 5, version 4. Wayland.

Open Geospatial Consortium (2005) Web Feature Service Implementation Specification. Version: 1.1.0.http://portal.opengeospatial.org/files/index.php?artifact_id=4756&passcode=ty3ncbxhmqp4e0um190z

Open Geospatial Consortium (2006) The OpenGIS® Abstract Specification Topic 6: Schema for coverage geometry and functions. Version: 7.0.0 http://portal.opengeospatial.org/files/index.php?artifact_id=19820&passcode=ty3ncbxhmqp4e0um190z

Open Geospatial Consortium (2008) Web Coverage Service (WCS) Implementation Standard. Version: 1.1.2, Version 1.1 Corrigendum 2 release. http://portal.opengeospatial.org/files/index.php?artifact_id=27297&passcode=ty3ncbxhmqp4e0um190z

Open GIS Consortium (OGC) (2002) Styled Layer Descriptor Implementation Specification. Version: 1.0.0.

Open GIS Consortium (OGC) (2001) Filter Encoding Implementation Specification. Version: 1.0.0.

Open Society Institute (2001) Budapest Open Access Initiative. http://www.soros.org/openaccess/ (último acceso Enero 2011)

Ordnance Survey (2001) Master Map real-world object catalogue. http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/faqs/Docs/realWorldObjectCatalogue.pdf

Oren, R., Delbru E., Decker, S. (2006) Textending faceted navigation for RDF data. In the International Semantic Web Conference 06, pages 559–572.

Ostländer, N., Tegtmeyer, S., Foerster, T. (2005) Developing an SDI for time-variant and multilingual information dissemination and data distribution. In Proceedings of EC-GI & GIS Workshop, ESDI: Setting the Framework, Alghero, Sardinia.

OTM Conferences 2007 Additional Reviewers

Ouksel, A. M., Sheth, A. P. (1999) Semantic Interoperability in Global Information Systems: A Brief Introduction to the Research Area and the Special Section. SIGMOD Record,28/1:5–12

Padgham, L., Winikoff, M. (2002) "Prometheus: A Methodology for Developing Intelligent Agents," in Proceedings of the Third International Workshop on Agent-Oriented Software Engineering (AAMAS 2002), Bologna, Italy.

Palopoli, L., Pontieri, L., Terracina, G., Ursino, D. (2000) Intensional and extensional integration and abstraction of heterogeneous databases. Data & Knowledge Engineering, 35, 201-237.

Palopoli, L., Sacca, D., Ursino, D. (1999) Semi-automatic techniques for deriving interscheme properties from database schemes. Data and Knowledge Engineering, 30(3):239–273.

Papakonstantinou Y., Abiteboul, S., Garcia-Molina, H. (1996a) Object fusion in mediator systems. In Proc. Of the Int. Conf. on Very Large Data Bases (VLDB), Bombay, India.

Papakonstantinou, Y., Borkar, V., Orgiyan, M., Stathatos, K., Suta, L., Vassalos, V., Velikhov, P. (2003) XML queries and algebra in the Enosys integration platform Data & Knowledge Engineering, Vol. 44 Issue 3. Elsevier Science Publishers B. V.

Papakonstantinou, Y., Garcia-Molina, H., Ullman, J. (1996b) Medmaker: A mediation system based on declarative specifications. In International Conference on Data Engineering, pages 132–141, New Orleans.

Paradela, L. (2001) PhD Thesis: Una Metodología para la Gestión del Conocimiento. Spain. Universidad Politécnica de Madrid.

Parent, C., Spaccapietra, S., Zimányi, E. (2006) The MurMur project: modeling and querying multi-representation spatio-temporal databases. Inf. Syst. 31, 8; p. 733-769.

Peng, Z. R. (2005) A proposed framework for feature-level geospatial data sharing: a case study for transportation network data. International Journal of Geographical Information Science, 19(4): 459-481.

Peters, W., Montiel-Ponsoda, E., Aguado de Cea, G. (2007) Localizing Ontologies in OWL. In Proceedings of the OntoLex07 Workshop, co-located at the ISWC 2007 Busan, South Corea.

Piasecki, M., Beran, B. (2005) Developing a domain ontology for the hydrologic community. In: Proceedings of the American Geophysical Union, 86(52), San Francisco,

Pinto, H.S., Tempich, C., Staab, S. (2004) DILIGENT: Towards a fine-grained methodology for DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies. In Ramón López de Mantaras and Lorenza Saitta, Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), pp. 393-397. IOS Press, Valencia, Spain. ISBN: 1-58603-452-9.

Poe, V., Klauer, P., Brost, S. (1998) Building a Data warehouse for Decision Support. Prentice Hall.

Poveda Villalón, M. (2009) Red de ontologías para el Camino de Santiago. Trabajo fin de carrera. Universidad Politécnica de Madrid.

Pressmann, R.S. (2000) Software Engineering. A Practitioner's Approach. 5'th Edition. McGraw-Hill, New York

Priyatna, F. (2009) RDF-based Access To Multiple Relational Data Sources. Master thesis. Universidad Politécnica de Madrid, Madrid.

Prud'hommeaux, E., Seaborne, A. (2008) SPARQL query language for RDF. W3C Recommendation.

Quak, W., Vries, M., Tijssen, T., Stoter, J., Oosterom, P. (2002) GML for exchanging topographic data. 5th AGILE Conference on Geographic Information Science, Palma (Balearic Islands, Spain).

Qin, H., Dou, D., LePendu, P. (2007) Discovering executable semantic mappings between ontologies. In: Proc. of the OTM Confederated Intl. Conf. CoopIS, DOA, ODBASE, GADA, and IS 2007, Vilamoura, Portugal, Springer, 832-849

Rahm, E., Do, H. H. (2000) Data Cleaning: Problems and Current Approaches. IEEE Data Eng. Bull., Vol. 23, No. 4, pp. 3-13.

Ram, S., Khatri, V., Hwang, Y., Yool, S. R. (2000) "Semantic modeling and decision support in hydrology," Photogrammetric Engineering and Remote Sensing, Vol. 66, No. 10, pp. 1229-1239.

Ram S., Khatri, V. Limin, Z., Dajun Zeng, D. (2001) GeoCosm: A semantics-based approach for information integration of geospatial data. Conceptual modeling for new information systems technologies. ER 2001 workshop. Lecture Notes in Computer Science. pp. 152-165. Yokohama , Japon. ISSN 0302-9743.

Ramos, J.A., Gómez-Pérez, A. (2008) Mappings for the Semantic Web. IEEE Intelligent Systems - Trends & Controversies. November-December 2008, pp. 79-81.

Raskin, R. G., Pan, M. J. (2005) Knowledge representation in the semantic web for Earth and environmental terminology (SWEET) Computers & Geosciences Vol.31, Issue 9, 1119-1125

Rawat, S., Roy, P.S., Lemmens, R., Saran, S. (2004) GML-based model of NSDE format for Geo-Spatial Data Interoperability. Global Spatial Data Infrastructure 7th Conference, Bangalore, India.

Ressler, J., Dean, M. (2007) Geospatial Ontology Trade Study. Proceedings of the Second International Ontology for the Intelligence Community Conference OIC-2007.

Rich, E., Knight, K. (1994) "Inteligencia Artificial". Editorial McGraw Hill. 2a. Edición.

Ristad, E. S., Yianilos, P. N. (1998) Learning string edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(5):522–532.

Rodríguez, A., Abad, P., Alonso, J.A., Sánchez, A. (2006) La Infraestructura de Datos Espaciales de España (IDEE): Un proyecto colectivo y globalizado. III Jornadas Técnicas de la IDEE. En Avances en las Infraestructuras de Datos Espaciales. Granell, C., Gould, M. (eds.). Col•lecció "Treballs d´Informàtica i Tecnología". Núm. 26. Universitat Jaume I. ISBN: 84-8021-590-9.

Rodríguez Pascual, A.F., García Asensio, L. (2005) 'A fully integrated information system to manage cartographic and geographic data at a 1:25,000 scale', XXII International Cartographic Conference (ICC2005). A Coruña, Spain. ISBN: 0-958-46093-0

Rodríguez Pascual, A.F., López Romero, E., Abad Power, P., Sánchez Maganto, A. (2006) 'Modelo de Nomenclátor de España. v1.2'. Disponible en: http://www.idee.es/resources/recomen dacionesCSG/MNEv1_2.pdf accedido 20081111

Roman, D., Klien, E. (2007) SWING - a semantic framework for geospatial services. In K. T. Arno Scharl, editor, The Geospatial Web. In the Advanced Information and Knowledge Processing Series, pages 227–237. Springer.

Roth, M.T., Schwarz, P. (1997) Don't Scrap It, Wrap It! A wrapper Architecture for Legacy Data Sources. Proceedings of the International Conference on Very Large DataBases, 266–275.

Roussey, C., Corcho, O., Vilches-Blázquez, L.M. (2009) A catalogue of OWL ontology antipatterns. Proceedings of the fifth international conference on Knowledge (KCAP'09) Pages 205-206. Redondo Beach, California, USA. ISBN:978-1-60558-658-8

Rusher, J. (2001) TripleStore. Semantic Web Advanced Development for Europe (SWAD-Europe), Workshop on Semantic Web Storage and Retrieval - Position Papers. http://www.w3.org/2001/sw/Europe/events/20031113-storage/positions/rusher.html

Sarawagi, S., Bhamidipaty, A. (2002) Interactive deduplication using active learning. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), pages 269-278.

Salton, G., McGill, M. J. (1986) Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA.

Salvadores, M., Correndo, G., Rodriguez-Castro, B., Gibbins, N., Darlington, J., Shadbolt, N. (2009) LinksB2N: Automatic Data Integration for the Semantic Web. In: International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2009).

Schade, S., Cox, S. (2010) "Linked Data in SDI or How GML Is Not about Trees," In Proceedings of the 13th AGILE International Conference on Geographic Information Science - Geospatial Thinking. Guimarães, Portugal.

Shekhar, S. (2004) Spatial Data Mining and Geo-spatial Interoperability. Report of the NCGIA Specialist Meeting on Spatial Webs, Santa Barbara. National Center for Geographic Information and Analysis, University of California.

Shekhar, S., Chawla, S. (2003) Spatial databases: A tour. Prentice Hall. ISBN 013-017480-7.

Shekthar, S., Chawla, S., Ravada, S., Fetterer, A., Liu, X., tien Lu, C. (1999) Spatial databases – accomplishments and research needs. IEEE Transactions on Knowledge and Data Engineering, 11(1).

Shepherd, I. D. H. (1991) Information Integration in GIS. In D. J. Maguire & M. F. Goodchild & D. W. Rhind (Eds.), Geographical Information Systems: Principles and applications. London, UK: Longman.

Sheth, A. P., (1998) Changing focus on interoperability in information systems: From system, syntax, structure to semantics. In Interoperating Geographic Information Systems, Goodchild, M. F., Egenhofer, M. J., Fegeas, R., Kottman, C. A. (eds), pp. 5–30, Kluwer.

Sheth, A., Kashyap, V. (1992). So for (schematically), yet so near (semantically). In. Proc. IFIP DS-5 Semantics of Interoperable Database Systems. Lorne, Australia: Chapman & Hall.

Sheth A. P., Larson, J.A. (1990) Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. ACM Computing Surveys, 22(3).

Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., van de Velde, W., Wielinga, B. (1999) Knowledge engineering and management. The CommonKADS Methodology. MIT press, Cambridge, Massachusetts.

Scheider, S., Devaraju, A., Janowicz, K., Maué, P., Schade, S., Keßler, C., Ortmann, J., Bishr, M., Fincke, T., Weigel, T., Kuhn, W. (2009) "Grounding Geographic Information," in The European Information Society – Taking Geoinformation Science One Step Further. Proceedings of the 12th AGILE. Hannover (Germany)

Schmidt, D. (1998). Distributed object computing with CORBA middleware. Disponible en: http://www.cs.wustl.edu/~schmidt/corba.html.

Schuster, G., Stuckenschmidt, H. (2001) Building shared ontologies for terminology integration. Paper presented at the KI-01 Workshop on Ontologies, Vienna, Austria.

Schwering, A. (2008) Approaches to semantic similarity measurement for geo-spatial data: A survey. Transactions in GIS 12(1), 5-29.

Schwering, A., Raubal, M. (2005) Measuring semantic similarity between geospatial conceptual regions. In: Proceedings of the 1st International Conference on GeoSpatial Semantics, Mexico City, Mexico, 90-106.

Seaborne, A. (2004) RDQL - a query language for RDF. W3C member submission, W3C. http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/.

Seaborne A. Prudhommeaux E. (2008) SPARQL query language for RDF. W3C Recommendation - 15 January 2008. http://www.w3.org/TR/rdf-sparql-query/

Sehgal, V., Getoor, L., Viechnicki, P.D. (2006) Entity resolution in geospatial data integration. Proceedings of the 14th annual ACM international symposium on advances in geographic information systems. 83-90. Arlington, Virginia, USA.

Sevilla Sánchez, C., Rodríguez Pascual, A.F., González Matesanz, F.J., Blanco Ortega, L.M., Vilches-Blázquez, L.M. (2006) 'Un SIG corporativo en el IGN para la gestión integrada, publicación y análisis de datos geográficos', Proceedings of XII Congreso Nacional de Tecnologías de la Información Geográfica. Camacho Olmedo, M.T.; Cañete Pérez, J.A.y Lara Valle, J.J. Ed. ISBN: 84-338-3944-6 Depósito Legal: GR-1855-2006

Singh, M. P., Cannata, P., Huhns, M. N., Jacobs, N., Ksiezyk, T., Ong, K., Sheth, A. P., Tomlinson, C., Woelk, D. (1997) "The carnot heterogeneous database project: Implemented applications," Distributed and Parallel Databases, vol. 5, no. 2, pp. 207–225.

Smith, B., Ceusters, W. (2010) Ontological realism: A methodology for coordinated evolution of scientific ontologies. Applied Ontology 5:139–188

Smith, B., Mark, D. M. (1998) Ontology and Geographic Kinds. In T. K. Poiker and N. Chrisman (eds.), Proc. 8th International Symposium on Spatial Data Handling (SDH'98), Vancouver: International Geographical Union, pp. 308-320.

Smits, P. (Ed.) (2002) INSPIRE Architecture and Standards Position Paper.Architecture and Standards Working Group. Document available at http://inspire.jrc.it/documents/inspire_ast_pp_v4_3_en.pdf

Spaccapietra, S., Parent, C. (1994) View integration: A step forward in solving structural conflicts. IEEE Transactions on Knowledge and Data Engineering, 6(2):258–274.

Souza, D., Salgado, A. C., Tedesco, P. (2006) Towards a context ontology for geospatial data integration. In: International Workshop on Semantic-based Geographical Information Systems, LNCS - Proc of SeBGIS. Heidelberg: Springer Verlag, v. 4278. p. 1576-1585. Montpellier, France.

Srinivasan, U., Ngu, A.H.H., Gedeon, T. (2000) Managing heterogeneous information systems through discovery and retrieval of generic concepts. Journal of the American Society for Information Science, 51(8), 707-723.

Staab, S., Schnurr, H.P., Studer, R., Sure, Y. (2001) Knowledge Processes and Ontologies. IEEE Intelligent Systems 16(1):26–34.

Steudler, D. (2006) Swiss cadastral core data model—experiences of the last 15 years. Computers, Environment and Urban Systems, 30: 600-613.

Stoter, J.E., Morales, J.M., Lemmens, R.L.G., Meijers, B.M., van Oosterom, P.J.M., Quak, C.W., Uitermark, H.T., van den Brink, L. (2008) A Data Model for Multi-Scale Topographic Data. In Headway in Spatial Data Handling. 13th SDH'08. Lecture Notes in Geoinformation and Cartography, Ruas, A., Gold, C. (Eds.) pp. 233-254. ISBN: 978-3-540-68565-4. Montpellier, France.

Strauch, J.C.M., de Souza, J.M., Mattoso, M. (1998) A methodology for GIS Databases integration. IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'98), IEEE CS Press, pp.151-159. Taipei, Taiwan.

Stuckenschmidt, H., Wache, H., Vögele T., Visser, U. (2000a) 'Enabling technologies for interoperability', in Workshop: Information Sharing: Methods and Applications at the 14th International Symposium of Computer Science for Environmental Protection, eds., Ubbo Visser and Hardy Pundt, volume 20, pp. 35–46, Bonn, Germany.

Studer, R., Benjamins, V.R., Fensel, D. (1998) "Knowledge Engineering: Principles and Methods". Data and Knowledge Engineering. 25, pp. 161-197.

Suárez-Figueroa, M.C. (2010) NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse. PhD Thesis. Universidad Politécnica de Madrid.

Suárez-Figueroa, M.C., Aguado de Cea, G., Buil, C., Caracciolo, C., Dzbor, M., Gómez-Pérez, A., Herrero, G., Lewen, H., Montiel-Ponsoda, E., Presutti, V. (2007) NeOn D5.3.1: NeOn Development Process and Ontology Life Cycle. NeOn project. http://www.neon-project.org.

Suárez-Figueroa, M.C., Dellschaft, K., Montiel-Ponsoda, E., Villazón-Terrazas, B., Yufei, Z., Aguado de Cea, G., García, A., Fernández-López, M., Gómez-Pérez, A., Espinoza, M., Sabou, M. (2008) NeOn D5.4.1: NeOn Methodology for Building Contextualized Ontology Networks. NeOn project. http://www.neon-project.org.

Subgrupo de Trabajo del Núcleo Español de Metadatos (2005) Núcleo Español de Metadatos (NEM v1.0). http://www.idee.es/resources/recomendacionesCSG/NEM.pdf

Sui, D.Z., Maggio, R.C., (1999) Integrating GIS with hydrological modeling: practices, problems, and prospects. Computers, Environment and Urban Systems. 23, 33–51.

Swedish Standard (2006) SS637008:2006 Geographic information – Surface Water Systems Conceptual and Application Schema. Disponible en: http://www.sis.se/stanli/

Tanasescu, V., Gugliotta, A., Domingue, J., Gutiérrez Villarías, L., Davies, R., Rowlatt, M., Richardson, M., Stincic, S. (2007) Geospatial Data Integration with Semantic Web Services: the eMerges Approach. In A. Scharl & K. Tochtermann (Eds.), The Geospatial Web. London: Springer.

Tejada, S.; Knoblock, C. A.; Minton, S. (2001) Learning object identification rules for information integration. Information Systems 26(8):607–633.

Ter Bekke, J.H. (1991) Semantic data modeling in relational environments. PhD. Thesis, University of Delft.

The European Parliament and Council of the EU (2007) INSPIRE Directive (Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community - INSPIRE).

The Object Management Group (OMG) (1992). The common object request broker: Architecture and specification. OMG Document 91.12.1. Revision 1.1.92.

Tolk, A., Diallo, S. Y., Turnitsa, C. D. (2007) "Applying the Levels of Conceptual Interoperability Model in Support of Integratability, Interoperability, and Composability for System-of-Systems Engineering," Journal of Systemics, Cybernetics and Informatics, Volume 5 Number 5, pp. 65-74.

Tomai, E., Kavouras, M. (2004) From "Onto-GeoNoesis" to "Onto-Genesis": The Design of Geographic Ontologies. Geoinformatica, 8(3), 285-301.

Tripathi, A., Babaie, H. A. (2008) Developing a modular hydrogeology ontology by extending the sweet upper-level ontologies. Computers & Geoscience, 34(9):1022-1033.

Truter, P. (2003) A Central Geospatial Data Warehouse to Support Decision Support Systems in Government and Mapping of Africa to Support NEPAD. In Proceedings of 21st International Cartographic Conference. Durban (South Africa).

Uitermark, H. T. (1996) The integration of geographic databases: Realising geodata interoperability through the hypermap metaphor and a mediator architecture, Proceedings of the Second Joint European Conference & Exhibition on Geographical Information (JEC-GI'96) (Vol. 1) : from research to application through cooperation: from research to application through cooperation, IOS Press, p.92-95, Barcelona, Spain.

Uitermark, H., van Oosterom, P. J. M., Mars, N. J. I., Molenaar, M. (1999a) Ontology-Based Geographic Data Set Integration, Proceedings of the International Workshop on Spatio-Temporal Database Management. Böhlen, M. H., Jensen, C. S., Scholl, M. O. (Eds.) (STDBM'99), LNCS 1678, p. 60-78, Springer-Verlag Berlin Heidelberg.

Uitermark, H.T., Van Oosterom, P.J.M., Mars, N.J.I., Molenaar, M. (2005) Ontology - based integration of topographic data sets. In: International Journal of Applied Earth Observation and Geoinformation;7(2): pp. 97-106.

Uitermark, H.T.J.A. (2001) Ontology-Based Geographic Data Set Integration. Ph.D. thesis, ISSN 1381-3617 / ISBN 90-365-1617-x, 155 pp. University of Twente. The Netherlands.

Uitermark, H., Vogels, A., van Oosterom, P. (1999b) Semantic and geometric aspects of integrating road networks. In Vckovski, A., Brassel, K. E., and Schek, H.-J., editors, Interoperating Geographic Information Systems, Second International Conference, INTEROP '99 (LNCS 1580), pages 177–188.

Ullman, J. D. (1997) Information integration using logical views. In Proc. of the 6th Int. Conf. on Database Theory (ICDT'97), volume 1186 of Lecture Notes in Computer Science, pages 19–40. Springer, London, UK.

UNECE (1998) The Aarhus Convention. The United Nations Economic Commission for Europe http://www.unece.org/env/pp/ (último acceso Enero 2001)

Uschold, M., Grüninger, M. (1996) "Ontologies: Principles, Methods and Applications". Knowledge Engineering Review 11(2):93–155

Valerhaugen, H. (2004) Working with Generic GML through Schema Analysis, Master Thesis, Ostfold Univ., Halden Norway.

Van Heijst, G. Schreiber, A.Th., Wielinga, B.J. (1997) Using explicit ontologies in KBS development. International Journal of Human-Computer Studies 45: p.183-292.

Vatsavai, R. R., Shekhar, S., Burk, T. E., Lime, S. (2006) UMN-MapServer: A High-Performance, Interoperable, and Open Source Web Mapping and Geo-Spatial Analysis System. Lecture Note on Computer Science, 4197: 400-417.

Vazquez Maure, F.; Martin Lopez, J. (1987) Vocabulario de términos geográficos. Madrid: Ministerio de Obras Públicas y Urbanismo. Instituto Geográfico Nacional. 89 p. ISBN: 84-505-4929-9.

Vckovski, A. (1998) Interoperable and Distributed Processing in GIS. Taylor & Francis, London. ISBN: 0-203-26991-8 http://www.amazon.com/gp/reader/0748407936/ref=sib_fs_top/002-8685054-7033639?ie=UTF8&p=S00M&checkSum=3PHfnixlZowmTCvkDs24Cz96HRpP9FEEqtkzkAjfeCQ%3D#reader-link

Vilches Blázquez, L.M., Bernabé Poveda, M.A., Suárez Figueroa, M.C., Gómez-Pérez, A., Rodríguez Pascual, A. F. (2007b) "Towntology & hydrOntology: Relationship between Urban and Hydrographic Features in the Geographic Information Domain". Ontologies for Urban Development Ed. Teller, J.; Roussey, C.; Lee, J. Springer-Verlag, 2007. ISBN: 978-3-540-71975-5.

Vilches Blázquez, L.M.; Cañete Pérez, J.A.; Corcho, O. y Bernabé Poveda, M.A. (2008) Interrelaciones entre las tecnologías de la información geográfica y la ingeniería ontológica para la mejora de la gestión de los recursos geoespaciales. Actas del XIII Congreso Nacional de Tecnologías de la Información Geográfica. Hernández Calvento, L.; Parreño Castellano, J.M. (eds.). Las Palmas de Gran Canaria. ISBN: 978-84-96971-53-0

Vilches-Blázquez, L. M., Ramos, J. A., López-Pellicer, F. J., Corcho, O., Nogueras-Iso, J. (2009a) An approach to comparing different ontologies in the context of hydrographical information. Fourth International Workshop "Information Fusion and Geographic Information Systems" (IF&GIS'09) Popovich et al., (eds.) Lecture Notes in Geoinformation and Cartography (LNG&C) Springer-Verlag. Pages: 193-207. ISBN: 978-3-642-00303-5 May 17th - 20th, 2009 St. Petersburg, Russia.

Vilches-Blázquez, L.M., Corcho, O. (2009b) A heuristic approach to generate good-quality linked data about hydrography. In 8th International Workshop on Web Semantics (WebS'09) in conjunction with 20th edition of DEXA. 31 August - 4 September. Linz, Austria.

Vilches-Blázquez, L.M., Martins, B., Wyttenbach, A., Bernabé, M.A., Álvarez, M., Luzio, J., Borbinha, J. (2007c) 'Geographical and historical thesauri: The state of the art', DIGMAP Deliverable D2.1. ECP-2005-CULT-038042, DIGMAP. Disponible en: http://www.digmap.eu/lib/exe/ fetch.php?id=wiki%3Aproject_documents%3Adeliverables& cache=cache&media=wiki:ec:digmap-d2.1_v1_2.pdf

Vilches Blázquez L.M., Rodríguez Pascual, A.F., Bernabé Poveda, M.A. (2006). "Ingeniería ontológica: El camino hacia la mejora del acceso a la información geográfica en el entorno web". III Jornadas Técnicas de la Infraestructura de Datos Espaciales de España. En Avances en las Infraestructuras de Datos Espaciales. Granell, Carlos & Gould, Michael (editores). Col•lecció "Treballs d´Informàtica i Tecnología". Núm. 26. Universitat Jaume I. pág 95 – 103. ISBN: 84-8021-590-9. Castellón de la Plana.

Vilches-Blázquez, L.M., Rodríguez Pascual, A. F., Mezcua Rodríguez, J., Bernabé Poveda, M.A., Corcho, O. (2007a) "An approach towards a harmonized framework for hydrographic features domain". In Proceedings of XXIII Int. Cartographic Conference. Moscow, Russia.

Vilches-Blázquez, L.M., Villazón-Terrazas, B., De Leon, A., Priyatna, F., Corcho, O (2010a) An Approach to Publish Spatial Data on the Web: The GeoLinked Data Use Case. Workshop on Linked SpatioTemporal Data 2010 in conjunction with the 6th International Conference on Geographic Information Science (GIScience 2010) Zurich (Switzerland).

Vilches-Blázquez, L.M., Villazón-Terrazas, B., Saquicela, V., De Leon, A., Corcho, O., Gómez-Pérez, A. (2010b) GeoLinked Data and INSPIRE through an Application Case. 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS 2010). San Jose, California (USA).

Visser, U. (2005) Intelligent Information Integration for the Semantic Web. Lecture Notes in Artificial Intelligence. Springer-Verlag New York, LLC. ISBN-13: 9783540229933, 172pp.

Visser, U., Stuckenschmidt, H. (2002) Interoperability in GIS - Enabling Technologies, In Proceedings of 5th AGILE Conference on Geographic Information Science, (Eds, Ruiz, M., Gould, M. and Ramon, J.) Palma de Mallorca, Spain, pp. 291-297

Visser, U., Stuckenschmidt, H., Schuster, G., Vögele, T., (2002) Ontologies for geographic information processing, Computers & Geosciences, Vol. 28, pp. 103-117.

Visser, U., Stuckenschmidt, H., Wache, H., Vögele, T. (2001) Using Environmental Information Efficiently: Sharing Data and Knowledge from Heterogeneous Sources. Environmental Information Systems in Industry and Public Administration. C. Rautenstrauch and S. Patig. Hershey, USA & London, UK, IDEA Group: 41-73.

Volz, J., Bizer, C., Gaedke, M., Kobilarov, G. (2009) Silk - A Link Discovery Framework for the Web of Data. In 18th International World Wide Web Conference 2009.

Volz, R., Handschuh, S., Staab, S., Stojanovic, L., Stojanovic, N. (2004) Unveiling the hidden bride: Deep annotation for mapping and migrating legacy data to the semantic web. Journal on Web Semantics, 1(2):187–206.

Volz, R., Oberle, D., Staab, D., Studer, R. (2003) Ontolift prototype. WonderWeb Deliverable D11. http://wonderweb.semanticweb.org.

W3C (2009) Publishing Open Government Data. W3C Working Draft. http://www.w3.org/TR/gov-data/

W3C (2001) XML Linking Language (XLink) Version 1.0. W3C Recommendation 27 June 2001.

Wache, H.; Scholz, Th.; Stieghahn, H.; König-Ries, B. (1999) An integration method for the specification of rule-oriented mediators. Proceedings of the International Symposium on Database Applications in Non-Traditional Environments, (DANTE'99), pp. 109 – 112. Kyoto, Japan.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G, Neumann, H., Hübner, S. (2001) Ontology-based integration of information – A survey of existing approaches. In IJCAI-01 Workshop: Ontologies and Information Sharing, pages 108-117.

Walter, V., Fritsch, D. (1999) Matching Spatial Data Sets: Statistical Approach, International Journal of Geographical Information Science 13(5), 445–473

Wang, Y. R., Madnick, S. E. (1989) The inter-database instance identification problem in integrating autonomous systems. In Proceedings of the Fifth IEEE International Conference on Data Engineering (ICDE 1989), pages 46-55.

Waterman, D. A. (1986) A Guide to Expert Systems. Addison-Wesley, Boston, Massachusetts.

Weis, M., Naumann, F. (2005) Dogmatix Tracks Down Duplicates in XML. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, MD, USA, ACM Press, 431-442

Weis, M., Naumann, F., Jehle, U., Lufter, J., Schuster, H. (2008) Industry-scale duplicate detection. In Proceedings of the VLDB Vol. 1, Issue 2, pp. 1253-1264. Auckland (New Zealand)

Widom, J. (1995) Research problems in data warehousing. Fourth International Conference on Information and Knowledge Management, pp.25-30.

Widom, J., (1996) Integrating Heterogeneous Databases: Lazy or Eager? ACM Computing Surveys. 28A(4).

Wiederhold, G. (1992) Mediators in the architecture of future information systems. IEEE Computer, Vol. 25, No. 3, pp 38-49.

Wiederhold, G. (1994). An algebra for ontology composition. In. Proceedings of 1994 Monterey Workshop on formal Methods (pp. 56-61), U.S. Naval Postgraduate School, Monterey CA, 1994.

Wilson, M. L. Schraefel, M. C. (2007) Bridging the gap: Using ir models for evaluating exploratory search interfaces. In SIGCHI 2007 Workshop on Exploratory Search and HCI. San Jose, CA, USA.

Winkler, W.E. (1999) The state of record linkage and current research problems. Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Winkler, W.E. (2006) Overview of Record Linkage and Current Research Directions. Statistical Research Division U.S. Census Bureau Washington, DC 20233

Worboys, M.F., Deen, S.M. (1991) Semantic Heterogeneity in Distributed Geographic Databases. SIGMOD Record, Vol. 20, Nº 4.

Wu, Z., Chen, H., Wang, H., Wang, Y., Mao, Y., Tang, J., Zhou, C. (2005) Dartgrid: a Semantic Web Toolkit for Integrating Heterogeneous Relational Databases, in Proc. of 4th International Semantic Web Conference (ISWC'05).

Xing, W., Corcho, O., Goble, C., Dikaiakos, M. (2007) Active Ontology: An Information Integration Approach for Highly Dynamic Information Sources. In European Semantic Web Conference 2007 (ESWC-2007), Innsbruck, Austria.

Xunta de Galicia (2004) 'Diccionario de datos de base topográfica 1:5.000 v_1.0. Noviembre 2004', Xunta de Galicia, Consellería de Política Territorial, Obras Públicas e Vivenda.

Yeung, A.K.W., Hall, G.B. (2007) Spatial Database Systems. Design, Implementation and Project Management. Series: GeoJournal Library, Vol. 87. Springer. ISBN: 978-1-4020-5391-7

Yu, C. Popa, L. (2004) Constraint-based xml query rewriting for data integration. In SIGMOD Conference, pages 371-382.

Yu, J., Benatallah, B., Casati, F., Daniel, F. (2008) Understanding Mashup Development. IEEE Internet Computing, Volume 12 (5), p. 44-52. ISSN:1089-7801.

Yu, J., Benatallah, B., Saint-Paul, R., Casati, F., Daniel, F., Matera, M. (2007) "A Framework for Rapid Integration of Presentation Components," Proc. Int'l World Wide Web Conf. (WWW 07), ACM Press, pp. 923–932.

Zaslavsky, I., Baru, C. (2003) Grid-enabled mediation services for geospatial information. In Proc of workshop on Next Generation Geospatial Information.

Zhang, J., Gruenwald, L. (2001) A GML-Based Open Architecture for Building a Geographical Information Search Engine Over the Internet, Proceedings of the 1st International Workshop on Web Geographic Information Systems (WGIS 2001) in conjunction with the 2nd International Conference on Web Information Systems Engineering (WISE 2001).

Ziegler, P., Dittrich, K. R. (2004) "Three Decades of Data Integration - All Problems Solved?". In 18th IFIP World Computer Congress (WCC 2004), volume 12, Building the Information Society, pages 3-12, Toulouse, France. Kluwer.

Ziegler, P., Dittrich, K. R. (2007) Data Integration — Problems, Approaches, and Perspectives. In John Krogstie, Andreas L. Opdahl, and Sjaak Brinkkemper, editors, Conceptual Modelling in Information Systems Engineering, pages 39–58. Springer, Berlin.

Anexo

Los tres problemas de investigación más importantes en bases de datos solían ser 'rendimiento', 'rendimiento' y 'rendimiento'; en los próximos años los tres problemas más importantes y que verdaderamente supondrán un desafío para la investigación serán 'la semántica', 'la semántica' y 'la semántica'.

Stefano Ceri (Borgida *et al.*, 2004)

Conforme a la Directiva Europea INSPIRE (The European Parliament *et al.*, 2007) "*... la información espacial... debe concebirse de forma que se garantice el almacenamiento, disponibilidad y mantenimiento de datos espaciales al nivel más adecuado; que sea posible combinar, de forma coherente, datos espaciales de diversas fuentes en toda la comunidad, y puedan ser compartidos entre distintos usuarios y aplicaciones; que sea posible que los datos espaciales recogidos a un determinado nivel ... sean compartidos con* otros usuarios*; que pueda darse difusión a los datos espaciales en condiciones que no restrinjan indebidamente su utilización generalizada; que sea posible localizar los datos espaciales disponibles, evaluar su adecuación para un determinado propósito y conocer las condiciones de uso*". Sin embargo, la existencia de múltiples productores hace común que en la práctica la información geoespacial sea capturada, gestionada y actualizada con diversos niveles de granularidad, calidad y estructura (Gómez-Pérez *et al.*, 2008). Además, con frecuencia, diferentes bases de datos describen, al menos parcialmente, el mismo espacio geográfico, aunque los datos son recogidos para distintos propósitos y, por tanto, en principio pueden ser diferentes de una fuente a otra (Devogele *et al.*, 1998). Esta situación de diversidad provoca la construcción de múltiples conjuntos de bases de datos espaciales caracterizados por diferentes niveles y tipologías de heterogeneidad, descritas en detalle en el Capítulo 2 (sección 2.1.1) de este trabajo.

Este anexo presenta el contexto de información considerado en este trabajo conforme a las diversas tareas propuestas en la Actividad 2 (Análisis de fuentes) de la metodología de integración descrita en el Capítulo 5. En la sección 1 se realiza una descripción de las bases de datos estudiadas, siendo clasificadas en diferentes niveles (europeo, nacional y local). El análisis de las bases de datos conforme a los aspectos mencionados en la metodología propuesta se presenta en la sección 2.

1. Contexto de información

En este trabajo se consideran diferentes bases de datos geográficas conformes con el modelo relacional, con diferente granularidad, escalas de captura (desde 1:10.000.000 a 1:5.000) e información. Dichas bases de datos proceden de diversas instituciones o productores de

información. Gran parte de las fuentes estudiadas poseen información de diferentes dominios, aunque este trabajo se centrará en la información hidrográfica relacionada con instancias de fenómenos geográficos españoles (componente común del conjunto de bases de datos). Además, producto de la riqueza de las distintas modalidades lingüísticas de España[1], aunque con algunas excepciones, estas fuentes poseen información multilingüe en los diferentes idiomas oficiales nacionales (castellano, gallego, catalán, vasco y aranés). A continuación, se realiza una descripción de alto nivel de las diferentes bases de datos agrupándolas en diversos niveles (europeas, nacionales y regionales). En la Figura A.1. se muestra una visión general de las bases de datos consideradas.

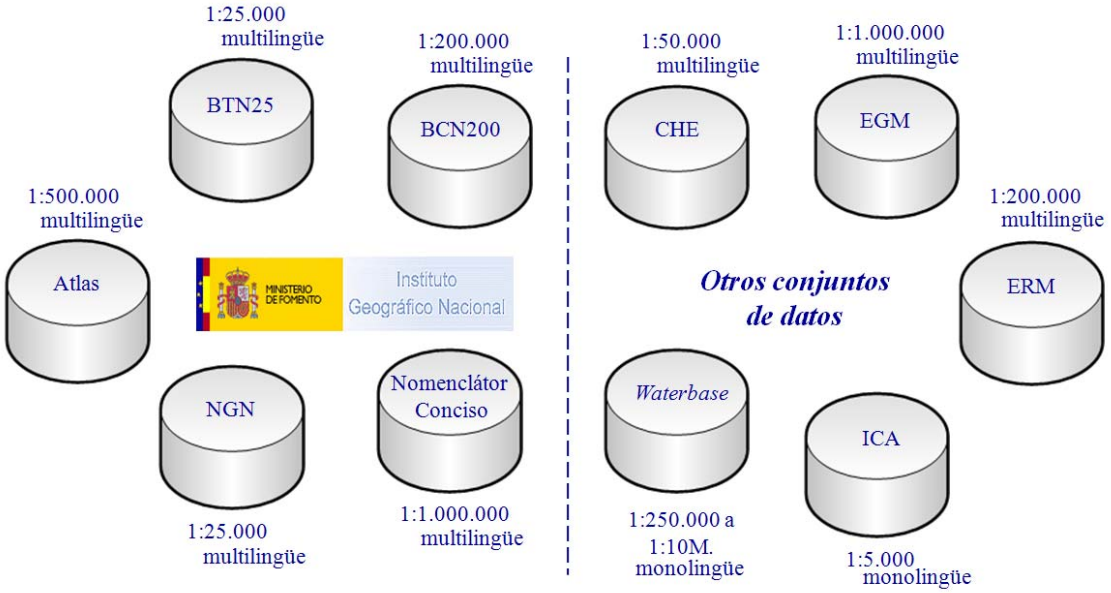


Figura A.1. Visión general del contexto de las bases de datos estudiadas

Bases de datos europeas. A nivel europeo, este trabajo considera tres bases de datos distintas: *Waterbase*, *EuroGlobalMap* y *EuroRegionalMap*.

- La base de datos ***Waterbase*** se creó dentro del proyecto *Water Information System for Europe* (WISE). Esta base de datos contiene datos previamente no disponibles e información recogida en la Unión Europea (EU) por diversas instituciones o confederaciones. WISE fue desarrollado por cuatro socios de la EU: *Environment Directorate-General*[2], *Eurostat* y *Joint Research Centre*[3] (JRC) (pertenecientes a la Comisión Europea) y la Agencia Europea de Medio Ambiente (EEA)[4]. Vinculado a este proyecto aparece *Waterbase*, nombre genérico dado a la base de datos de la Agencia Europea de Medio Ambiente con información sobre el estado y la calidad de los ríos, lagos, masas de agua subterráneas, aguas de transición, costeras y marinas. Asimismo, contiene información sobre la cantidad de los recursos de agua europeos. En referencia a algunos datos estadísticos, esta base de datos contiene más de 44.000 instancias referidas a fenómenos hidrográficos nacionales, por ejemplo, repartidas en más de 1.500 registros pertenecientes a lagunas, más de 3.500 ríos, etc. Además, esta base de datos se caracteriza por incluir únicamente información monolingüe.

[1] La Constitución Española en su artículo 3 afirma que "*la riqueza de las distintas modalidades lingüísticas de España es un patrimonio cultural que será objeto de especial respeto y protección*".
[2] *DG Environment*. http://ec.europa.eu/dgs/environment/index_en.htm
[3] *JRC*. http://ec.europa.eu/dgs/jrc/index.cfm
[4] *The European Environment Agency* (EEA). http://www.eea.europa.eu/

EuroGeographics representa a casi todos los Organismos Cartográficos y Catastrales Nacionales de Europa (*National Mapping and Cadastral Agencies* – NMCAs). Esta asociación supervisa, entre otros, los proyectos EuroGlobalMap y EuroRegionalMap. A continuación se lleva a cabo una descripción general de estos proyectos y de sus bases de datos asociadas.

- El proyecto ***EuroRegionalMap*** (ERM) establece la creación de una base de datos topográfica vectorial paneuropea a escala 1:250.000, que se usará como base geográfica de referencia para la presentación, visualización y análisis espacial. Esta base de datos está basada en la armonización de las bases de datos geográficas nacionales de las NMCAs ya existentes. Además, esta base de datos contiene información multilingüe y más de 36.400 hidrónimos referidos al ámbito nacional.
- ***EuroGlobalMap*** (EGM) es un conjunto de datos topográficos que cubre toda Europa a escala 1:1.000.000. Estos datos son producidos en cooperación con los Organismos Cartográficos Nacionales (National Mapping Agencies – NMAs) de Europa usando bases de datos oficiales. La información resultante de este proyecto proporciona la primera infraestructura de información geoespacial que estará mantenida en el nivel de la fuente por los respectivos Organismos Cartográficos Nacionales (NMAs) con objeto de proporcionar un acceso armonizado a dicha información. Asimismo, esta base de datos se caracteriza por contener más de 3.500 hidrónimos e información multilingüe.

Bases de datos nacionales. Con respecto a las fuentes de información que recogen datos sobre el conjunto nacional, este trabajo considera cinco bases de datos pertenecientes al Instituto Geográfico Nacional (IGN). De forma genérica, estas bases de datos (una Base Cartográfica Numérica, una Base Topográfica Numérica, dos fuentes vinculadas a respectivos nomenclátores y una correspondiente a un atlas) se caracterizan por recoger información múltiple a diferentes escalas y multilingüe.

- La Base Cartográfica Numérica a escala 1:200.000 (**BCN200**) se caracteriza por poseer más de 85.000 registros vinculados a tablas relacionadas con información hidrográfica. De este volumen de datos más de 59.000 topónimos distintos están asociados a diferentes fenómenos hidrográficos.
- La Base Topográfica Numérica a escala 1:25.000 (**BTN25**) es una base de datos orientada a SIG que presenta datos bidimensionales y continuos que cubren toda España. Posee más de 470.000 entradas en las diferentes tablas relacionadas con la información hidrográfica. De esta cifra, 192.000 registros corresponden a topónimos hidrográficos distintos.
- El **Nomenclátor NOMGEO** (NGN) alberga información desde 1:50.000 a escalas superiores. Tiene más de 490.000 entradas en su base de datos y más de 74.000 pertenecen a hidrónimos distintos con coordenadas geográficas, que pertenecen a 44 fenómenos diferentes con instancias multilingües.
- El **Nomenclátor Conciso** (NC) recoge información a escala $1:1_1$. Esta fuente constituye un corpus básico de topónimos estandarizados creado por la Comisión Española de Nombres Geográficos. Esta base de datos tiene una única tabla donde aparece una miscelánea de fenómenos geográficos. En la versión controlada de la mencionada base de datos el número de entradas asciende a 3.667 registros pertenecientes a 22 fenómenos. De este número total de registros 765 corresponden a hidrónimos no duplicados en la base de datos.
- La base de datos del **Atlas Nacional** contiene información a escala 1:500.000. Esta fuente proporciona una visión global sintetizada de la geografía física y humana de España y contribuye al conocimiento territorial mediante la integración de datos temáticos. La base de datos tiene más de 1.100 instancias hidrográficas, entre las que aparece información multilingüe.

Bases de datos regionales. Finalmente, en este trabajo se recogen diversas bases de datos con información de ámbito local. A continuación se describen brevemente algunas características de estas fuentes de información:

- La base de datos del **Instituto Cartográfico de Andalucía** (ICA) posee información a escala 1:5.000. Esta fuente contiene más de 17.000 topónimos (monolingües) relacionados con diferentes fenómenos hidrográficos.
- La base de datos de la **Confederación Hidrográfica del Ebro** (CHE) recoge información a escala 1:50.000. En la base de datos de este productor temático aparecen más de 1.400 instancias relacionadas con los fenómenos hidrográficos de la cuenca del río Ebro. Entre el número de instancias mencionado se recoge información de carácter multilingüe.

2. Análisis de fuentes

El objetivo de esta sección, conforme a la Actividad 2 (sección 5.5.2) de la metodología propuesta en este trabajo, es obtener una descripción y comparación de los diversos esquemas y datos. En estas tareas la información presente y la escala de las fuentes son factores clave. Este hecho se manifiesta, fundamentalmente, en el proceso de comparación. Este proceso se lleva a cabo bajo la condición previa de la existencia de solapamientos de información entre fuentes. Seleccionadas las fuentes que cumplen este criterio, el factor escala es el elemento principal del proceso de comparación. Así, en primer lugar, se compararán las fuentes que poseen idéntica escala. Finalizado este proceso, se procede, con el objetivo de comprobar y enriquecer las diferencias y semejanzas entre diferentes fuentes, a un proceso de comparación del tratamiento de la información geoespacial desde una visión transescalar.

La información obtenida en este proceso de análisis ayuda a comprender cada una de las fuentes de información y contribuye a identificar similitudes y diferencias que favorecen el proceso de integración propuesto en la metodología. A continuación, se profundiza en el análisis de las características de las diferentes fuentes mediante la descripción de diversos ejemplos.

Descripción de esquemas

En la tarea destinada a la descripción de esquemas, descrita en detalle en la sección 5.5.2 (tarea 2.1), se lleva a cabo el análisis de cuatro aspectos fundamentales (*análisis de las tablas, análisis de componentes, análisis de las relaciones y análisis del aspecto multilingüe*) del modelado de los esquemas de las bases de datos desde la óptica del proceso de integración. A continuación se recoge una descripción genérica en el ámbito de las bases de datos nacionales.

Con respecto a las consideraciones tecnológicas, estas bases de datos están gestionadas a través de diferentes gestores de bases de datos. Así, BTN25, BCN200 y Atlas Nacional son gestionadas con Oracle, mientras que las correspondientes al Nomenclátor Conciso y Nomenclátor NOMGEO están gestionadas por MySQL, aunque originalmente provenían en formato de Microsoft Access.

La base de datos de **BCN200** contiene 106 tablas en las que se recoge toda la información vinculada a su catálogo de fenómenos correspondiente, descrito en el Capítulo 6 de este trabajo. Sobre esta fuente se va a poner especial atención en las siguientes tablas ELEMENTO_BCN03 (49.373 entradas) y EJES_RIOS (36.423 entradas). Estas tablas se caracterizan por no poseer un nombre descriptivo respecto de la información contenida y por contener una miscelánea de fenómenos. La información de estas tablas está distribuida en campos idénticos. El resto de tablas con información relacionada con el dominio hidrográfico son descartadas por no poseer información relativa a coordenadas. A continuación se describen los campos más significativos de estas tablas.

- *ID_CODIGO*. Identifica la clase de entidad a la que pertenece en referencia al catálogo de fenómenos.
- *ETIQUETA*. Recoge el nombre del topónimo y, en ocasiones, también aparece un nombre alternativo.
- *N_ETIQUETAS*. Contiene el número de topónimos recogidos dentro del campo ETIQUETA.
- *LONGITUD*. Contiene la información relativa a la longitud del topónimo.
- *PERIMETRO*. En este campo se recoge idéntica información al campo LONGITUD.
- *GEOMETRY*. Este campo recopila la información relacionada con la geometría de los topónimos tratados.

La base de datos vinculada a **BTN25** presenta 411 tablas en las que se distribuye la información relacionada con el catálogo de fenómenos de BTN25, descrito en el Capítulo 6 de este trabajo. De este conjunto de tablas se pone especial atención en aquellas que contienen información hidrográfica, es decir, BCN0301L_RIO, BCN0304L_CAU_ART, BCN0307L_TUBERI, BCN0310L_RAMBLA, BCN0313L_RIA, BCN0316S_LAGUNA, BCN0319S_HUMEDA, BCN0334P_SURGEN, BCN0337P_CASCAD, BCN0355L_ISLA y BCN0358S_SALINA. De forma general, estas tablas se caracterizan por poseer nombres descriptivos conformes al contenido y por presentar, en algunos casos, contenidos heterogéneos. A continuación se describen, brevemente, algunos de los campos comunes de las tablas anteriormente mencionadas.

- *ID_ELEMENTO*. La información almacenada en este campo corresponde al identificador único del elemento dentro de la hoja.
- *ID_CODE*. Identifica la clase de entidad a la que pertenece en referencia al catálogo de fenómenos.
- *ETIQUETA*. Recoge el nombre del topónimo y, en ocasiones, también aparece un nombre alternativo. Resulta frecuente que las entradas de este campo aparezcan sin información.
- *N_ETIQUETAS*. Contiene el número de topónimos recogidos dentro del campo.
- *GEOMETRY*. Este campo recopila la información relacionada con la geometría de los topónimos tratados.

La base de datos del **Nomenclátor NOMGEO**, en su versión 24.07, contiene 20 tablas en las que se distribuye la información de esta base de datos. La tabla ENTIDAD recoge el conjunto de fenómenos vinculados al listado del catálogo de esta fuente descrito en el Capítulo 6. En esta tabla se recoge gran diversidad de información. En cuanto a los campos presentes en la misma, destacan:

- *Nombre*. Cada entrada de este campo contiene un topónimo.
- *Entidad*. Contiene el tipo de fenómeno geográfico vinculado con el topónimo.
- *Común*. Este campo ofrece un registro común a dos provincias e inscrito en ambas con el mismo nombre. Se caracteriza por ser un booleano.
- *XUTM*. Coordenada X en proyección UTM. Sistema geodésico ED-50 en la Península Ibérica e Illes Balears, ETRS89 en Ceuta y Melilla, y REGCAN95 en las Islas Canarias.
- *YUTM*. Coordenada Y en proyección UTM. Sistema geodésico ED-50 en la Península Ibérica e Illes Balears, ETRS89 en Ceuta y Melilla, y REGCAN95 en las Islas Canarias.
- *Huso*. Este campo contiene la zona en proyección UTM.

La base de datos del **Nomenclátor Conciso**, en su versión 1.0, posee una única tabla que recoge toda la información relacionada con esta fuente. Por tanto, resulta obvio afirmar que existe gran diversidad en su contenido. Entre la información recogida en esta base de datos merece destacarse la referida a los siguientes campos:

- *Nombre*. En este campo se recoge la instancia o nombre de los fenómenos geográficos en su variante monolingüe (por ejemplo, Turia, Río) y multilingüe (Alcoy/Alcoi) según el organismo competente o el documento indicado en el campo FUENTE NOMBRE.
- *Idioma Nombre*. Lengua en que está expresado el nombre de cada elemento geográfico según el código dado por la Norma ISO 639-1 y 2, Códigos para la Representación de los Nombres de Lenguas.
- *Variante*. En este campo aparecen nombres de topónimos alternativos o variantes en otros idiomas. La información presente en este campo es relativamente escasa.
- *Idioma Variante*. Aparecen en modo abreviado las diferentes lenguas de la denominación incluida en el campo VARIANTE, de acuerdo con lo establecido para el campo IDIOMA NOMBRE.
- *Grupo*. En este campo se delimita el dominio al que pertenece el topónimo.
- *Tipo Entidad*. Aquí se recoge la información correspondiente a la tipología del fenómeno asociada a cada topónimo de esta base de datos.
- *Latitud*. En este campo aparece la información de latitud representada en grados y minutos.
- *Longitud*. En este campo aparece la información de longitud reflejada en grados y minutos. Las coordenadas reflejan la localización de los topónimos mediante coordenadas geográficas, según el sistema de referencia ED50 (European Datum 1950). Se indican los grados y minutos, redondeando al minuto entero más próximo. No se incluye el hemisferio porque todas las latitudes son septentrionales. Las longitudes están referidas al meridiano de Greenwich, siendo negativas las que están situadas al Oeste y positivas al Este.

La base de datos del **Atlas Nacional** contiene 44 tablas diferentes en la que se distribuye la diferente información. De este conjunto de tablas se pone especial atención en aquellas que contienen información hidrográfica, es decir, ANE_AFLUENTES, ANE_EMBALSES, ANE_HUMEDALES, ANE_ISLAS, ANE_LONGITUD_LINEA_COSTA, ANE_RIOS, ANE_SUBAFLUENTES_1 y ANE_SUBAFLUENTES_2. Este conjunto de tablas no presenta miscelánea de fenómenos y presentan nombres identificativos respecto a su contenido. A continuación se describe, brevemente, algunos de los campos comunes de las tablas anteriormente mencionadas.

- *Nombre*. En este campo se recoge la instancia o topónimo de los diversos fenómenos hidrográficos recogidos.
- *Longitud*. En este campo se recoge el valor de la distancia asociado a una instancia determinada.
- *Superficie*. En este campo se recoge el valor asociado a la extensión de una instancia determinada.

Por otro lado, se puede afirmar que el conjunto de estas bases de datos descritas identifican claves primarias (*primary keys*) y extranjeras (*foreign keys*) en sus diferentes tablas. Asimismo, destaca la cardinalidad 1 a 1 entre las relaciones establecidas en las tablas y el hecho de que todas estas bases de datos consideran información multilingüe a nivel de instancia, aunque no se realiza un tratamiento especial de esta información.

Comparativa entre esquemas

En la tarea destinada a la comparación de esquemas, descrita en detalle en la sección 5.5.2 (tarea 2.2), se lleva a cabo dicha comparación considerando tablas, componentes especificados, relaciones y multiligualidad. A continuación se recoge, de forma breve, algunos aspectos del proceso de comparación entre bases de datos de escalas distintas (Nomenclátor NOMGEO y Confederación Hidrográfica del Ebro).

- *Análisis de las tablas*.
  - En lo que respecta al número de tablas, la base de datos de la CHE contiene 117, mientras que la perteneciente al NGN contiene 4 tablas.
  - El nombre de las tablas y su contenido resultan más descriptivos en la fuente del NGN.
  - El contenido de las tablas en ambas fuentes se caracteriza por una gran heterogeneidad, aunque la información presenta menos diversidad en las tablas de la CHE.

- *Análisis de los componentes especificados.*
  - El número de columnas varía entre las diferentes tablas de ambas fuentes, aunque aparecen nombres de campos comunes, por ejemplo, nombre o tipo de entidad.
  - Se identifican claves primarias en las diversas tablas que componen ambas bases de datos. Sin embargo, en la base de datos del NGN se lleva a cabo una distinción de claves extranjeras, en la fuente que concierne a la CHE no se reflejan este tipo de claves.

- *Análisis de las relaciones.*
  - Las relaciones establecidas en las diversas tablas de ambas fuentes de información se caracterizan por disponer de una cardinalidad 1 a 1.

- *Análisis del aspecto multilingüe*.
  - Ambas bases de datos consideran información multilingüe a nivel de instancia. Asimismo, la fuente correspondiente a la CHE contempla un campo para recoger nombres alternativos a los topónimos originales.

Descripción de datos

Esta tarea, descrita en detalle en la sección 5.5.2 (tarea 2.3), tiene por objetivo la descripción de los datos contenidos en las diferentes fuentes de información. Para alcanzar este objetivo se considera el análisis y descripción de ciertos parámetros cartográficos, características de los datos y la consideración del multilingüismo asociada a los datos de las diversas fuentes de información. A continuación se recoge la descripción de estos aspectos para BTN25.

- El listado de metadatos recopilado en la Tabla A.1 describe los parámetros cartográficos de BTN25.
- *Características de los datos*.
  - Originalmente, la información de BTN25 se encuentra en formato DGN[5]. Esta información tiene coordenadas ED50-UTM a escala 1:25.000 con resolución decimétrica. Con posterioridad, estos ficheros se transforman a coordenadas ED50-Geográficas y son cargados en una base de datos Oracle.
  - Las unidades de medida utilizadas en BTN25 son metros y se concretan en los siguientes campos:
    - campo COTA_0201 de la tabla BCN0201L_CUR_NIV
    - campo COTA_0204 de la tabla BCN0204P_PUN_ACO
    - campo ZMAX_0325 de la tabla BCN0325L_LIN_EMB
  - La información de BTN25 tan sólo controla que el cruce entre entidades se materialice en un nodo dentro de los nodos de ambas entidades, pero no tiene implementada topología arco-nodo.

[5] *DGN*. Es el nombre utilizado por los formatos de ficheros CAD (*Computer Aided Design*) de los software CAD *MicroStation* e *Intergraph's Interactive Graphics Design System* (IGDS).

- Se identifican omisiones en algunos registros (aparecen en blanco) y en algunos atributos, por ejemplo: Longitud, Superficie, Perímetro.

- La *consideración del multilingüismo*.
  - No se realiza ningún tratamiento de este tipo sobre los datos, aunque entre la información de esta fuente se pueden encontrar topónimos en castellano, catalán, euskera y gallego.

Tabla A.1**.** Listado de metadatos pertenecientes a la base de datos de BTN25

Elementos de metadatos	Información
Recurso (datos)	
Título	BTN25
Nombre de la organización (responsable)	Instituto Geográfico Nacional
Resumen descriptivo del conjunto de los datos	Base Topográfica Nacional a Escala 1:25.000
Localización geográfica del conjunto de los datos	España (Excepto las islas canarias)
Extensión geográfica (*Boundary Box*)	
Propósito	Producción cartográfica en entornos SIG
Fuente	Ficheros DGN de la BTN25
Escala	1:25.000
Descripción del nivel	
Tipo de representación espacial	
Fecha de creación (de los metadatos)	Enero 2010
Fecha (de referencia para el recurso mencionado)	2005
Revisión	Bianual
Formato de distribución	DUMP (BBDD completa), SHP (hoja)
Fecha de referencia para el conjunto de los datos	
Idioma del conjunto de los datos	Español
Categoría del tema del conjunto de los datos	
Sistema de referencia	Geográficas ED50
Identificador de sistema de referencia	EPSG:4230
Información del Datum vertical	
DQ_Completeness	
Calidad	No se recoge información sobre parámetros de calidad en esta fuente de información
Resolución	
Resolución espacial del conjunto de los datos	
Exactitud Posicional Externa Absoluta	
Descripción de la medida	
Resultado de la medida	
Calidad de los datos	
Exactitud posicional	
Consistencia topológica	
Identificador geográfico	
Información de agregación	
Exactitud de los atributos no cuantitativos	
Exactitud de los atributos cuantitativos	
Resultado cuantitativo	
Paso de proceso	
Información sobre la calidad de los datos	
Organización	
Autor	Instituto Geográfico Nacional
Conservador	Francisco Javier García García
Parte responsable del conjunto de los datos	

Comparativa entre datos y revisión de expertos

En estas tareas, descritas en detalle en la sección 5.5.2 (tarea 2.4 y 2.5), se considera un proceso de análisis y comparación de aspectos léxico-sintácticos, unidades de medida, ausencias generalizadas y/o lagunas puntuales de información (omisiones), exceso de información (comisiones) y presencia-tratamiento de información multilingüe. Asimismo, se realiza un

proceso de revisión de los datos basada en la retroalimentación de los expertos. Una visión detallada de los problemas detectados durante el proceso de comparación de los datos que conforman las diferentes fuentes de información se presenta en la siguiente sección.