Escuela Técnica Superior de Ingenieros Informáticos

Departamento de Inteligencia Artificial

TESIS DOCTORAL

**ENRIQUECIMIENTO DE FUENTES DE
DATOS CON DESCRIPCIONES
LINGÜÍSTICAS MEDIANTE BASES DE
DATOS GEOGRÁFICAS EXTERNAS**

Realizada por:

Javier Sánchez Soriano. Ingeniero en Informática

Dirigida por:

Martín Molina González. Doctor en Informática

Oscar Corcho García. Doctor en Informática

2016

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

Presidente
Vocal
Vocal
Vocal
Secretario
Suplente
Suplente

Realizado el acto de defensa y lectura de la Tesis el día de
de en la Escuela Técnica Superior de Ingenieros Informáticos.

Calificación:

EL PRESIDENTE

LOS VOCALES

EL SECRETARIO

A mi familia, que me
ha apoyado siempre.

Agradecimientos

Me gustaría agradecer en estas líneas la ayuda que muchas personas y amigos me han prestado durante el proceso de investigación y redacción de este trabajo. En primer lugar a mi antiguo compañero del grupo de investigación *Intelligent Systems & Knowledge Engineering* de la Universidad Politécnica de Madrid, el Dr. Victor Flores Fonseca (Universidad Católica del Norte) cuya tesis sembró la semilla de la mía. De la misma forma me gustaría agradecer y reconocer su ayuda al resto de mis compañeros de dicho grupo así como a los del grupo de investigación *Hidroinformática y Gestión del Agua*, también de la Universidad Politécnica de Madrid.

Así mismo, deseo expresar mi reconocimiento al Dr. David Atauri Mezquida y al Dr. Enrique Puertas Sanz, por su amistad, los sabios consejos así como por la ayuda que me dieron siendo compañeros en la Universidad Europea de Madrid. Igualmente, me gustaría mostrar mi afecto y el más profundo agradecimiento al resto de mis compañeros en la Universidad Europea de Madrid.

Me gustaría agradecer especialmente al Dr. Oscar Corcho (Universidad Politécnica de Madrid), su orientación y su inestimable ayuda, así como el fuerte impulso que proporcionó a los trabajos de la presente tesis cuando más lo necesitaba, como codirector de dicho trabajo.

Por ultimo, quisiera expresar mi sincera gratitud al Dr. Martín Molina González (Universidad Politécnica de Madrid), director de este trabajo. Por haberme animado a realizar estudios de doctorado y tras haber completado previamente y bajo su dirección tanto la tesis de master como el proyecto de final de carrera. Agradecerle el haberme orientado en todos los momentos que necesité sus consejos. Sin su amistad, ayuda y su interés, esta tesis difícilmente habría llegado a concluirse en los plazos y forma que hoy tiene.

Y concluyo: Muchas gracias a mi familia, por su paciencia y comprensión así como al resto de amigos y gente querida.

"En vez de intentar producir un programa que simule la mente adulta, ¿por qué no tratar de producir uno que simule la mente del niño? Si ésta se sometiera entonces a un curso educativo adecuado se obtendría el cerebro de adulto."

Alan Mathison Turing

Índice

I	INTRODUCCIÓN	1
1.	Introducción	3
1.1.	Planteamiento general y objetivos	4
1.2.	Organización de la memoria	8
II	ANTECEDENTES	11
2.	Fundamentos	13
2.1.	Información Geográfica	13
2.1.1.	Sistemas georreferenciados	14
2.1.2.	Sistemas de Información Geográfica	19
2.1.3.	Infraestructuras de Datos Espaciales	24
2.2.	Redes de Sensores	26
2.2.1.	SAIH: Sistema Automático de Información Hidrológica	29
2.2.2.	Sensor Web	36
2.2.3.	Ontologías sobre sensores	38
2.3.	Datos geográficos abiertos	41
2.3.1.	Características de las fuentes de datos en Internet	42
2.3.2.	Datos abiertos en Internet	44
3.	Estado del arte	49
3.1.	Curación de datos	49
3.2.	Generación de lenguaje natural	53
3.2.1.	Expresiones de referencia	53
3.2.2.	El caso de las expresiones de referencia geográficas	57
4.	Discusión	61
III	PROPUESTA	63

5. Método Propuesto	65
5.1. Hipótesis de partida	65
5.2. Planteamiento general del método	66
5.3. Metodología aplicada	69
5.4. Modelo de tareas y métodos	71
5.5. Modelo del Dominio	76
5.5.1. Información dinámica de entrada	76
5.5.2. Información intermedia y de salida	78
5.5.3. Información estática de entrada (bases de conocimiento)	80
5.6. Modelo de Inferencias	82
5.6.1. Inferencia Preprocesar-datos	83
5.6.2. Inferencia Componer-descripción	84
5.6.3. Inferencia Generar-hipótesis-de-descripción	85
5.6.4. Inferencia Seleccionar-fuente-de-datos	86
5.6.5. Inferencia Buscar-en-fuente-de-datos	87
5.7. Ejemplo de razonamiento	88
6. Evaluación	95
6.1. Experimento 1. Uso de una base de datos externa	96
6.1.1. Aplicación del método	97
6.1.2. Procedimiento de evaluación	102
6.1.3. Resultados de la evaluación	103
6.1.4. Conclusiones del experimento 1	104
6.2. Experimento 2. Uso de varias fuentes externas	105
6.2.1. Aplicación del método	105
6.2.2. Procedimiento de evaluación	111
6.2.3. Resultados de la evaluación	112
6.2.4. Conclusiones del experimento 2	116
6.3. Experimento 3. Referencias temporales	118
6.3.1. Aplicación del método	119
6.3.2. Conclusiones del experimento 3	121

IV	CONCLUSIONES	123
7.	Conclusiones	125
7.1.	Resumen	125
7.2.	Aportaciones	126
7.3.	Líneas futuras	127
	Referencias	129
V	ANEXOS	141
A.	Especificación CommonKADS	143
A.1.	Tarea <i>Enriquecer-datos-sobre-sensores</i>	144
A.2.	Tarea <i>Planificar-contenido-de-la-descripción</i>	145
A.3.	Método <i>Enriquecimiento-con-descripciones</i>	146
A.4.	Método <i>Planificar-contenido-con-Open-Data</i>	147
A.5.	Inferencia <i>Preprocesar-datos</i>	148
A.6.	Inferencia <i>Componer-descripción</i>	149
A.7.	Inferencia <i>Generar-hipótesis-de-descripción</i>	150
A.8.	Inferencia <i>Seleccionar-fuente-de-datos</i>	151
A.9.	Inferencia <i>Buscar-en-fuente-de-datos</i>	152
B.	Datos para evaluación	153

Índice de figuras

1.1. Presentación generada por la aplicación VSAIH.	5
1.2. Texto generado describiendo la situación hidrológica.	6
2.1. Coordenadas geográficas.	14
2.2. Meridianos y Paralelos.	16
2.3. Longitud y latitud de un determinado lugar de observación. . .	17
2.4. Husos y Zonas UTM.	19
2.5. Ejemplo de los datos utilizados en un SIG	20
2.6. Ejemplo de capas (Ráster y Vectorial) en un SIG.	21
2.7. Captura de pantalla del Geoportal Europeo <i>Inspire</i>	25
2.8. Tecnologías de medición y comunicación en meteorología. . . .	26
2.9. Ecosistema de redes de sensores <i>Libelium</i>	27
2.10. Placas <i>Beaglebone</i> , <i>Raspberry Pi</i> e <i>Intel Edison</i>	28
2.11. Aplicación Web acerca de datos hidrológicos del SAIH.	31
2.12. Aplicación Web del Ebro sobre datos hidrológicos del SAIH. .	33
2.13. Sensores y centro de control en el sistema SAIH.	34
2.14. Distribución de estaciones de volumen en la cuenca del Ebro. .	35
2.15. Visión de conjunto de SWE.	36
2.16. Visión general de los estándares de redes de sensores.	37
2.17. <i>Linked Data</i> Geográfico.	43
2.18. Los tres tipos de geometrías en OSM.	47
3.1. Elementos clave del DCC <i>Curation Lifecycle Model</i>	50
5.1. Generación de descriptores lingüísticos.	66
5.2. Procesos principales del método propuesto.	68
5.3. Arquitectura <i>CommonKADS</i>	69
5.4. Modelo de conocimiento <i>CommonKADS</i> en dominio médico. .	70
5.5. Descomposición en tareas del método.	71
5.6. Método <i>Enriquecimiento-con-descripciones</i>	73
5.7. Método <i>Planificar-contenido-con-Open-Data</i>	75
5.8. Componentes principales del modelo del dominio.	76
5.9. <i>Mapping</i> dominio inferencia <i>Preprocesar-datos</i>	83

5.10. <i>Mapping</i> dominio inferencia <i>Componer-descripción</i>	85
5.11. <i>Mapping</i> dominio inferencia <i>Generar-hipótesis-de-descripción</i>	86
5.12. <i>Mapping</i> dominio inferencia <i>Seleccionar-fuente-de-datos</i>	87
5.13. <i>Mapping</i> dominio inferencia <i>Buscar-en-fuente-de-datos</i>	88
5.14. Visión completa del método.	89
5.15. Ejemplo de generación de descripción.	94
6.1. Método aplicado al caso geográfico de la red SAIH.	105
6.2. Distribución de sensores SAIH por centro de control.	112
6.3. Resultados de evaluación por centro de control.	114
6.4. Principales causas de descripciones incorrectas.	116
6.5. Dualidad espacio-tiempo en referencias.	118

Índice de Algoritmos

5.1. Método Enriquecimiento-con-descripciones.	72
5.2. Método Planificar-contenido-con-Open-Data.	74
5.3. Inferencia <i>Componer-descripción</i>	84
A.1. Tarea <i>Enriquecer-datos-sobre-sensores</i>	144
A.2. Tarea <i>Planificar-contenido-de-la-descripción</i>	145
A.3. Método Enriquecimiento-con-descripciones.	146
A.4. Método Planificar-con-Open-Data.	147
A.5. Inferencia <i>Preprocesar-datos</i>	148
A.6. Inferencia <i>Componer-descripción</i>	149
A.7. Inferencia <i>Generar-hipótesis-de-descripción</i>	150
A.8. Inferencia Seleccionar-fuente-de-datos.	151
A.9. Inferencia Buscar-en-fuente-de-datos.	152

Índice de Tablas

2.1. Número de sensores de la red SAIH por área geográfica.. . . .	32
6.1. Patrones lingüísticos para construir descripciones de sensores.	99
6.2. Resultados evaluación calidad lugar geográfico.	103
6.3. Categorías de atributos de OSM.	110
6.4. Atributos temporales usados por [Molina and Stent, 2010]. . .	120

Resumen

El presente trabajo se ha centrado en la investigación de soluciones para automatizar la tarea del enriquecimiento de fuentes de datos sobre redes de sensores con descripciones lingüísticas, con el fin de facilitar la posterior generación de textos en lenguaje natural. El uso de descripciones en lenguaje natural facilita el acceso a los datos a una mayor diversidad de usuarios y, como consecuencia, permite aprovechar mejor las inversiones en redes de sensores.

En el trabajo se ha considerado el uso de bases de datos abiertas para abordar la necesidad de disponer de un gran volumen y diversidad de conocimiento geográfico. Se ha analizado también el enriquecimiento de datos dentro de enfoques metodológicos de curación de datos y métodos de generación de lenguaje natural. Como resultado del trabajo, se ha planteado un método general basado en una estrategia de generación y prueba que incluye una forma de representación y uso del conocimiento heurístico con varias etapas de razonamiento para la construcción de descripciones lingüísticas de enriquecimiento de datos.

En la evaluación de la propuesta general se han manejado tres escenarios, dos de ellos para generación de referencias geográficas sobre redes de sensores complejas de dimensión real y otro para la generación de referencias temporales. Los resultados de la evaluación han mostrado la validez práctica de la propuesta general exhibiendo mejoras de rendimiento respecto a otros enfoques. Además, el análisis de los resultados ha permitido identificar y cuantificar el impacto previsible de diversas líneas de mejora en bases de datos abiertas.

Abstract

This work has focused on the search for solutions to automate the task of enrichment sensor-network-based data sources with textual descriptions, so as to facilitate the generation of natural language texts. Using natural language descriptions facilitates data access to a wider range of users and, therefore, allows better leveraging investments in sensor networks.

In this work we have considered the use of open databases to address the need for a large volume and diversity of geographical knowledge. We have also analyzed data enrichment in methodological approaches and data curation methods of natural language generation. As a result, it has raised a general method based on a strategy of generating and testing that includes a representation using heuristic knowledge with several stages of reasoning for the construction of linguistic descriptions of data enrichment.

In assessing the overall proposal three scenarios have been addressed, two of them in the environmental domain with complex sensor networks and another real dimension in the time domain. The evaluation results have shown the validity and practicality of our proposal, showing performance improvements over other approaches. Furthermore, the analysis of the results has allowed identifying and quantifying the expected impact of various lines of improvement in open databases.

Parte I

INTRODUCCIÓN

1. Introducción

En general, los datos recogidos por redes de sensores para la gestión de sistemas dinámicos son potencialmente útiles no sólo para expertos en el dominio, sino también a amplias audiencias con una gran diversidad de usuarios.

Por ejemplo, las redes de sensores para la monitorización del agua en ambientes naturales pueden ayudar a una gran variedad de usuarios, tales como organismos de protección civil, ayuntamientos, consultores, compañías de seguros, investigadores científicos, periodistas, educadores y ciudadanos en general. Esta información puede ayudar a tomar decisiones de gestión (por ejemplo, acerca de la agricultura, la producción de energía hidroeléctrica, el riesgo de inundaciones, o el cambio climático) y también puede contribuir a aumentar la sensibilización del público en general sobre temas importantes (por ejemplo, la disponibilidad limitada de los recursos naturales como el agua para promover su mejor uso con prácticas sostenibles) [Molina et al., 2015].

Otro ejemplo de redes de sensores podrían ser los relativos a la medición de la calidad del aire. En la normativa vigente desde el 21 de mayo de 2008, relativa a la calidad del aire ambiente y a una atmósfera más limpia en Europa (Directiva 2008/50/CE del Parlamento Europeo) se recoge todo lo relativo a umbrales, acciones, etc. Además, se establece que los datos deben ofrecerse al público. La Directiva relativa a la calidad del aire ambiente y a una atmósfera más limpia en Europa, entre otras, incluye medidas destinadas a: *Asegurar que esa información sobre calidad del aire ambiente se halla a disposición de los ciudadanos*. En concordancia con ello se establece que la información se facilitará de forma gratuita por cualquier medio de comunicación de fácil acceso, incluido Internet, y tendrá en cuenta las disposiciones de la Directiva 2007/2/CE por la que se establece una infraestructura de información espacial en la Comunidad Europea (Portal *Inspire*, ver Figura 2.7).

Sin embargo, no siempre resulta fácil para las comunidades de usuarios el acceso y comprensión de los datos de las redes de sensores. Dichos conjuntos de datos normalmente se representan utilizando terminología técnica, así como jerga profesional (dispositivos sensores, códigos de identificación, magnitudes físicas, etc.), que puede resultar desconocida para muchos usuarios.

Por ejemplo, en España, el sistema SAIH (Sistema Automático de Información Hidrológica) es una red nacional de sensores hidrológicos que ha estado operando durante más de 20 años [DGA, 2009]. Esta red recoge datos hidrológicos en tiempo real (caudales de agua, niveles de agua, lluvia, etc.) grabados por unos 1.800 sensores en diferentes lugares de la geografía nacional. La información recogida por la red SAIH resulta útil para diferentes tareas, como pueden ser la alerta temprana de inundaciones o la gestión del agua. Los sensores en la red SAIH se identifican mediante un conjunto de convenios acordados por los operadores técnicos. Por ejemplo, el sensor con identificador *C001L85PQUIN* se refiere a un sensor que mide la precipitación acumulada en intervalos de una hora en un lugar determinado de la cuenca hidrográfica del Ebro. Este tipo de referencias puede ser utilizado con eficacia por los expertos de dominio. Sin embargo, para el público en general, estos identificadores y listas de códigos representan una barrera importante del idioma que impide que el personal no experto realice una fácil comprensión de los datos [Molina et al., 2015].

Además, el uso de estos conjuntos de datos requiere la familiaridad con las tecnologías informáticas (formatos de datos y las herramientas *software*) lo que también supone dificultades a los usuarios no expertos que estén interesados en los datos. Algunas aplicaciones *software* ayudan a los usuarios a interpretar estos conjuntos de datos mediante navegación gráfica y sistemas de visualización. Sin embargo, estas soluciones ofrecen un apoyo limitado para el público general porque dichas tecnologías normalmente asumen la familiaridad y el conocimiento del dominio para la interpretación de datos y visualizaciones gráficas y requiere que los usuarios tomen la iniciativa.

1.1. Planteamiento general y objetivos

Para facilitar el acceso a datos de sensores a grandes comunidades de usuarios, una posible solución es usar sistemas *data-to-text* [Hunter et al., 2008, Reiter and Dale, 2009, Reiter et al., 2005]. La idea básica es utilizar métodos computacionales que obtienen datos de las redes de sensores y automáticamente generan descripciones textuales en lenguaje natural que ayudan a los usuarios a comprender el significado de los datos con las interpretaciones y explicaciones pertinentes.

Un ejemplo de la solución *data-to-text* fue implementado para el sistema VSAIH (Vigilancia Hidrológica Automática - Sistema de Información SAIH) [Molina, 2012]. Este sistema analiza los datos de la red de sensores hidrológicos nacional SAIH y genera noticias en estilo periodístico para el público en general. El sistema VSAIH genera automáticamente noticias relacionadas principalmente con análisis descriptivos (¿Qué pasó?) así como análisis de diagnóstico (¿Por qué ocurrió?) sobre el riesgo de inundaciones, la disponibilidad de los recursos hídricos así como sobre los fallos del sensor (ver [Molina and Flores, 2012] para más detalles).

Figura 1.1: Presentación generada por la aplicación VSAIH.

La Figura 1.1 muestra un ejemplo de una presentación generada por el sistema VSAIH. La presentación incluye una descripción textual en estilo periodístico con su titular y el cuerpo junto con gráficos interactivos (mapas, gráficos, animaciones, etc.). La evaluación práctica del sistema VSAIH [Molina and Flores, 2012] mostró que este tipo de descripciones son capaces de facilitar la toma de decisiones y ahorrar tiempo crítico en situaciones especiales de emergencia.

Caudal por encima de lo normal en río Ebro en Ascó

En *río Ebro en Ascó* se registra un caudal de $362 \text{ m}^3/\text{s}$ lo que supone un incremento de $10.0 \text{ m}^3/\text{s}$ con respecto a la última hora. El caudal habitual en este punto del río es de $208 \text{ m}^3/\text{s}$.

En relación con lo anterior, se puede destacar el siguiente comportamiento hidrológico. Se han registrado variaciones de volumen en 3 embalses en río Ebro en las últimas 24 horas. La máxima disminución de volumen se ha producido en el embalse de Ribarroja con un decremento de 4.18 Hm^3 en las últimas 24 horas.

Figura 1.2: Texto generado describiendo la situación hidrológica.

En general, más allá del tipo de presentación construido por el sistema VSAIH, la generación de descripciones en lenguaje natural para redes de sensores también puede proporcionar más flexibilidad en combinación con convertidores *text-to-speech* (texto a voz) para ser utilizado en contextos de comunicación multimodal (por ejemplo, la comunicación de voz con redes de sensores, informar a las personas con discapacidad visual con mensajes de audio sobre los datos del sensor, etc.).

La generación de este tipo de descripciones textuales requiere que el problema específico de la generación de expresiones de referencia geográficas para ubicaciones geográficas se resuelva [Reiter and Dale, 1995] (por ejemplo, la generación de referencias tales como: "el río Ebro en Ascó" o "el río de Ribarroja"). En general, los métodos para generación de expresiones de referencia han evolucionado desde las primeras obras [Appelt, 1985, Appelt and Kronfeld, 1987, Kronfeld, 1990] a marcos más recientes, como la búsqueda gráfica [Krahmer et al., 2003] o restricción de la satisfacción [Garden, 2002].

Sin embargo, el problema de generar descripciones textuales para sensores geográficamente distribuidos es diferente de otros enfoques generales [Turner et al., 2009]. Uno de los principales problemas es que los conjuntos de datos de sensores contienen poca información geográfica explícita (generalmente sólo coordenadas geográficas) y el proceso natural de generación de lenguaje tiene que utilizar una gran cantidad de conocimiento geográfico para generar las descripciones. Por ejemplo, es necesario conocer los nombres de las diferentes entidades geográficas (lugares poblados, regiones administrativas, bosques, ríos, cuencas hidrográficas, montañas, lugares históricos, etc.) junto

con diversas relaciones espaciales (ubicaciones geográficas, formas espaciales, etc.) para generar descripciones útiles.

Expertos de dominio (por ejemplo, los expertos en la red de sensores SAIH en España) generan descripciones de texto utilizando referencias geográficas en función de su conocimiento del dominio de la zona espacial. La variedad de posibles referencias geográficas utilizadas por los expertos (por ejemplo, nombres de lugares poblados, formaciones naturales, monumentos históricos, etc.) muestra que esto no es un problema trivial. Por ejemplo, podemos pensar que es suficiente si seleccionamos el lugar poblado más cercano a la ubicación geográfica del sensor (dada por su latitud y longitud). Sin embargo, algunos sensores pueden estar situados en determinados lugares despoblados (bosques, montañas, parques, etc.), o en otro lugar poblado que puede ser más importante (por ejemplo, cuenta con mayor número de habitantes o es más conocido por razones históricas), o del lugar poblado no es lo suficientemente relevante en comparación con otras referencias (monumentos, construcciones, etc.). En esos casos, se utiliza otro nombre correspondiente a otro lugar poblado, tal como un monumento, una construcción, etc. [Molina et al., 2015].

En consecuencia, la generación de presentaciones en lenguaje natural es un problema difícil que requiere un gran volumen de conocimiento (geográfico y temporal). Para llevar a cabo esta tarea de forma automática, se pueden distinguir dos etapas:

1. En primer lugar, de forma anticipada, se pueden enriquecer los datos sobre redes de sensores con expresiones en lenguaje natural parciales (por ejemplo, sobre localización) mediante algoritmos sin restricciones de tiempo de respuesta que pueden hacer uso de un gran volumen de información.
2. Después, en tiempo real (durante la medición de datos), se pueden generar descripciones elaboradas con partes obtenidas previamente en la etapa anterior, con algoritmos eficientes con restricciones de tiempo de respuesta y que necesitan menor volumen de conocimiento.

El trabajo de investigación que se presenta en esta memoria tiene como fin dar respuesta a la necesidad planteada en una primera etapa sobre enriquecimiento de datos. Se ha considerado como hipótesis de trabajo el uso de bases

de datos abiertas en Internet para abordar la necesidad de disponer de un gran volumen y diversidad de conocimiento geográfico.

El trabajo de investigación analiza la viabilidad de esta solución teniendo en cuenta la actual disponibilidad de bases de datos abiertas que, en principio, pueden presentar ciertos grados de imperfección en los datos. Como metodología para hacer este trabajo, se han considerado los siguientes objetivos específicos:

1. Análisis del estado del arte respecto a las soluciones existentes potencialmente aplicables al problema considerado.
2. Diseño de un método general para enriquecimiento de datos que hace uso de bases de datos abiertas.
3. Evaluación del método propuesto mediante aplicación a redes sensores de dimensión real y uso de bases de datos abiertas representativas.
4. Análisis de resultados de la evaluación. Se distinguen dos tipos de conclusiones sobre dicho análisis:
 - Ventajas y limitaciones del método propuesto
 - Identificación de posibles mejoras en las bases de datos abiertas

1.2. Organización de la memoria

El presente trabajo se encuentra organizado en cinco partes. La primera parte contiene el capítulo 1, correspondiente a la Introducción. La segunda parte está dividida en tres capítulos: Fundamentos, Estado del arte y Discusión (capítulos 1, 2 y 3 respectivamente). La tercera parte es la de la propuesta y se compone de los capítulos Método propuesto y Evaluación (capítulos 4 y 5). En la cuarta parte se tiene el capítulo número 7, correspondiente a las conclusiones y finalmente la quinta parte corresponde a los Anexos.

El capítulo 2 describe el contexto del problema. Para ello se revisan los fundamentos en los que se apoya el presente trabajo: Información geográfica, redes de sensores y los datos abiertos.

En el capítulo 3 se realiza un estudio del estado del arte actual en las áreas de investigación de curación de datos y de generación del lenguaje natural (consideradas básicas para el desarrollo de esta tesis). En cada una de éstas se hace un estudio de las técnicas y propuestas (herramientas) *software* más representativas, relacionadas con el objetivo de este trabajo.

En el capítulo 4 se presenta una discusión que se focaliza en la relación de las técnicas y tendencias de estas áreas de investigación con la motivación de la tesis y se puntualiza sobre la adecuación de éstas al objetivo planteado.

En el capítulo 5 se detalla la propuesta de método para la generación de descripciones lingüísticas. Esto orientado a ofrecer una solución técnica para el enriquecimiento de fuentes de datos con descripciones lingüísticas mediante bases de datos geográficas externas, acorde a los objetivos planteados.

En el capítulo 6 se instancia la propuesta general en dos dominios. Uno de ellos real y complejo, para validar su aplicabilidad a sistemas de estas características y su utilidad práctica, tanto de representación de los elementos del dominio como los que posibilitan la generación de descripciones lingüísticas acorde a las exigencias descritas. El otro es un dominio simulado, el del tiempo, en el que también se valida la aplicabilidad del método propuesto. Además se incluye en este capítulo la descripción de los dominios de aplicación de la propuesta, así como las descripciones de las diferentes validaciones del prototipo de herramienta de generación de descripciones lingüísticas.

El capítulo 7 contiene el resumen de las contribuciones alcanzadas y se exponen las líneas principales de trabajo futuro a desarrollar. Finalmente la memoria contempla las reseñas bibliográficas y un conjunto de anexos.

Parte II

ANTECEDENTES

2. Fundamentos

En este capítulo se presentan los tres fundamentos necesarios para abordar el problema del enriquecimiento de fuentes de datos sobre redes de sensores con descripciones lingüísticas con el enfoque seguido en el presente trabajo respecto al uso de bases de datos abiertas. Se trata de los siguientes fundamentos:

1. Fundamentos de información geográfica
2. Fundamentos de redes de sensores
3. Fundamentos de datos abiertos

En las próximas secciones del capítulo se describen las bases y las características de cada uno de los fundamentos.

2.1. Información Geográfica

La información geográfica (o IG) constituye uno de los tres pilares fundamentales de este trabajo y al hablar de ella se refiere a datos espaciales georreferenciados, pudiendo ser estos de cualquier naturaleza. Se entiende como datos espaciales aquellos que permiten definir la localización y relación entre objetos en base a un sistema de referencia establecido. Los sistemas de referencia espacial pueden ser de dos tipos [de Lerma et al., 2008]:

- **Georreferenciados:** Se establecen sobre la superficie terrestre. Son los más habituales al tratarse de un dominio tangible, perceptible y muy utilizado como referencia. Este tipo de datos pueden tener una posición implícita, como por ejemplo la población de una sección censal, una referencia catastral, etc. o bien una posición explícita como en el caso de las coordenadas capturadas con GPS.
- **No georreferenciados:** Son sistemas que tienen valor físico y que pueden ser útiles en determinadas situaciones.

Los sistemas de información georreferenciados (abreviados como SIG en su acrónimo inglés, *Geographic Information Systems*) son los que tienen un mayor interés para este trabajo y serán aquellos que se utilicen en adelante.

Figura 2.1: Coordenadas geográficas¹.

2.1.1. Sistemas georreferenciados

Los sistemas de coordenadas permiten determinar unívocamente la posición de un punto haciendo uso de tuplas de números, denominadas coordenadas. El orden en que se escriben las coordenadas suele ser relevante identificándose por su posición en una tupla ordenada, del estilo $\langle x, y \rangle$ o $\langle x, y, z \rangle$.

El estudio de los sistemas de coordenadas es objeto de la geometría analítica y permite formular los problemas geométricos de forma numérica. Un ejemplo habitual es el uso de las coordenadas geográficas (ver Figura 2.1), en las cuales se asigna la tupla $\langle \text{longitud}, \text{latitud} \rangle$ para localizar los puntos. El uso de un sistema de coordenadas que permita describir puntos en el espacio en el dominio de la física recibe el nombre de sistema de referencia.

Existen diferentes sistemas de referencia, aunque no existe uno que sea el más

¹Imagen obtenida de <http://web.educastur.princast.es>.

adecuado de forma general. Cada uno presenta unas ventajas y unas desventajas dependiendo de las operaciones que se quiera realizar de acuerdo a sus normas. En cualquier caso, es posible cambiar de unos a otros mediante operaciones matemáticas. Algunos de los sistemas de referencia más utilizados son:

- *Sistema de coordenadas geográficas*
- *Sistema de coordenadas universal transversal de Mercator (UTM)*
- *Sistema de coordenadas cartesianas*
- *Sistema de coordenadas polares*
- *Sistema de coordenadas cilíndricas*
- *Sistema de coordenadas esféricas*
- *Sistema de coordenadas curvilíneas generales*
- *Sistema de coordenadas curvilíneas ortogonales*

Todos ellos permiten definir sistemas de coordenadas, resultando muy habitual el uso del *sistema de coordenadas geográficas* así como del *sistema de coordenadas universal transversal de Mercator (UTM)*. Se describen ambos sistemas de referencia a continuación:

Sistema de coordenadas geográficas

El sistema de coordenadas geográficas utiliza una superficie esférica para definir ubicaciones en la Tierra. Para ello se utilizan sus valores de latitud (que abarca la vertical, de Norte a Sur) y longitud (desplazamientos en horizontal, entre Este y Oeste). La longitud y la latitud son ángulos medidos desde el centro de la Tierra hasta un punto de la superficie de la Tierra. Los ángulos se suelen medir en grados.

- *Latitud*: Mide el ángulo entre cualquier punto y el ecuador sobre el meridiano que pasa por dicho punto. La distancia equivalente a un grado en dichos meridianos depende de la latitud, dado que a mayores valores para latitud se reducen los kilómetros por grado. Las líneas

Figura 2.2: Meridianos y Paralelos ².

de latitud paralelas al ecuador se llaman paralelos (pueden verse en la Figura 2.2). En el paralelo Ecuador, cuya circunferencia mide 40.076km, un grado correspondería aproximadamente con 111,32 km. Respecto a la latitud se puede añadir lo siguiente:

- La unidad para medir la latitud son grados sexagesimales
 - Puntos ubicados sobre el mismo paralelo tienen la misma latitud
 - Latitudes por encima de Ecuador se conocen como Norte (N)
 - Latitudes por debajo de Ecuador se conocen como Sur (S)
 - Valores comprendidos entre 0° y 90°
 - Valor para el ecuador es 0°
 - Polos Norte y Sur con latitudes 90° N y 90° S, respectivamente
- *Longitud:* Mide el ángulo a lo largo del Ecuador desde cualquier punto de la Tierra. El meridiano de *Greenwich* (el que pasa por Londres) es el establecido como longitud cero grados. Las líneas de longitud perpendiculares al ecuador se llaman meridianos (Figura 2.2). Un par de meridianos que abarcan la vuelta a la tierra forman una circunferencia

²Imagen obtenida de <http://www.aristasur.com/>.

Figura 2.3: Longitud y latitud de un determinado lugar de observación³.

de 40.007,161 km, con lo que un grado equivale a 111,131 km. Respecto a la longitud se puede añadir lo siguiente:

- La unidad para la longitud también son grados sexagesimales
- Puntos en el mismo meridiano tienen el mismo valor longitud
- El valor para el meridiano de *Greenwich* es 0°
- Longitudes a la derecha de *Greenwich* (según se mira a la superficie terrestre) se conocen como Este (E)
- Longitudes a izquierda de *Greenwich* se conocen como Oeste (O)
- Los valores están comprendidos entre 0° y 180°

Combinando estos dos ángulos, se puede expresar la posición de cualquier punto de la superficie de la Tierra. Ejemplos son *Las Rozas, Madrid, España*

³Imagen obtenida de <http://www.elcielodelmes.com>.

con latitud 40.25° Norte y longitud 3.42° Oeste. Es decir, un vector dibujado desde el centro de la Tierra al punto 40.25° N, 3.42° O pasará por *Las Rozas*.

Sistema de coordenadas universal transversal de Mercator

Para el sistema de coordenadas universal transversal de Mercator, la proyección UTM es del tipo cilíndrica donde el factor de escala en la dirección del paralelo y en la dirección del meridiano son iguales. Los meridianos se proyectan sobre el plano con una separación proporcional a la del modelo, así hay equidistancia entre ellos. Sin embargo los paralelos se van separando a medida que nos alejamos del Ecuador, por lo que al llegar al polo las deformaciones serán infinitas. No se representa todo el globo con este sistema si no que está limitado a la región comprendida entre los paralelos 84° y 80° S. La proyección es compuesta, dado que la esfera se representa en partes en lugar de entera. Para ello se realiza una división la Tierra en husos de 6° de longitud cada uno (ver Figura 2.4).

El gran valor de este sistema es que ningún punto está demasiado lejos del meridiano central de su zona, lo que implica que las distorsiones son pequeñas y como consecuencia se tiene una buena precisión en general. Esta ventaja tiene un coste asociado, que es la discontinuidad: Un punto en el límite de una zona se proyecta en coordenadas distintas propias de cada Huso. Para evitar esta discontinuidad, se extienden las zonas, para que el meridiano tangente sea el mismo. Esto permite mapas continuos casi compatibles con los estándares. Sin embargo, en los límites de esas zonas, las distorsiones son mayores que en las zonas estándar.

Las coordenadas UTM se establecen en base a los usos y las zonas, que se describen de la siguiente forma:

- *Husos UTM*. La Tierra está dividida en 60 husos, cada uno de ellos equivalente a 6° de longitud contenidos en la zona de proyección UTM (definida entre los paralelos 80° S y el 84° N). Cada huso está numerado entre el 1 y el 60, estando el primer huso limitado entre las longitudes 180° y 174° O y centrado en el meridiano 177° O, el segundo huso entre las longitudes 174° y 168° O centrado en el meridiano 171° O y así sucesivamente. Cada huso tiene asignado un meridiano central, que

Figura 2.4: Husos y Zonas UTM⁴.

es donde se sitúa el origen de coordenadas, junto con el ecuador. Los husos se numeran en orden ascendente hacia el este. Por ejemplo, la Península Ibérica está situada en los husos 29, 30 y 31 mientras que las islas Canarias están situada en los husos 27 y 28.

- *Bandas UTM.* La Tierra está dividida en 20 bandas, cada una de ellas equivalen a 8° de latitud. Cada banda se nombra con letras que van desde la *C* hasta la *X* excluyendo las letras *I* y *O* por su parecido con los números uno y cero y la letra *Ñ*. La zona *C* coincide con el intervalo de latitudes que va desde 80°S (o -80° latitud) hasta 72°S (o -72° latitud). Las bandas polares no están consideradas en este sistema de referencia. Si una banda tiene una letra igual o mayor que *N* pertenece al hemisferio norte, en otro caso queda en el hemisferio sur.

2.1.2. Sistemas de Información Geográfica

Los Sistemas de Información Geográfica (SIG) surgen como resultado de la aplicación de las Tecnologías de la Información (TI) a la gestión de la Información Geográfica (IG). Los SIG consisten en un conjunto de herramientas que integran y relacionan diversos componentes (usuarios, *hardware*, *software* y procesos) que permiten la organización, almacenamiento, manipulación, aná-

⁴Imagen obtenida de <https://commons.wikimedia.org/wiki/File:Utm-zones.jpg>.

Figura 2.5: Ejemplo de los datos utilizados en un SIG. ⁵

lisis y modelización de grandes cantidades de datos procedentes del mundo real que están vinculados a una referencia espacial, facilitando la incorporación de aspectos sociales-culturales, económicos y ambientales que conducen a la toma de decisiones de una manera más eficaz.

En los SIG se utilizan diferentes técnicas o procesos en función de lo que se quiera hacer, para las cuales existen diferentes herramientas o *suites* de *software*. Entre las diferentes técnicas o procesos se podrían destacar:

1. Creación y captura de los datos y cartografía automatizada
2. Representación y conversión de datos
3. Proyecciones, sistemas de coordenadas y reproyecciones
4. Análisis espacial y superposición de mapas
5. Geocodificación y geoestadística

⁵Imagen obtenida de U.S. Government Accountability Office.

Figura 2.6: Ejemplo de capas (Ráster y Vectorial) en un SIG ⁶.

Como se ha comentado, hay disponibles en el mercado varias aplicaciones *software* SIG. Pueden establecerse varias clasificaciones de las mismas en función de diferentes características:

1. Clasificación en función del coste

- a) **Comerciales.** Se caracteriza por ser de pago. Hay muchas soluciones disponibles con todo tipo de modelos de negocio asociados a las mismas.
- b) **Open Source.** Existe multitud de *software* libre (más de 300 proyectos, según [Steiniger and Bocher, 2009]) dado el interés y la gran actividad al respecto por parte de la comunidad.

2. Clasificación en función del tipo de aplicación

- a) **Clientes de escritorio.** Los clientes de escritorio ofrecen un amplio rango de aplicaciones, desde simples visualizadores a *software*

⁶Imagen obtenida de <http://sabia.tic.udc.es>.

de creación de mapas y análisis y tecnología punta en sistemas de edición y análisis profesional:

- ABACO DbMAP
- ArcGIS
- Autodesk Map
- qGIS
- Bentley Map
- GRASS
- etc.

b) **Bases de datos geográficas.** Son bases de datos optimizadas para almacenar y consultar datos que representan los objetos definidos en un espacio geográfico. La mayoría de las bases de datos espaciales permiten representar objetos geométricos simples, tales como puntos, líneas y polígonos. Algunas bases de datos espaciales manejan estructuras más complejas, tales como objetos en tres dimensiones, coberturas topológicas, redes lineales y TIN:

- PostGIS
- SpatiaLite
- Oracle Spatial
- MySQLSpatial
- etc.

c) **Servidores Web.** Capa intermedia de cualquier sistema de información, también conocida como middleware. Se trata de cualquier aplicación que se ejecuta en un servidor y que tiene como objetivo proporcionar uno o más servicios que serán consumidos por clientes, independientemente de si estos son otras aplicaciones de servidor, clientes web, clientes de escritorio o dispositivos móviles:

- GeoServer
- MapServer
- deegree
- QGIS Server
- etc.

d) **Clientes ligeros Web.** Software que se pueden utilizar desde dispositivos móviles del tipo de los smartphone y tabletas:

- OpenLayers
 - Mapbender
 - Leaflet
 - etc.
- e) **Catálogo de metadatos.** Estos servidores implementan estándares de descubrimiento de datos como CSW. Mediante estos servidores usuarios y otros componentes pueden encontrar juegos de datos y otros servicios mediante protocolos estandarizados. Estos servidores son el corazón de las Infraestructuras de Datos Espaciales:
- GeoNetwork
 - CatMDEdit
 - etc.
- f) **Bibliotecas geoespaciales.** Paquetes que ofrecen funcionalidades avanzadas para cualquier sistema de información en la forma de librerías. Algunos de estos proyectos ofrecen además de API de programación conjuntos de herramientas propias:
- GeoTools
 - Sextante
 - Shapely
 - GeoScript
 - JTS
 - etc.
- g) **Herramientas espaciales.** Herramientas ETL (*Extract, Transform and Load*: Extracción, Transformación y Carga) orientadas al uso de metadatos y con funcionalidades espaciales dedicadas a la integración de diversos orígenes de datos para la construcción y/o actualización de bases de datos espaciales y almacenes de datos espaciales:
- GeoKettle
 - FM
 - etc.

2.1.3. Infraestructuras de Datos Espaciales

Actualmente se dispone de un importante volumen de datos georreferenciados gracias a dos aspectos:

1. Los avances tecnológicos
2. La posibilidad de trabajar cómodamente con los datos georreferenciados mediante el uso de *software* e infraestructuras SIG

El objetivo de las Infraestructuras de Datos Espaciales (o IDE) es catalogar y acercar al público general esta información que, con frecuencia, es desconocida o no tiene canales adecuados para darse a conocer. Esta catalogación y acercamiento de la información implica:

- Consensuar estándares de metadatos
- Definir las referencias espaciales comunes
- Promover estándares de interoperabilidad para intercambios
- Proporcionar las herramientas y conocimientos para poner al alcance del público toda esa información espacial

La definición clásica de una IDE es básicamente tecnológica, ya que la presenta como una red descentralizada de servidores, que incluye datos y atributos geográficos; metadatos; métodos de búsqueda, visualización y valoración de los datos (catálogos y cartografía en red) y algún mecanismo para proporcionar acceso a los datos espaciales.

También podría definirse desde un punto de vista organizativo, que vendría a decir que el término IDE se utiliza para denotar el conjunto básico de tecnologías, políticas y acuerdos institucionales destinados a facilitar la disponibilidad y el acceso a la información espacial. En este sentido se entiende que el término infraestructura lo que quiere es enfatizar la existencia de un entorno solvente y sostenido que garantice el funcionamiento del sistema. Las IDE pueden ser implementadas en:

- Empresas
- Centros de investigación

Figura 2.7: Captura de pantalla del Geoportal Europeo *Inspire*⁷.

- Organismos oficiales
- Como ayuda para la gestión de su propia información espacial
- Como servicio público creado exprofeso y/o enlazando otras IDE

El objetivo es facilitar el acceso y la integración de la información espacial, tanto a nivel institucional y empresarial como de la propia ciudadanía, lo que permitirá:

1. Extender el conocimiento y el uso de la información geográfica
2. La optimización de la toma de decisiones
3. Promover los metadatos estandarizados como método para documentar la información espacial
4. La reducción de costos y evitar la duplicación de esfuerzos
5. Animar a la cooperación entre los agentes, favoreciendo un clima de confianza para el intercambio de datos

⁷Imagen obtenida de <http://inspire-geoportal.ec.europa.eu>.

Figura 2.8: Tecnologías de medición y comunicación en meteorología⁸.

2.2. Redes de Sensores

Para llevar a cabo trabajos de supervisión y actuación basados en el comportamiento de sistemas físicos (por ejemplo, cuencas hidrográficas, áreas meteorológicas, etc.) se cuenta con la recogida de información procedente de diversas fuentes y accesible para tomadores de decisiones [Herold et al., 2005]. En las últimas décadas, se han desarrollado sistemas de este tipo de acuerdo con sus necesidades particulares (ver por ejemplo [Aloysius, 2001]). Dichos sistemas cubren diversos aspectos tales como: planificación de respuestas, sistemas de alertas temprana, planificación de prevención a largo plazo, etc.

La construcción de dichos sistemas se beneficia de los recientes avances en tecnologías de información en diferentes campos. Por ejemplo el uso de sensores remotos basados en satélites y otros como redes de estaciones de medidas como pluviómetros, sensores de nivel, etc. es una de las formas más efectivas para la observación en tiempo real y de recogida de información.

⁸Imagen obtenida de World Meteorological Organization.

Figura 2.9: Ecosistema de redes de sensores *Libelium*⁹.

La cantidad y diversidad de sensores así como su empleo ha ido creciendo a lo largo del tiempo hasta dispararse en la actualidad. Existen multitud de sensores, capaces de medir muy diferentes aspectos y magnitudes. Además se han ido desarrollando en una gran variedad de ámbitos, dando lugar a redes de sensores de muy diferente tipología así como capacidades. Las posibilidades de comunicación han ido también cambiando a lo largo del tiempo, comenzando por redes cableadas, inalámbricas conectadas mediante radiofrecuencia, enlaces por satélite, *bluetooth*, *zigbee*, *wi-fi*, etc. Actualmente predominan aquellas conectadas de forma inalámbrica, en cualquiera de las muchas posibilidades como recoge [Garbarino and Echeverria, 2010].

La gestión y la comunicación de dichos sensores se gestiona en muchos casos mediante elementos de concentración (elementos repetidores de señal, pasarelas, etc.) y elementos de almacenamiento y gestión de las mediciones, dando lugar a lo que conocemos como redes de sensores.

⁹Imagen obtenida de <http://www.libelium.com>.

Figura 2.10: Placas *Beaglebone*, *Raspberry Pi* e *Intel Edison*.

La creación y uso de redes de sensores hoy en día se encuentra al alcance de todos los públicos, desde grandes compañías, pasando por todo tipo de colectivos hasta finalmente llegar a personas que a título individual implantan y utilizan dichas redes de. Este último caso es posible gracias, entre otras cosas, a la eclosión del IoT (*Internet of Things* o Internet de las Cosas) [Gubbi et al., 2013].

Mediante el empleo de microcontroladores programables de uso general, como los de la familia *Arduino*, las diferentes versiones de *Raspberry Pi*, *Beaglebone* o los *Intel Edison* y *Curie* (pueden verse en la Figura 2.10), así como de *frameworks* de desarrollo como *Intel XDK* y plataformas de almacenamiento y soporte tipo *FIWARE* resulta fácil y cómodo habilitar nodos para medir y monitorizar cualquier tipo de variable, contribuyendo a la antes mencionada expansión.

Algunos ejemplos del uso este tipo de microcontroladores son el sistema de [Sarango and Fabricio, 2014] para medir el consumo de agua potable con *Raspberry*, los sistemas de [Lourenço et al., 2015, Dubey et al., 2015] o el de [Kim and Myung, 2015], para medir el pulso de conductores, monitorizar de

salud y monitorizar microbiología del agua con placas *Intel Edison*. Otros ejemplos en el campo de la apicultura utilizando placas *Arduino* son los propuestos por [Mezquida and Martínez, 2009] y [Kviesis et al., 2015].

2.2.1. SAIH: Sistema Automático de Información Hidrológica

Como ejemplo representativo de red de sensores se puede citar el caso de la red SAIH. España es un país con grandes problemas de tipo hidrológico, debido a sus condiciones climatológicas, a la irregularidad de sus precipitaciones, tanto en el tiempo como en el espacio, y a su configuración orográfica, que producen situaciones contradictorias, a veces simultáneas: sequías asoladoras en unas zonas e inundaciones catastróficas en otras. Para paliar estos efectos, además de la infraestructura hidráulica, se han realizado una serie de actuaciones estructurales y no estructurales que se completan con varios sistemas de información y previsión. Uno de estos sistemas es el llamado Sistema Automático de Información Hidrológica (SAIH).

El Sistema Automático de Información Hidrológica (SAIH) [DGA, 2009] es un sistema de información en tiempo real, estructurado según las grandes cuencas hidrográficas peninsulares, y planteado para facilitar la toma de decisiones en la gestión de los recursos hídricos y la explotación de las infraestructuras hidráulicas en situación de avenida.

El SAIH es un potente instrumento de trabajo que permite conocer en tiempo real la situación hidro-meteorológica e hidrológica de las cuencas hidrográficas. Este conocimiento se consigue mediante la captación, transmisión, elaboración y presentación de forma útil para el usuario, de las variables hidrológico-hidráulicas en una serie de puntos representativos del funcionamiento. Para llevar a cabo estas funciones el SAIH capta en tiempo real los datos hidrológicos, los hidráulicos (situación de las infraestructuras) y otros datos meteorológicos básicos de cada cuenca, y los transmite al correspondiente centro de decisiones, donde son procesados y aplicados a la solución de los problemas propios de la gestión del agua de la cuenca, tanto en circunstancias normales (explotación), como en situaciones extraordinarias (avenidas).

Los objetivos iniciales del SAIH se orientan fundamentalmente a la ayuda a

la toma de decisiones en la gestión de los embalses en situación de avenida. Posteriormente este objetivo se extiende, convirtiendo al SAIH en un sistema multipropósito de gran utilidad para:

- La optimización de la gestión de los recursos hídricos
- Gestión de sequías
- Control de caudales ecológicos
- Verificación de caudales concesionales
- Control de aprovechamientos
- Control de caudales circulantes por canales de riego
- Planes de Emergencia y Protección Civil ante riesgo de inundación
- Suministrando información durante los episodios críticos

La mayor parte de las aplicaciones pueden agruparse de la forma siguiente:

- Gestión de recursos hídricos
- Control de los sistemas de explotación
- Verificación del cumplimiento de caudales concesionales
- Información de los caudales circulantes en los ríos
- Estadísticas hidrológicas
- Control de avenidas y seguridad de presas
- Captación de datos en tiempo real como niveles, caudales entrantes y salientes, cantidad de lluvia, etc.
- Comprobación del funcionamiento de los elementos de desagüe
- Información visual
- Ayuda a la toma de decisiones en avenidas
- Inundaciones

Figura 2.11: Aplicación Web del Ministerio español de Medio Ambiente acerca de datos hidrológicos del SAIH ¹⁰.

- Sistemas de alerta hidro-meteorológica
- Seguimiento en tiempo real de las variables hidro-meteorológicas
- Previsiones de niveles y caudales mediante modelación hidrológica

El Programa Nacional SAIH se inició en España a finales de los años ochenta. El objetivo de este programa fue la instalación de dispositivos sensores y redes de telecomunicaciones en las principales cuencas hidrológicas para obtener información hidrológica sobre el estado de los ríos en tiempo real en los centros de control.

El sistema SAIH incluye diferentes tipos de sensores tales como pluviómetros que registran información como la lluvia en algunos lugares, los sensores de nivel de agua, y sensores para la descarga en los embalses y de caudal en los canales fluviales. En total, entre los diferentes tipos de sensores instalados se

¹⁰Imagen obtenida de <http://sig.magrama.es/saih/>.

<i>Área</i>	<i>lluvia</i>	<i>caudal</i>	<i>volumen</i>	<i>Total</i>
<i>Ebro</i>	223	128	48	399 (22 %)
<i>Duero</i>	182	37	18	237 (13 %)
<i>Tajo</i>	153	40	39	232 (13 %)
<i>Guadalquivir</i>	168	7	45	220 (12 %)
<i>Júcar</i>	121	16	21	158 (9 %)
<i>Guadiana</i>	130	5	11	146 (8 %)
<i>Norte</i>	81	26	20	127 (7 %)
<i>Andaluza</i>	92	9	13	114 (6 %)
<i>Catalana</i>	55	32	10	97 (5 %)
<i>Segura</i>	52	16	13	81 (4 %)
TOTAL	1257 (70 %)	316 (17 %)	238 (13 %)	1811

Tabla 2.1: Número de sensores de la red SAIH por área geográfica..

alcanza la cifra de unos 1.800 sensores. Hay nueve centros de control SAIH en España, uno por cada cuenca principal (Ebro, Tajo, Júcar, Segura, etc.) Utilizando el sistema SAIH, la información se registra periódicamente y se envía a los centros de control (por ejemplo, cada hora, 30 minutos o 15 minutos).

Los centros de control procesan y almacenan los datos en bases de datos locales. Además, el Ministerio de Medio Ambiente de España coordina e integra los datos registrados en una base de datos global. Parte de esta información es accesible a través de aplicaciones Web (Figura 2.12 y 2.13).

La Tabla 2.1 contiene un resumen del número de sensores instalados en las distintas cuencas hidrográficas. Las características de cada uno de los tipos de sensores son los siguientes:

- **Pluviómetro (P):** Es un dispositivo encargado de medir la precipitación de lluvia en el punto de la cuenca en la que se encuentra. Las unidades son milímetros por hora. Estos son la mayoría de los sensores de SAIH y representan casi el 55 % del conjunto total de los sensores.
- **El sensor de caudal (Q):** Una estación de caudal es un dispositivo ubicado en el lecho del río para medir su caudal. Se mide en metros cúbicos por segundo.
- **La estación de nivel (N, C):** La estación de nivel es un dispositivo ubicado en un embalse o un río. Un sensor de nivel que mide el nivel

Figura 2.12: Aplicación Web de la Confederación Hidrográfica del Ebro sobre datos hidrológicos del SAIH¹¹.

de agua del río o la presa en la que se encuentra. Se mide en metros con respecto al nivel del mar.

- **Estación de Volumen (Y):** Una estación de volumen se encuentra en la presa de un embalse para medir el volumen de agua almacenada en él. Por razones prácticas, se considera como valor de sensor, pero en realidad se deduce a nivel local desde el nivel del embalse. Se mide en hectómetros cúbicos.

En muchos casos, sensores de diferentes tipos comparten ubicación y comunicaciones, así como otras funciones (agua y la calidad del aire, etc.). Los embalses suelen tener pluviómetros y sensor de nivel entre otros. Los lechos de los ríos por lo general tienen pluviómetros junto con sensores de caudal o de nivel. La infraestructura SAIH también incluye los dispositivos específicos de telecomunicaciones (sistemas emisores-receptores de radio, redes de fibra óptica, etc.) que establecen la comunicación entre los sensores y el centro de control de la cuenca.

¹¹Imagen obtenida de <http://www.saihebro.com>.

Las medidas de comportamiento en tiempo real de los elementos de una cuenca hidrográfica obtenida por sensores se reciben en el SAIH para ser analizadas y utilizadas en función al alcanzar los propósitos descritos anteriormente. Para la medición del estado de los ríos y embalses hay diferentes dispositivos distribuidos a lo largo de la cuenca, destinados a recoger los datos de los caudales, niveles y precipitaciones ocurridos en diferentes instantes de tiempo. Se dispone de datos recogidos por el sistema SAIH cada T minutos (por ejemplo, $T = 15$ min. ó $T=30$ min.).

Figura 2.13: Sensores y centro de control en el sistema SAIH.

Los tipos de sensores habituales son pluviómetros, sensores de nivel (cota) y estaciones de aforo (además de otros como para medir apertura de compuertas de embalses, espesor de nieve, etc.). Los sensores tienen localización geográfica que indican su posición con respecto a los ejes en el mapa.

A modo de ejemplo, en la Figura 2.14 se muestran las estaciones de volumen sobre la cuenca hidrográfica del Ebro.

Actualmente, gracias a los avances tecnológicos en el campo de la teledetección, los modelos hidrológicos empiezan a estar preparados para asimilar

Figura 2.14: Distribución de estaciones de volumen en la cuenca del Ebro ¹².

información distribuida de diferentes variables geofísicas derivadas de sensores activos o pasivos instalados en distintas plataformas (por ejemplo radares y satélites).

En general, se identifican como fuentes de información adicionales:

- Fuentes de datos derivados del Radar Meteorológico
- Fuentes de datos derivados de la Teledetección Satelital
- Pronósticos derivados de los Modelos Numéricos de Tiempo (NWP) o modelos meteorológicos. Estos modelos con sus predicciones cuantitativas de precipitación (QPF), nos aumentan la ventana de pronóstico de nuestros modelos hidrológicos entre 24 y 96 horas.

Todo ello proporciona al usuario el valor del parámetro deseado en el punto observado por el SAIH, sea este una medida obtenida por un sensor o un pronóstico adquirido mediante los modelos citados.

¹²Imagen obtenida de <http://www.saihebro.com>.

Figura 2.15: Visión de conjunto de SWE¹³.

2.2.2. Sensor Web

Como ya se ha dicho las redes de sensores son cada vez más comunes, motivo por el cual la necesidad de asistencia en la gestión y en la consulta a dichas redes requiere la utilización de normas o estándares que faciliten el razonamiento automático. Los trabajos del grupo SWE (*Sensor Web Enablement*), perteneciente al OGC (*Open Geospatial Consortium*) han proporcionado una arquitectura [Botts et al., 2008] así como una colección de estándares basados en servicios entre los que se encuentran los utilizados en el IDE planteado por [Molina and Bayarri, 2011]: *Catalog Service for the Web* (CSW) y *Web Map Service* (WMS). Entre dichos estándares se encuentran lenguajes para descripción de sensores, de capacidades, de medidas y de otros aspectos relevantes.

Estos estándares facilitan la catalogación de los sensores y la comprensión de los procesos que permiten realizar las mediciones y el intercambio de datos basado en XML u otras etiquetas estandarizadas. Sin embargo, estos trabajos no proporcionan interoperabilidad semántica ni una base para el razonamiento que pueda facilitar el desarrollo de aplicaciones avanzadas. SWE ofrece a los usuarios y desarrolladores:

¹³Imagen obtenida de <http://www.opengeospatial.org>.

Figura 2.16: Visión general de los estándares de redes de sensores ¹⁴.

- Interfaces para aplicaciones de *Sensor Web*
- Conectores para los estándar *IEEE 1451*, *TML*, *CAP*, *WS-N*, *ASAP*
- Participar en procesos para la elaboración de normas
- Ubicación de los sensores vinculada a estándares geospaciales
- Fusión de datos de sensores con otros datos espaciales
- Vinculado al *IEEE* y otras organizaciones de estándares

Los estándares adoptados así como los pendientes de aprobación por parte del OGC en el *framework* del SWE incluyen los siguientes:

- **Observaciones y Medidas** (*Observations & Measurements* - O&M). Modelos generales así como codificaciones XML.
- **Protocolo PUCK** (*PUCK Protocol Standard*). Define un protocolo para recuperar una descripción *SensorML*, el código del controlador del sensor así como otra información desde el propio dispositivo, lo que permite la instalación, configuración y operación automática del sensor.

¹⁴Imagen obtenida de <http://www.w3.org>.

- **Lenguaje de modelado de sensores** (*Sensor Model Language* - SensorML). Modelo estándar y esquema XML para describir los procesos dentro de los sistemas de procesamiento de sensores.
- **Servicio de Observación de Sensores** (*Sensor Observation Service* - SOS). Interfaz abierta para un servicio web que permita obtener observaciones descripciones de la plataforma de uno o más sensores.
- **Servicio de planificación del sensor** (*Sensor Planning Service* - SPS). Interfaz abierta para servicios web que permitan a los clientes 1) determinar la viabilidad de la recogida de datos de uno o más sensores o modelos y 2) presentar solicitudes de cobro.
- **Modelo de datos común SWE** (*SWE Common Data Model*). Modelos de datos de bajo nivel para el intercambio de datos de sensores relacionados mediante nodos pertenecientes al marco SWE del OGC.
- **Modelo de servicio SWE** (*SWE Service Model*). Define los tipos de datos para su uso común en los servicios *SWE* del OGC.

Otras normas SWE están en discusión o en diversas etapas de desarrollo.

2.2.3. Ontologías sobre sensores

Como indican [Sánchez Alonso et al., 2008] las ontologías (entre otras tecnologías semánticas) pueden:

1. Mejorar la interoperabilidad y la integración semántica
2. Facilitar el razonamiento
3. Facilitar la clasificación

Así como otros tipos de aseguramiento y automatización no abordados en los estándares OGC. Además, la semántica permitiría que la red, sus sensores y sus datos pudieran a través de especificaciones de alto nivel:

- Organizarse
- Instalarse

- Administrarse
- Consultarse
- Ser entendidos
- Ser controlados

Los sensores resultan diferentes a otras tecnologías como las SOA (arquitecturas orientadas a servicios - *Service Oriented Architecture*), debido a la naturaleza basada en eventos así como de las relaciones temporales y espaciales que son propias de los sensores y de las redes de sensores. Cuando debe tenerse en cuenta el razonamiento sobre sensores, las limitaciones físicas complejas (disponibilidad limitada de potencia, memoria limitada, la calidad de datos variables) o la conectividad, la representación mediante ontologías permite aplicar técnicas de inferencia más fácilmente. Ontologías útiles para trabajar con redes de sensores, recogidas por [Compton et al., 2012]:

- *OWL for Observations*. Representación OWL para el esquema *Información geográfica - Observación y medidas*. Tiene una dependencia mínima de otras ontologías, al incorporar soporte para tiempo, geometrías y medidas [Probst, 2006].
- *CSIRO Sensor Ontology*. Para describir y razonar sobre sensores, mediciones y modelos científicos. La ontología proporciona una descripción semántica de sensores para su uso en *workflows* [Compton et al., 2009b, Holger and Compton, 2009].
- *OntoSensor*. Ontología utilizada para construir una base de conocimiento de sensores [Goodwin and Russomanno, 2006, Goodwin et al., 2007, Russomanno et al., 2005b, Russomanno et al., 2005a].
- *Sensor Web for Autonomous Mission Operations (SWAMO) ontology*. Habilita interoperabilidad dinámica de productos y servicios *Sensor Web*. Describe agentes para la distribución de recursos en el sistema, la toma de decisiones distribuida y otras operaciones autónomas [Underbrink et al., 2008].
- *Sensor Data Ontology (SDO)*. Para buscar datos de sensores a través de redes de sensores distribuidas y heterogéneas [Eid et al., 2006, Eid et al., 2007].

- *MMI Device Ontology*. Ontología de dispositivos oceanográficos, incluidos sensores (miden cosas) y *samplers* (recogen cosas). La prioridad es caracterizar en términos generales los dispositivos (clasificarlos en grupos). Un uso de dichas caracterizaciones es ayudar a los usuarios (o aplicaciones web) descubrir sensores o datos de interés [Compton et al., 2009a].
- *SensorML Processes*. Para representar el modelo *SensorML* y servir como ontología de partida para *MMI Device Ontology* [Robin and Botts, 2006].
- *Coastal Environmental Sensor Networks (CESN) ontology*. Creada como parte del proyecto *CESN Semantic Data Reasoner* [Calder et al., 2010].
- *Wireless Sensor Networks Ontology (WISNO)*. Para proporcionar términos relacionados con sensores y actuadores [Hu et al., 2007].
- *Agent-based Middleware Approach for Mixed Mode Environments (A3ME) ontology*. Clasificación básica para la descripción y descubrimiento de dispositivos y sus capacidades en redes heterogéneas con recursos limitados [Herzog et al., 2008, Herzog and Buchmann, 2009].
- *Ontonym - Sensor*. Para describir los sensores y los datos que generan. Es un conjunto de ontologías superiores que representan conceptos básicos en computación ubicua como son la hora, lugar, gente, detección, procedencia, eventos, dispositivos o los recursos [Stevenson et al., 2009].

Se quiere destacar la ontología *Semantic Sensor Network Ontology* como un caso destacado, al encontrarse actualmente estandarizándose en el W3C/OSG *Working Group on Spatial Data on the Web*¹⁵. Dicha ontología describe los sensores, sus mediciones así como los conceptos relacionados [Compton et al., 2012]. No describe conceptos del dominio, tiempo o localizaciones entre otras, dado que se encuentran disponibles y pueden obtenerse de otras ontologías.

¹⁵La ontología se encuentra publicada en <http://purl.oclc.org/NET/ssnx/ssn> (último acceso 17/11/2015).

2.3. Datos geográficos abiertos

En los últimos años se ha sucedido una gran recogida de datos espaciales con el consiguiente aumento de información georreferenciada, que en muchos casos se encuentra disponible para su acceso y consulta. Esta explosión ha sido posible gracias a los avances en las tecnologías web, que han permitido a usuarios ajenos a estas labores (más propias de gobiernos y grandes organizaciones) participar de forma activa en la contribución y mejora de este tipo de bases de datos (BBDD) de información geográfica, promoviendo lo que se conoce como VGI (*Volunteer Geographic Information*) que a día de hoy se está convirtiendo en un paradigma, según [Jiang and Thill, 2015].

Estas BBDD tienen un gran interés, por la gran diversidad de colaboradores así como por su rápida expansión pero también presentan toda una serie de problemas y desafíos como son:

- La fragmentación de la información
- Problemas de coherencia
- Errores de etiquetado

A ello hay que unir el manejo de un ingente volumen de información que puede disponer de una capa que proporcione un acceso más inteligente a los contenidos geográficos. Esto lo proponen [Ballatore and Bertolotto, 2013], quienes hacen una revisión de las BBDD de conocimiento abierto centrándose en el geoespacial y más concretamente en *Open Street Map* (OSM), sobre el que elaboran la base de conocimiento *Open Street Map Semantic Network*, para una mejora de la integración así como de la recuperación de información. OSM ha supuesto toda una revolución en el mundo de los datos abiertos, con una excelente acogida por parte de la comunidad [Mashhadi et al., 2015].

Existe un especial interés por parte de las administraciones respecto al acercamiento del *Open Data* a los ciudadanos. A raíz de ello ha ido creciendo el marco normativo vigente en Europa y en España sobre *Reutilización de la Información del Sector Público* (RISP) [Clabo and Ramos-Vielba, 2015, Schade et al., 2015], referido por la Directiva 2003/98 de 17 del noviembre, la Ley 37/2007 del 16 de noviembre y el Real Decreto 1495/2011 del 8 de noviembre

así como posteriores modificaciones y ampliaciones.

Recientemente se ha publicado en el BOE la Ley 18/2015 por la que se modifica la Ley 37/2007 sobre reutilización de la información del sector público. El nuevo texto persigue facilitar la creación de productos y servicios de información basados en datos del sector público, garantizar la eficacia en el uso transfronterizo de documentos del sector público por empresas privadas y ciudadanos y promover la libre circulación de información y la comunicación, garantizando el respeto a la seguridad jurídica, la protección de los datos personales, así como la propiedad intelectual e industrial.

2.3.1. Características de las fuentes de datos en Internet

Las características de las fuentes de datos en Internet podrían resumirse en las siguientes:

- **Múltiples fuentes de datos.** La diversidad de datos así como sus características son prácticamente ilimitadas. Se tienen datos de todas las áreas, desde aquellos de carácter hidro-meteorológico, datos de desplazamientos de vehículos o todo tipo de información de las administraciones. El acceso masivo a Internet hoy en día y su acercamiento a todo tipo de dispositivos en la tendencia conocida como IoT (*Internet of Things*) ha permitido que multitud de dispositivos generen información, entre la que se encuentra la de los sensores de los teléfonos y relojes inteligentes (fotografías, coordenadas, ritmo cardiaco, temperatura,...), los coches conectados, o microcontroladores programables, entre otros.
- **Muchas de dichas fuentes son de acceso abierto.** La tendencia es que las colecciones de datos comentadas en el punto anterior se hagan públicas. Existe multitud de plataformas de carácter específico así como otras generales para dicho fin. Los datos de la administración son un ejemplo, dadas las exigencias de los ciudadanos a poder acceder a datos generados con dinero público se ha convertido en tendencia que se pongan a disponibilidad de los usuarios. El acceso abierto viene regulado por unas licencias de uso, con diferentes limitaciones y propiedades pero en cualquier caso con la idea de crecer y ser reutilizados por el bien de la comunidad.

- **Gran parte tienen referencias geográficas.** La generación de este tipo de información ha crecido de la misma forma que se ha extendido el uso de receptores de posición por satélite tipo GPS, hoy en día presentes en multitud de dispositivos de medida. En la Figura 2.17 puede verse un mapa de *Linked Data* geográfico.

Figura 2.17: *Linked Data* Geográfico ¹⁶.

- **Tendencia creciente a los aportes de voluntarios a dichas fuentes.** Usuarios de a pie, de forma voluntaria, contribuyen al enriquecimiento de fuentes de datos, en muchos casos de acceso abierto (*Open Data*). Esto ha supuesto un cambio de tendencia, dado que no son sólo las grandes empresas o la administración los que generan la información sino que también son individuos a título personal quienes contribuyen con su trabajo al poblado de las fuentes de datos, generalmente en aquellas de acceso abierto y en muchos casos en las de carácter geográfico. Resulta interesante combinar dicha información VGI y la convencional, como identifican [Upton et al., 2015] en la gestión forestal.

¹⁶Imagen obtenida de <http://linkeddata.org>.

- **Fuentes de datos accesibles mediante conectores de BBDD, URIs, API REST, etc.** Existe una gran variedad de formas de acceso a las fuentes de datos, entre las que se encuentran BBDD, URI's (*Uniform Resource Identifier* o identificador de recursos uniforme, API REST (Interfaz de programación de aplicaciones para Transferencia de Estado Representacional, o *Application Programming Interface for Representational State Transfer*). El caso de las API REST se está convirtiendo en el más usual, dada la facilidad para integrarlo y escalarlo en aplicaciones de tipo *mashup* (aplicaciones Web que combinan información de varias fuentes). De acuerdo con esta tendencia, [Rettig et al., 2015] presentan un API REST *Open Source* para redes de sensores medioambientales. En la misma línea trabajan [Yu and Liu, 2015], en la gestión de recursos hídricos.

2.3.2. Datos abiertos en Internet

En los últimos años se han hecho muchos esfuerzos para la publicación de conjuntos de datos a gran escala en Internet bajo licencias abiertas. Ejemplos de tales *datasets* son *Open Street Map*, *Geonames*, *DBpedia*, y muchos otros dentro y fuera del ámbito geográfico (por ejemplo, las bases de datos disponibles en la nube LOD - *Linked Open Data*). Algunos autores [Roth and Frank, 2009, Schuldes et al., 2011] han utilizado con éxito este tipo de *datasets* en problemas geográficos (direcciones de ruta).

Un buen número de estos *datasets* se han publicado de acuerdo a los principios del *Linked Data* [Bizer et al., 2009] y pueden o no estar registrados en la mencionada nube LOD. Dichos principios pueden resumirse en

1. Permitir subir los datos a la Web
2. Definir URI para los datos
3. Hacer esas URI referenciables

La cantidad de fuentes de datos geográficos LOD compatibles es bastante significativa, con enfoques internacionales como es el caso de *Geonames* o bien enfoques nacionales y regionales, como la iniciativa *GeoLinked Data* Española [Vilches-Blázquez et al., 2010]. Este movimiento del *Open Data* también

está siendo adoptado por muchas ciudades, que están incorporando sus *datasets* en portales (propios o de terceros) dedicados al *OpenData* y también están publicando APIs, que ofrecen el acceso a todo este tipo de información geográfica.

Algunas de las necesidades de las aplicaciones de datos respecto a la información geográfica pueden ser consultar:

- Nombres de las diferentes entidades geográficas: lugares poblados, regiones administrativas, bosques, ríos, cuencas, montañas, lugares históricos, etc.
- Relaciones espaciales: coordenadas geográficas, formas espaciales, etc.
- Datos relacionados con la relevancia de localizaciones espaciales: número de habitantes, número de fotografías para evaluar la popularidad de ciertos puntos de referencia, etc.

Algunos ejemplos de fuentes de datos geográficos en Internet pueden ser: *Geonames*, *Open Street Map*, *DBpedia* o *Panoramio*

Geonames

Se trata de un servidor de nombres geográficos cuyo funcionamiento es relativamente simple. Mediante el envío de unas coordenadas geográficas se solicita que proporcione un nombre adecuado para la ubicación especificada. El nombre se selecciona en función de diferentes características (lugar poblado, carreteras, accidentes montañosos, hidrografía, etc.). Tiene su propia ontología, para organizar el conocimiento [Vatant and Wick, 2006].

- Incluye más de 7.5 millones de ubicaciones
- Contiene nombres de ubicaciones en muchos idiomas
- Las coordenadas manejadas se expresan en el formato WGS84 estándar
- Bajo licencia *Creative Commons* (*Atribución 3.0 Unported CC*)

Por ejemplo, para encontrar el lugar poblado más cercano a una ubicación se utiliza la etiqueta ADMINISTRATIVE. En este caso, *Geonames* devuelve

una lista de los lugares más poblados ordenados por distancia. Para las zonas montañosas se utiliza la etiqueta HYPISOGRAPHIC.

Las consultas a *Geonames* se pueden realizar mediante consultas a los *Web Services* ofrecidos por la plataforma (también se pueden bajar unos ficheros XML con todas las características de un país). Para la utilización de los *Web Services* se requiere el uso de unas bibliotecas (se encuentran disponibles para varios lenguajes de programación) que permiten acceder a la funcionalidad.

En cualquier caso es un requisito imprescindible estar registrado en el sistema para controlar las cuotas o número de consultas realizadas. Para ello se utiliza un sistema de créditos, que van asociados a cada tipo de consulta mediante un coste así como con la cuenta del usuario, que puede ser gratuita (nº de créditos diarios reducido) pasando por varios tipos de cuentas con diferentes cantidades de créditos en función del coste.

Open Street Map (OSM)

Es un servidor de datos geográficos que permite el acceso libre a mapas y a varios conjuntos de datos. Estos datos se pueden descargar por completo o se puede acceder mediante una API REST. *Open Street Map* ha sido creado por miles de voluntarios [Goodchild, 2007, Haklay et al., 2010], como la *Wikipedia* (en el contexto geográfico, conocido como VGI). Tiene una licencia de uso del tipo *Open Database (ODbL)*.

Como resultado de las consultas a *Open Street Map* se obtiene un conjunto de etiquetas dentro de una región de búsqueda (*bounding box*), definida por cuatro parámetros: longitud mínima x_1 , latitud mínima y_1 , longitud máxima x_2 , latitud máxima y_2 . Las etiquetas se devuelven en un archivo XML con la información geográfica almacenada en los servidores de OSM para la región solicitada. Esta información geográfica consiste en nodos (*node*), formas (*way*) y relaciones (*relation*):

- **Node:** Punto singular definido por sus coordenadas geográficas. Puede contener cualquier número de atributos. Ejemplo de un nodo es la definición de un pico en una formación de montañosa.
- **Way:** Conjunto de nodos, que puede ser una polilínea o un polígono

Figura 2.18: Los tres tipos de geometrías en OSM.

(polilínea cuyos primer y el último nodos tienen la misma posición). Al igual que los *node*, pueden contener cualquier número de atributos. Ejemplos de *ways* serían un tramo de carretera (polilínea), una sección de un río (polilínea también) o el perímetro de un embalse (polígono).

- **Relation:** Predicado parte-de (*part-of*). Indica que una lista de *nodes* o *ways* pertenecen a alguna forma o región más grande. Esta región es contenedor de estos elementos, así como la parte (o la totalidad) de la región de búsqueda.

Estos tres elementos conforman lo que se conoce como el modelo de geometrías de *Open Street Map*. Dicho modelo resulta especialmente útil así como de gran interés en el presente trabajo por su utilidad para la extracción de información espacial. A continuación puede verse un diagrama explicativo de las citadas geometrías en la Figura 2.18, dónde se aprecia el aspecto de las diferentes geometrías con sus *Tags* asociadas.

DBPedia

El proyecto *DBpedia* extrae información estructurada de la *Wikipedia* y permite que esta información sea accesible en la Web. Lo hace de acuerdo a los

términos de la licencia *Reconocimiento – Compartir* 3.0 de *Creative Commons* y la Licencia de *Documentación Libre* de GNU.

DBpedia permite a los usuarios consultar las relaciones así como las propiedades asociadas a los recursos de la *Wikipedia*, incluyendo enlaces a otros conjuntos de datos relacionados. *DBpedia* está considerada como el núcleo de la nube de datos enlazados (LOD cloud) [Auer et al., 2007].

Panoramio

Es un sitio web para compartir fotos orientado a la geolocalización. Se puede acceder a las fotos subidas por sus usuarios manejándolas como una capa de *Google Earth* o de *Google Maps*. El objetivo de la plataforma es permitir a los usuarios la visualización de las fotos que otros usuarios han tomado en un determinado lugar.

Fue una base de datos abierta (por ejemplo, cuando fue utilizada en los experimentos que se presentan en este documento), aunque actualmente no se encuentra disponible tras haber cerrado *Google* el servicio en el año 2014, para sustituirlo por *Views*.

3. Estado del arte

En este capítulo se describe el estado actual de técnicas relacionadas con el enriquecimiento de fuentes de datos con descripciones lingüísticas mediante bases de datos geográficas externas en la forma que se plantea en este trabajo. Se consideran dos áreas principales:

1. La curación de datos
2. La generación de lenguaje natural (GLN)

3.1. Curación de datos

Actualmente existe una enorme cantidad de fuentes de datos (de todo tipo) disponibles a través de Internet, entre las que se encuentran las sociológicas, estadísticas, contables o geográficas, siendo estas últimas las más interesantes desde el punto de vista del presente trabajo.

Por lo general, en muchas áreas no existen estándares de catalogación ni de etiquetado para los datos, o bien no se utilizan aun existiendo, como viene sucediendo en muchos casos en el área de los datos de sensores o en el de información geográfica. Una tercera alternativa consiste en etiquetados en formatos propios, como los utilizados en *GeoNames* (GN) o en *Open Street Map* (OSM) donde han desarrollado sus propias taxonomías de datos. Además, puede darse el caso de que los datos no tengan un etiquetado correcto, sean erróneos o ambiguos, caso especialmente significativo en aquellos sistemas colaborativos o Wikis, como son *Freebase*, *DBPedia* o nuevamente OSM, que se nutren de las aportaciones altruistas de sus colaboradores.

En dichas situaciones se requiere un procedimiento de curación de datos que permita que los datos sean utilizables para su proceso así como para preservarlos a lo largo del tiempo para futuras consultas o procesos sobre dichos datos.

Existen *frameworks* generales como el Digital Curation Centre - *Curation Lifecycle Model* [Cervone, 2008, Higgins, 2008], que define un proceso completo

con toda una serie de procedimientos así como listas de tareas y comprobación para realizar tal actividad, como se muestra en la Figura 3.1.

Figura 3.1: Elementos clave del DCC *Curation Lifecycle Model*¹⁷.

Cualquier información en formato digital (los datos) es el centro del ciclo de vida de la curación (*Curation Lifecycle*). Entre dichos datos se incluye cualquier *objeto digital sencillo* como pueden ser los archivos de texto, archivos de imágenes o de sonido (con sus identificadores y metadatos relacionados) y también los *objetos digitales complejos*, que consisten en la combinación de objetos digitales sencillos, como son los sitios web. También se incluyen las *bases de datos*. Los diferentes elementos del ciclo son:

- Acciones del ciclo de vida completo:
 - **Descripción y representación de la información.** Asignar metadatos administrativos, descriptivos, técnicos, estructurales y de preservación mediante la utilización de las normas apropiadas para asegurar la descripción y el control adecuado a largo plazo.

¹⁷Imagen obtenida de <http://www.dcc.ac.uk>.

- **Plan de conservación.** Plan para la conservación del material digital durante todo el ciclo de vida de curación. Incluye planes para la gestión y la administración de todas las acciones del ciclo de vida.
 - **Observación y participación de la comunidad.** Mantener una vigilancia adecuada sobre las actividades comunitarias y participar en el desarrollo de estándares, herramientas y *software*.
 - **Curación y preservación.** Ser consciente y comprometerse con la gestión y las acciones administrativas previstas para promover la curación y conservación durante todo el ciclo de vida.
- Acciones secuenciales:
- **Conceptualizar.** Concebir y planificar la creación de datos, incluyendo el método de captura y las opciones de almacenamiento.
 - **Crear o recibir.** Dos opciones, compatibles entre sí:
 - Crear datos incluyendo metadatos administrativos, descriptivos, estructurales y técnicos. La preservación de metadatos también se puede añadir en el momento de la creación.
 - Recibir datos provenientes de creadores de datos (conforme a las políticas de recogida), archivos, repositorios o centros de datos, y si es necesario asignarles los metadatos adecuados.
 - **Evaluación y selección.** Evaluar y seleccionar los datos para la curación y conservación a largo plazo. Adherirse a las guías documentadas, políticas o requisitos legales.
 - **Ingerir.** Transferir los datos a un archivo, repositorio, centro de datos u otro método de custodia. Adherirse a las guías documentadas, políticas o los requisitos legales.
 - **Acción de preservación.** Emprender acciones para garantizar la conservación y la autoría de los datos a largo plazo. Deben asegurarse de que los datos siguen siendo auténticos, confiables y útiles manteniendo su integridad. Entre las acciones de preservación se incluyen:
 - Limpieza de datos
 - Validación

- Asignación de metadatos de preservación
 - Asignación de información de representación
 - Garantizar el uso de estructuras de datos o formatos de archivo aceptables
 - **Almacenar.** Almacenar los datos de manera segura adhiriéndose a las normas pertinentes.
 - **Acceso, uso y reutilización.** Asegurarse de que los datos son accesibles para los usuarios autorizados. Deben aplicarse controles de acceso robustos así como procedimientos de autenticación.
 - **Transformar.** Crear nuevos datos a partir de los originales, como por ejemplo mediante la migración a un formato diferente, o mediante la creación de un subconjunto (por ejemplo tras la ejecución de una *query*) con motivo de su publicación.
- Acciones ocasionales:
- **Desechar.** Desechar aquellos datos que no han sido seleccionados para la curación y preservación a largo plazo, de acuerdo con las políticas, la orientación o los requisitos legales documentados. Normalmente los datos pueden ser transferidos a otro archivo, repositorio, centro de datos u otro método de custodia. En algunos casos se destruyen los datos dado que por la naturaleza de los mismos se puede requerir, por razones legales, su destrucción segura.
 - **Reevaluar.** Reevaluar los datos que fallan en los procedimientos de validación para una nueva evaluación y reelección.
 - **Migrar.** Migrar los datos a un formato diferente. Esto puede hacerse para adecuarlos a nuevos entornos de almacenamiento o para garantizar la inmunidad de los datos ante la obsolescencia del *hardware* o del *software*.

El uso de este tipo de herramientas, en el área de los datos geográficos y especialmente en el de los datos geográficos voluntarios o VGI puede permitir un mejor aprovechamiento por parte de terceros al estandarizar formatos y garantizar la corrección de dicha información.

Enriquecimiento de datos

El enriquecimiento de datos se refiere a los procesos utilizados para mejorar, refinar o elaborar los datos en bruto [Machado et al., 2015]. Esta idea (así como otros conceptos similares) contribuyen a hacer de los datos un activo valioso para cualquier negocio o empresa. También muestra de forma proactiva el imperativo común utilizando estos datos de varias formas.

Existen varias motivaciones para hacer frente a estos procesos de enriquecimiento de datos, que se pueden resumir en las principales características de la gran cantidad de datos originados por las aplicaciones Web 2.0. Por lo tanto, estamos hablando de *big data* y los diferentes demandas de los usuarios y empresas de tratar con ellos, con las implicaciones que ello tiene [Aguiar et al., 2015]. En este contexto se está tratando con diferentes características de los datos como el volumen, diversidad o variedad, la velocidad, la variabilidad y la complejidad.

Ejemplo de experiencias de enriquecimiento de datos son el presentado por [Ritze and Eckert, 2014], donde describen un *plugin JavaScript* así como un API para el enriquecimiento de datos del lado del cliente (independiente de la plataforma). En el contexto del IoT [Qanbari et al., 2015] muestran un *middleware* para descubrimiento de patrones ocultos en base a variables correlacionadas, que enriquece los datos con la información descubierta en los datos en bruto, proporcionando una mayor utilidad a dichos datos.

3.2. Generación de lenguaje natural

3.2.1. Expresiones de referencia

Las expresiones de referencia indican cómo referirse a entidades o elementos que se presentan en textos. Existen diferentes algoritmos para la generación de expresiones de referencia. Uno de los algoritmos más conocidos es el algoritmo incremental de [Reiter and Dale, 1992], que ha servido como base a otras propuestas como por ejemplo el algoritmo Sensible al Contexto de [Krahmer and Theune, 2002].

Generación de Expresiones de Referencia (GER) es una de las tareas o mó-

dulos de la Generación de Lenguaje Natural (GLN) en torno a la cual más trabajos de investigación se ha desarrollado. Esto se debe principalmente al papel principal que la actividad juega en el proceso de comunicación entre entidades. Además, numerosos estudios apuntan a que una gran mayoría de Sistemas GLN incorporan un módulo de GER independientemente de su propósito.

La GLN es un campo de la Lingüística Computacional y la Inteligencia Artificial, que a su vez interacciona con otras muchas áreas como la Ciencia Cognitiva y la Interacción Hombre-Máquina.

Tal y como se define en [Reiter et al., 2000], la GER es “lo que respecta a como se produce una descripción de una entidad de forma que el receptor u oyente sea capaz de identificar esa identidad en un contexto dado”. Esto conlleva a un problema no trivial que es el de identificar la forma en la que se va a hacer referencia a esa entidad de forma que sea identificable así como que no contengan información innecesaria de acuerdo a las pautas definidas en las máximas de [Grice, 1975].

Siguiendo las directrices establecidas en [Reiter and Dale, 1995], existen diversas formas de llevar a cabo una expresión de referencia dando lugar a una clasificación en base a las condiciones del discurso actual, es decir, si el objeto a referenciar (*target*) se ha presentado en el discurso o no:

- Introducción inicial: No existe demasiados trabajos realizados en torno a este tipo de expresión. Habitualmente, se lleva a cabo la estrategia de la utilización de nombres propios o la localización del objeto o entidad para referirse a él e introducirlo en el discurso.
- Pronombres: El uso de los pronombres ha sido ampliamente estudiado. Sin embargo, su generación carece de investigación, por lo que la estrategia habitual es el uso de pronombres en caso de que la entidad haya sido introducida en el discurso o referenciada en sentencias anteriores y no exista otra entidad a la que se pueda adjudicar esa referencia. El uso de pronombres es una estrategia común en referencias posteriores.
- Descripciones precisas: Se trata del área con mayores resultados en investigación. Existen numerosos algoritmos y paradigmas asociados a

esta tarea.

A continuación se va a revisar el algoritmo incremental a modo de ejemplo del funcionamiento de los algoritmos de GER. También se revisará el algoritmo sensible al contexto, consecuencia de este primero.

Algoritmo Incremental de Reiter y Dale

[Reiter and Dale, 1992] describen un algoritmo para la GER dentro de un sistema de GLN. Este algoritmo es una mejora de otro previamente existente, basado en evidencias psicolingüísticas y en el análisis de transcripciones de diálogos entre humanos. Este algoritmo supone una línea de base para comparar el funcionamiento y rendimiento de algoritmos similares.

El algoritmo de Reiter y Dale se basa en las siguientes suposiciones sobre la base de conocimiento subyacente empleada:

1. Toda entidad se caracteriza por una colección de atributos y sus respectivos valores
2. Toda entidad tiene, entre sus atributos, un tipo
3. La base de conocimiento puede organizar los valores de algunos atributos en forma de jerarquía de subsunción

Además, este algoritmo depende de la siguiente información para tomar sus decisiones finales:

- Cada objeto representado en el sistema debería tener un valor básico que se corresponde con el concepto preferido para referirse a dicho objeto. Esto se usa para ofrecer un punto de partida a partir del cual se puedan construir las referencias al objeto: los conceptos más cercanos a este valor básico en la taxonomía serán preferibles a conceptos más lejanos. Cada objeto puede tener asignadas diferentes clases de nivel básico para distintos usuarios. Con esta información es posible decidir si se quieren generar las referencias de manera más específica o más general.
- Para cada objeto debe haber un modo de determinar si el usuario (la persona para que el sistema está generando el texto) es consciente de

que un determinado par atributo-valor es aplicable a dicho objeto. Esto ayuda a determinar si la mención de una característica particular le resultaría útil al usuario para identificar el objeto.

- Se debe disponer de una lista de atributos preferidos o *preferred attributes* que describe el conjunto de atributos que los seres humanos prefieren usar al describir objetos (en su dominio específico, Reiter y Dale observaron que los atributos preferidos eran el tipo, el tamaño, la forma y el color).

Para construir una referencia a una entidad en particular, el algoritmo toma como entrada un símbolo correspondiente al referente objetivo (aquel para el que se quiere generar la referencia) y una lista de símbolos que corresponden a otras entidades dentro del foco de atención, basado en el referente objetivo y denominado conjunto de contraste. El algoritmo devuelve una lista de pares atributo-valor correspondiente al contenido semántico de la expresión de referencia que se está generando. El algoritmo opera iterando sobre la lista de atributos disponibles, buscando aquel que, siendo conocido por el usuario, elimina el mayor número de elementos restantes del conjunto de contraste. El valor básico considerado para los atributos se usa para dar preferencia a determinados atributos cuando esta decisión no sea clara.

Además, el algoritmo escoge el valor más apropiado para un cierto atributo a partir de un valor inicial. La idea es usar un valor para ese atributo que sea subsumido por el valor inicial, que describa con precisión al referente objetivo, que elimine el mayor número de distractores posibles y que, siguiendo estas restricciones, sea lo más próximo posible al nivel inicial de la taxonomía. Así, se evita usar una referencia como “*the chihuahua*” o “el chihuahua” en un contexto donde solo haya un único perro, en cuyo caso la referencia más adecuada sería simplemente “*the dog*” o “el perro” .

Algoritmo Sensible al Contexto de Krahmer y Theune

[Krahmer and Theune, 2002] presentan una variante del *Algoritmo Incremental de Reiter y Dale* que tiene como novedad que es sensible al contexto. Cada entidad del dominio tiene asignado un cierto grado de prominencia o *salience*, y las descripciones son generadas con respecto a un conjunto de contraste definido como el conjunto de entidades que tienen igual o mayor

prominencia que la entidad para la que se está generando la referencia. De esta manera el algoritmo permite la generación de referencias más escuetas para las entidades prominentes como podrían ser las entidades que han sido mencionadas recientemente en el discurso.

La principal diferencia entre el Algoritmo Incremental y esta aproximación es la condición de éxito. Mientras que en el Algoritmo Incremental la ejecución finaliza satisfactoriamente cuando el conjunto de contraste queda vacío, el Algoritmo Sensible al Contexto comprueba si la entidad referida es la más prominente entre aquellas que corresponden al conjunto de propiedades seleccionado para generar la referencia.

Si la referencia corresponde a la entidad más prominente, entonces se considera exitosa, incluso si corresponde también a otras entidades del contexto que son menos prominentes. De esta forma, si se asume que inicialmente todas las entidades tienen el mismo grado mínimo de prominencia, introducir una nueva entidad requeriría distinguirla de todos los elementos en el conjunto de contraste, mientras que volver a referirse a una entidad ya mencionada requiere solamente distinguirla de las entidades que tiene el mismo o mayor grado de prominencia.

3.2.2. El caso de las expresiones de referencia geográficas

En este apartado se hace una revisión del estado del arte de técnicas de generación de expresiones de referencia en dominios geográficos.

Sistemas que generan textos a partir de georreferencias [Dale et al., 2005], [Turner et al., 2006, Turner et al., 2008, Thomas and Sripada, 2008] o también en [Reiter and Dale, 2009] tienen en común que todos necesitan expresar eventos, sucesos o previsiones en determinadas regiones o zonas geográfica, con lo que la generación de descripciones o expresiones en lenguaje natural que hagan referencia a esas zonas supone un campo de investigación en el lenguaje natural. Ejemplos de este tipo de expresiones son los siguientes: *in some places in the southwest, in the rural areas, at the end of the street, in southern areas, etc.*

La mayoría de los sistemas que han trabajado en la generación de expresiones espaciales se han centrado en espacios pequeños: comunicación entre robots [Kelleher and Kruijff, 2006], animaciones en 3D [Lester and Bares, 1998] y escenas sencillas de visualizar [Viethenand and Dale, 2008, Ebert et al., 1996]. En algunos de estos sistemas se muestran aproximaciones para identificar contextos espaciales actuales.

Se han diseñado métodos para generar descripciones de rutas [Dale et al., 2005, Moulin and Kettani, 1999] así como también se han generado descripciones basadas en *landmarks* [Varges, 2005] haciendo uso del corpus de *maptask*. Sin embargo, cabe destacar el trabajo realizado en [Turner et al., 2009] siguiendo las líneas en [Turner et al., 2006] ya que además de generar expresiones en espacios de gran tamaño, llevan a cabo una filosofía rompiendo con los algoritmos de generación de expresiones de referencia hasta el momento. El área de expresiones Geográficas supone una gran oportunidad debido a la escasez de trabajo realizado y la aplicabilidad que tiene.

Tal y como se apunta en [Reiter and Dale, 2009], las técnicas clásicas de generación de expresiones de referencia, debido a las restricciones de brevedad que requieren los sistemas y el tamaño del contexto (miles de puntos) quedan limitadas a la solución del problema en la generación de este tipo de expresiones ya que requiere un coste computacional alto.

Los eventos asociados a estos sistemas se dan en regiones determinadas, no en puntos concretos, y el tratamiento de éstos se realiza de forma aproximada o vaga y no exacta. Por todo ello, se argumenta un enfoque más relajado sobre el concepto de expresión de referencia adoptando una forma de no necesariamente obtener una descripción que distinga por completo al referente (selección de propiedades que obtengan el menor número de distractores).

Las principales limitaciones de los paradigmas clásicos radican por tanto en que se trata de paradigmas que trabajan con contextos definidos, habitualmente de poco tamaño y que en contextos reales de mayor complejidad pierden eficiencia. Los paradigmas están asociados a generar información referencial y existe una carencia de información navegacional [Dale and Reiter, 1995]: se presupone que el referente se encuentra en el foco de atención del

oyente. Sin embargo, nuevas técnicas relacionales y en las que se definen *landmarks* en incluso se les asigna un determinado ratio de “capacidad de ser identificado” (*salience*) permiten generar expresiones en base a un punto de referencia, línea o región.

Uno de los retos en la generación de expresiones de referencia consiste en la generación de referencias en las cuales existen varios referentes. Este campo es uno de los más amplios en los que se puede trabajar, ya que no hay demasiadas aportaciones al respecto.

La gran mayoría de sistemas desarrollados hasta el momento llevan a cabo expresiones en las que los predicados se definen como relaciones uno a uno o incluso de 1 a 2. Sin embargo, existe un campo abierto a la generación de relaciones de tipo N-arias.

Los cuantificadores es otro de los temas en los que no se haya demasiada bibliografía. Sin embargo, resultados empíricos demuestran que el ser humano hace uso de cuantificadores incluso aunque sean innecesarios.

Otro de los retos asociados a la generación de expresiones de referencia, es la tendencia a la mejora o extensión de los algoritmos clásicos en alguna de las suposiciones realizadas pero la no integración de cada mejora. Existe una necesidad de estandarización y unificación de algoritmos.

Por último cabe mencionar la necesidad del estudio de una mayor interacción entre los sistemas y el usuario, de modo que se busque una adaptación al entorno en el que se encuentra el usuario, las necesidades de éste, los conocimientos y capacidad de percepción de los usuarios (principio de sensibilidad [Dale and Reiter, 1995]). En esta área, pocos son los trabajos que se han realizado. En el algoritmo clásico se realiza una aproximación a este tipo de principio mediante la función *UserKnows()*. Sin embargo, de ahí en adelante pocos son los avances en esta materia.

Por otro lado, la incorporación de *Fuzzy Ontologies* (basadas en razonamiento en lógica difusa), pueden aportar un nuevo enfoque en el razonamiento espacial y cualitativo, así como la cobertura de numerosos conceptos difusos.

4. Discusión

La revisión del estado del arte muestra que el desarrollo tecnológico y las propuestas de la comunidad de investigación han permitido disponer de un número de aportaciones para poder resolver el problema que se plantea en el presente trabajo. En concreto, dichas aportaciones permiten:

- Disponer de soluciones técnicas avanzadas para construir y operar con amplias y complejas redes de sensores, sobre las que se dispone de conocimiento formalizado y almacenado de una forma conveniente para ser utilizado con medios computacionales.
- Concebir etapas de curación de datos bajo un enfoque metodológico para enriquecer bases de datos con información útil para facilitar la realización eficiente de procesos de tratamiento de datos.
- Acceder a bases de datos abiertas con un gran volumen de información geográfica como por ejemplo *Open Street Map*, *Geonames*, etc.
- Hacer uso de algoritmos existentes para generar automáticamente textos en lenguaje natural y expresiones de referencia.

Sin embargo, se pueden citar los siguientes aspectos pendientes que todavía no están resueltos para el problema que se plantea en el presente trabajo:

- La generación de expresiones de referencia geográficas requiere de un gran volumen de conocimiento. No existen soluciones que abordan este problema de forma suficiente para redes complejas de sensores.
- Las bases de datos abiertas son imperfectas. Tienen información que puede ser incompleta, inconsistente o con importantes restricciones en los procedimientos de búsqueda, para asegurar un adecuado nivel de eficiencia en la generación de respuestas.

Parte III

PROPUESTA

5. Método Propuesto

En el presente capítulo se describe el método propuesto como solución al problema que se plantea en este trabajo. En primer lugar, se muestran las hipótesis de partida y una descripción general del método. Después se describe haciendo uso de un enfoque metodológico basado en la metodología *CommonKADS*. Finalmente se describe un ejemplo ilustrativo que muestra las características de dicho método.

5.1. Hipótesis de partida

Como se ha descrito en los capítulos anteriores, en el presente trabajo se aborda el problema de enriquecimiento de datos sobre redes de sensores con expresiones en lenguaje natural sobre localización espacial y temporal. Una de las principales utilidades de la solución a este problema es que con ello será posible disponer de información necesaria para generar presentaciones en lenguaje natural más complejas sobre redes de sensores de forma eficiente, facilitando con ello el acceso a los datos a un amplio número y diversidad de usuarios.

Las hipótesis de partida que se asumen para plantear la solución al problema considerado son las siguientes:

- La solución que se plantea debe ser general, es decir, transportable a distintas redes de sensores de diferentes dominios.
- Para poder disponer de conocimiento sobre localización geográfica y temporal, la solución propuesta hace uso de conocimiento existente en bases de datos abiertas. La solución propuesta debe tener en cuenta que la información disponible en dichas bases puede ser imperfecta (incompleta, incoherente, etc.).
- La solución puede operar sin restricciones importantes de tiempo, de ejecución (por ejemplo, sin restricciones de operación en tiempo real), dado que funciona como una fase de enriquecimiento de datos.
- El rendimiento de la solución propuesta debe mejorar significativamente el rendimiento de otros métodos existentes (por ejemplo, algoritmos básicos no informados).

Figura 5.1: Generación de descriptores lingüísticos.

5.2. Planteamiento general del método

En esencia, el objetivo del método es enriquecer la información sobre la red de sensores con descripciones lingüísticas haciendo uso de fuentes de información abiertas de Internet. En la Figura 5.1 puede observarse el contexto de operación del método propuesto.

El método recibe como entrada una colección de datos sobre redes de sensores con distribución espacio-temporal $\{D_1, D_2, \dots, D_n\}$. Dicha colección reúne información sobre la red de sensores, entre la que se encuentra la siguiente información:

- Naturaleza de la medida
- Medida obtenida en espacio y en tiempo
- Información topológica (subredes, etc.)

El método genera como salida un conjunto de descriptores lingüísticos $\{L_1, L_2, \dots, L_m\}$ que realizan el proceso de enriquecimiento de dichos datos. Para ello, el método utiliza también como entrada información de bases de datos abiertas como son:

- *Open Street Map*
- *GeoNames*
- *DBpedia*
- *Panoramio*
- etc.

Para definir el procedimiento que sigue el método, como se ha mencionado, se tiene en consideración que la información presente en las bases de datos abiertas es imperfecta y por tanto, la búsqueda de los datos requeridos puede no dar resultados. Por ejemplo, se puede buscar el nombre de un río en una base de datos como *Open Street Map* y puede no encontrarse dicho valor.

Para manejar esta situación, se plantea hacer uso de varias fuentes con información redundante de tal forma que si la información requerida no se encuentra (o no es de suficiente calidad) en una fuente de datos, se intenta encontrar en otras fuentes que puedan disponer de dicha información. El método que se plantea sigue una estrategia de generación y prueba que realiza una búsqueda dirigida por conocimiento heurístico (conocimiento que indica en qué orden buscar en las fuentes de datos) hasta encontrar una descripción satisfactoria.

Además de ello, se plantea que el método incluya conocimiento heurístico sobre tipos de sensores que exprese cómo describir en lenguaje natural dicha información. Así, se asume que es posible expresar de forma general para cada tipo de sensor (pluviómetro, sensor de temperatura, sensor de caudal, etc.) una “forma de hablar” del mismo basándose en sus características.

De acuerdo con ello, el método propuesto se divide en cinco procesos, apoyados en varias bases de conocimiento (ver Figura 5.2):

Figura 5.2: Procesos principales del método propuesto.

1. **Preprocesar datos.** Se trata del primer proceso, donde se realizan diferentes acciones o preprocesos con el fin de lograr que los datos de entrada sean uniformes.
2. **Generar hipótesis de descripción.** El objetivo de este paso de inferencia es generar un patrón de descripción lingüística sobre cómo expresar la información del sensor.
3. **Seleccionar fuente de datos.** El objetivo de este paso de inferencia es seleccionar qué fuentes abiertas de datos se deben consultar de acuerdo a la hipótesis de descripción.
4. **Buscar en fuente de datos.** El objetivo de este paso de inferencia

es acceder a fuentes abiertas de datos para obtener información correspondiente a las características de las referencias.

5. **Componer descripción.** Se combinan los resultados obtenidos por los procesos previos. De la combinación se obtiene la descripción lingüística en el formato apropiado.

5.3. Metodología aplicada

El método que se propone debe dirigir búsquedas en bases de datos abiertas y construir patrones de lenguaje natural con ayuda de conocimiento de naturaleza heurística. Debido a ello, se ha considerado conveniente hacer uso de la metodología de ingeniería de conocimiento *CommonKADS* que facilita la construcción de modelos de esta naturaleza. Además, para formalizar los contenidos de las bases de conocimiento se hace uso del lenguaje de la lógica.

Figura 5.3: Arquitectura *CommonKADS* ¹⁸.

La metodología *CommonKADS* permite realizar el análisis y la construcción de las fases necesarias en sistemas de conocimiento, como describe [Schreiber et al., 2000]. Para formalizar el método que se presenta en este trabajo se

¹⁸Imagen obtenida de [Schreiber et al., 2000].

hace uso de la parte de *CommonKADS* para definición de modelos de conocimiento. La metodología ofrece varias vistas para la definición de dicho modelo, clasificadas en tres categorías:

1. **Conocimiento de tareas.** Categoría que especifica las metas que persigue la aplicación y cómo se alcanzan esas metas mediante la realización de los trabajos gracias a la descomposición en subtarefas así como en inferencias finales.
2. **Conocimiento de inferencia.** Categoría que especifica los pasos básicos de inferencia que se quieren realizar haciendo uso del conocimiento del dominio.
3. **Conocimiento del dominio.** Categoría que especifica el conocimiento específico del dominio así como los tipos de la información que se maneja en dicho dominio.

Figura 5.4: Modelo de conocimiento *CommonKADS* en dominio médico ¹⁹.

En los próximos apartados se detallan cada una de las categorías del método propuesto para el problema de enriquecimiento planteado. En los anexos está el modelo completo con el lenguaje de especificación utilizado por la metodología *commonKADS*.

¹⁹Imagen obtenida de [Schreiber et al., 2000].

5.4. Modelo de tareas y métodos

Como se ha indicado previamente, se muestra a continuación el modelado del método de acuerdo a la metodología *CommonKADS*. En este apartado se aborda la descomposición en tareas.

Figura 5.5: Descomposición en tareas del método.

En la Figura 5.5 se muestra dicha descomposición. En esta figura pueden observarse los dos componentes utilizados para la representación de conocimiento de tareas: *tarea* y *método*. En concreto se tienen dos tareas, *Enriquecer-datos-sobre-sensores* y *Planificar-contenido-de-la-descripción*, las cuales se realizan con los métodos *Enriquecimiento-con-descripciones* y *Planificar-contenido-con-Open-Data*.

Cada una de estos dos métodos requiere a su vez la ejecución de varias inferencias, tres en el caso del segundo método y dos en el caso del primero. En la descripción de los algoritmos, que se ven más adelante, se muestran con la terminología de *CommonKADS* como *TASK* y *TASK-METHOD*, respectivamente. Seguidamente se describe cada una de las tareas.

La tarea *Enriquecer-datos-sobre-sensores* tiene una correspondencia al más alto nivel con el problema que se considera en este trabajo. Es la encargada de dirigir las acciones encaminadas a realizar el enriquecimiento de fuentes de datos. Esta tarea se lleva a cabo con el método *Enriquecimiento-con-descripciones*.

Algoritmo 5.1 Método Enriquecimiento-con-descripciones.

```

TASK-METHOD Enriquecimiento-con-descripciones;
REALIZES: Enriquecer-datos-sobre-sensores;
DESCOMPOSITION:
  INFERENCES:
    Preprocesar-datos, Componer-descripción;
ROLES:
  INTERMEDIATE:
    Datos-corregidos: "Datos de sensores en un
                      formato unificado";
    Datos-de-un-sensor: "Datos en formato unificado
                        de un único sensor";
    Plan-de-la-descripción: "Estrategia para construir
                            la descripción de un sensor";
    Descripción-de-un-sensor: "Descripción relativa
                              a un único sensor";
    Descripciones: "Agrupamiento de las descripciones de
                  los sensores";
CONTROL-STRUCTURE:
  Descripciones := [];
  Preprocesar-datos (
    Datos-en-bruto -> Datos-corregidos);
  FOR-EACH Datos-de-un-sensor IN Datos-corregidos DO
    Planificar-contenido-de-la-descripción (
      Datos-de-un-sensor -> Plan-de-la-descripción);
    Componer-descripción (Plan-de-la-descripción ->
      Descripción-de-un-sensor);
    Descripciones := Descripciones +
                    Descripción-de-un-sensor;
  END FOR-EACH
END TASK-METHOD Enriquecimiento-con-descripciones;

```

El Algoritmo 5.1 muestra la especificación *CommonKADS* del método con el lenguaje utilizado por dicha metodología. Esta especificación muestra el régimen de control seguido por el método con una iteración para cada sensor que incluye llamadas a las inferencias y a la tarea involucradas en el proceso de generación de descripciones.

Figura 5.6: Método *Enriquecimiento-con-descripciones*.

En la Figura 5.6 se muestra la estructura de inferencia para el método *Enriquecimiento-con-descripciones*. Dicho método describe cómo se realiza la tarea *Enriquecer-datos-sobre-sensores* mediante su descomposición en sub-tareas. Entre dichas sub-tareas se encuentran los pasos de inferencia *Preprocesar-datos* y *Componer-descripción* (que se ven más adelante) así como la tarea *Planificar-contenido-de-la-descripción*.

Cabe destacar que en la estructura de inferencia se utilizan dos bases de conocimiento (conocimiento estático de entrada) para que, dados como entrada los *Datos-en-bruto* obtener a la salida las *Descripciones*. La tarea *Planificar-contenido-de-la-descripción* es la encargada de planificar la generación de descripciones lingüísticas para el enriquecimiento de un sensor. Para ello hace uso del método *Planificar-contenido-con-Open-Data*. El Algoritmo 5.2 muestra el régimen de control del método que incluye una iteración para cada componente de la hipótesis de descripción con una secuencia de llamadas a las inferencias involucradas en el proceso de búsqueda en las bases de datos externas.

El método *Planificar-contenido-con-Open-Data*, describe cómo se realiza la tarea *Planificar-contenido-de-la-descripción* mediante su descomposición en pasos tres pasos de inferencia. La Figura 5.7 muestra la estructura de infe-

rencia para el método *Planificar-contenido-con-Open-Data*.

Algoritmo 5.2 Método Planificar-contenido-con-Open-Data.

```

TASK-METHOD Planificar-contenido-con-Open-Data;
REALIZES: Planificar-contenido-de-la-descripción;
DESCOMPOSITION:
  INFERENCES:
    Generar-hipótesis-de-descripción ,
    Seleccionar-fuente-de-datos ,
    Buscar-en-fuente-de-datos;
  ROLES:
    INTERMEDIATE:
      Estructura-de-la-hipótesis: "Patrón lingüístico que
        corresponde al caso del sensor tratado".
      Componentes-requeridos-de-la-hipótesis: "Componentes
        requeridos en la Estructura-de-la-hipótesis que
        hay que buscar en las fuentes de datos externas";
      Consulta-en-fuente-de-datos: "Consulta con los
        componentes requeridos para las fuentes de datos";
      Éxito-de-la-búsqueda: "Indicador del resultado de la
        consulta, si ha sido exitosa o ha fallado".
      Componentes-encontrados: "Los componentes que se han
        obtenido de las fuentes de datos".

CONTROL-STRUCTURE:
  Generar-hipótesis-de-descripción(
    Datos-de-un-sensor ->
    Estructura-de-la-hipótesis ,
    Componentes-requeridos-de-la-hipótesis );
  Componentes-encontrados = [];
  FOR-EACH C IN Componentes-requeridos-de-la-hipótesis DO
    REPEAT
      Seleccionar-fuente-de-datos(C ->
        Consulta-en-fuente-de-datos );
      Buscar-en-fuente-de-datos(C ->
        Éxito-de-la-búsqueda, Componente);
      Componentes-encontrados =
        Componentes-encontrados + Componente;
    UNTIL
      Éxito-de-la-búsqueda = false OR
      "no queden más fuentes de datos";
    END WHILE
  END FOR-EACH
  Plan-de-la-descripción :=
    Estructura-de-la-hipótesis + Componentes-encontrados;
END TASK-METHOD Planificar-contenido-con-Open-Data;

```

Dicho método describe cómo se realiza la tarea *Planificar-contenido-de-la-descripción* mediante su descomposición en subtareas. Entre dichas subtareas se encuentran los pasos de inferencia *Generar-hipótesis-de-descripción*, *Seleccionar-fuente-de-datos* y *Componer-descripción* (que se ven más adelante) así como la tarea *Buscar-en-fuente-de-datos*.

Figura 5.7: Método *Planificar-contenido-con-Open-Data*.

Se puede destacar en referencia a la estructura de inferencia *Planificar-contenido-con-Open-Data* que utiliza dos bases de conocimiento (conocimiento estático de entrada). A partir de los *Datos-de-un-sensor*, este método permite generar a la salida los *Componentes-encontrados* para dicho sensor. Para obtener dichos componentes se realiza toda una serie de procesos mediante consultas a bases de datos externas.

Como se puede observar, la organización en las diferentes tareas, bases de conocimiento e inferencias mostrada constituye una solución modular de carácter general, es decir, dicha organización es independiente del dominio específico y por tanto es potencialmente aplicable a diferentes tipos de redes de sensores.

5.5. Modelo del Dominio

Figura 5.8: Componentes principales del modelo del dominio.

En la Figura 5.8 se muestra la visión general del modelo del dominio. De acuerdo con dicha figura, los componentes del dominio son los siguientes:

- Información dinámica de entrada
- Información intermedia
- Información estática de entrada (bases de conocimiento)
- Información de salida

En las próximas secciones se describe cada uno de los elementos.

5.5.1. Información dinámica de entrada

Se asume que la información dinámica de entradas es la siguiente:

1. **Red-de-sensores.** Datos de entrada sobre la red de sensores. Entre la información sobre la red de sensores se encuentra la relativa a la topología de la red. Para cada sensor se incluye la siguiente información:
 - a) *TipoSensor*(x, y): El sensor x es de tipo y (por ejemplo, pluviómetro, sensor de caudal, etc.).
 - b) *CódigoSensor*(x, y): El sensor x tiene el código y (código alfanumérico utilizado por los operadores y otros sistemas informáticos para su identificación).
 - c) *FrecuenciaMedida*(x, y): La frecuencia de medida del sensor x es y (y en unidad de tiempo, por ejemplo en segundos).
 - d) Topología de la red. Otra información relacionada con la topología de la red expresada como relaciones entre sensores, subredes (por ejemplo, organizadas por zonas geográficas o centros de control locales), etc.
2. **Localización-espacio-temporal.** Datos de entrada sobre la localización espacio temporal de los sensores así como sobre las mediciones realizadas por dichos sensores de la red. Existen diferentes opciones, dependiendo de si los sensores son fijos o móviles:
 - a) *Tiempo*(s, m, x): El valor m del sensor s se ha obtenido en el instante x .
 - b) *Latitud*(s, m, x): El valor m del sensor s se ha obtenido en la latitud x .
 - c) *Longitud*(s, m, x): El valor m del sensor s se ha obtenido en la longitud x .
 - d) *Latitud*(s, x): La latitud del sensor s es x (para sensor estático).
 - e) *Longitud*(s, x): La longitud del sensor s es x (para sensor estático).

Las colecciones de datos pueden proceder de diferentes formatos, siendo posible utilizarlas mediante el empleo de los conectores e interfaces apropiados. Entre las más comunes se tienen las siguientes:

- **BBDD:** Bases de datos relacionales accesibles mediante sintaxis SQL. También es posible encontrar bases de dato no relacionales o no SQL.

- *XML*: Todo tipo de ficheros de datos etiquetados mediante el lenguaje de marcas extensible.
- *CSV* : Ficheros de datos tabulados separados por comas o cualquier otro tipo de delimitador acordado.
- *JSON* : Objetos en notación JavaScript.

5.5.2. Información intermedia y de salida

La información intermedia (dinámica) que se intercambia entre las inferencias del método y la información de salida es la siguiente:

1. ***Patrón-lingüístico***. Información que incluye *Características* así como relaciones entre ellas (ver más información más adelante).
2. ***Características***. Se trata de las características necesarias para conformar los patrones lingüísticos. Algunos ejemplos son *río*, *población* o *duración*.
3. ***Referencias***. Se trata de instancias o elementos concretos de las *Características*. Ejemplos de estas referencias son *Logroño*, *Júcar* o *Duero*.
4. ***Parámetros-de-búsqueda***. Parámetros requeridos para realizar una búsqueda en las fuentes de *Open Data* entre los que se encuentran:
 - a) Fuente de datos *Open Data*. Por ejemplo *Open Street Map*
 - b) *Latitud(x)*: La latitud es x
 - c) *Longitud(x)*: La longitud es x
 - d) Atributos. Los atributos solicitados en la búsqueda, como por ejemplo: *RIVER*, *STREAM*, *CANAL*, *DRAIN*, *WATER* o *RIVERBANK*
5. ***Éxito/Fracaso***. Indicador del estado de una solicitud de búsqueda de información, donde los dos posibles estados son:
 - a) *Éxito*. El resultado de la búsqueda ha sido un éxito
 - b) *Fracaso*. No ha sido posible encontrar los atributos solicitados en la búsqueda

6. **Descripciones-lingüísticas.** Los textos generados como resultado de la ejecución del método de enriquecimiento. Por ejemplo: *Río Vinalopó a su paso por Elda.*

En los próximos apartados se describe con más detalle la información correspondiente a los patrones lingüísticos.

Patrones lingüísticos

El método que se propone maneja expresiones de localización espacial/temporal, por lo que las descripciones que se utilizan incluyen:

- Referencias. Nombres de puntos de referencia (landmarks) relativos a características geográficas o temporales (por ejemplo, población, río, área montañosa, embalse, acontecimiento histórico, fiesta religiosa, etc.).
- Relaciones. Relaciones espacio-temporales entre puntos de referencia (por ejemplo, relaciones correspondientes a los términos “en”, “al lado de”, “a su paso por”, “entre”, etc.).

Las relaciones están tipificadas de acuerdo con algún método de razonamiento espacial tal como Region Connection Calculus (RCC8) u otras con mayor poder expresivo. La representación general que se maneja para el patrón de descripción es la siguiente:

Patrón(
 [Referencia(x_1 , característica – 1, parámetros – 1),
 Referencia(x_2 , característica – 2, parámetros – 2), ...],
 [Relación(relación – 1, argumentos – 1),
 Relación(relación – 2, argumentos – 2), ...])

Por ejemplo:

Patrón(
 [Referencia(x_1 , río, coordenadas(38,4831809997559, 0,796071231365204)),
 Referencia(x_2 , población, coordenadas(38,4831809997559, 0,796071231365204))],
 [Relación(en, x_1, x_2)])

Dicho patrón se refiere a una descripción en texto que se inicia con el nombre del río que se localiza en la coordenada geográfica indicada. La descripción

tiene a continuación la relación “en”. Finalmente la descripción incluye el nombre de la población que está en la coordenada geográfica indicada. Un ejemplo de descripción construida con este patrón sería la siguiente:

“Río Vinalopó a su paso por Elda”

Como alternativa a la descripción anterior, las relaciones del patrón pueden expresarse en forma de plantilla de texto en donde los huecos corresponden a puntos de referencia que es necesario completar. Por ejemplo:

Patrón(
[Referencia(x_1 , río, coordenadas(38,4831809997559, 0,796071231365204)),
Referencia(x_2 , población, coordenadas(38,4831809997559, 0,796071231365204))],
[Relación([x_1 “a su paso por” x_2]])]
)

Las referencias que ocupan los huecos se indican con semántica explícita, es decir, expresando qué clase de contenido ocupa cada posición. Para expresar dicha semántica puede manejar un conjunto de categorías geográficas y temporales de acuerdo con una ontología elegida, tales como río, embalse, área montañosa, etc.

5.5.3. Información estática de entrada (bases de conocimiento)

Las cuatro bases de conocimiento (información estática de entradas estáticas) utilizadas por las inferencias son las siguientes:

Clases-de-sensores. Base de conocimiento sobre las características de los sensores. Los siguientes predicados ilustran el posible contenido de esta base:

1. *MideCantidadSensor(x , y)*: El sensor de tipo x mide la cantidad y .
2. *UnidadMedida(x , y)*: La unidad de medida de la cantidad x es y (por ejemplo, mm/h, metros, etc.).
3. *IntervaloVálido(x , y , z)*: El intervalo válido de valores de la cantidad x tiene como valor mínimo y , y valor máximo z .
4. *Fiabilidad(x , y)*: La fiabilidad del tipo de sensor x es y .
5. *Tolerancia(x , y)*: La tolerancia del tipo de sensor x es y .

6. *Móvil*(x): El sensor de tipo x es móvil.

7. *Fijo*(x): El sensor de tipo y es fijo.

Selección-de-patrones-lingüísticos. Base de conocimiento que incorpora un conjunto de reglas que permiten seleccionar los patrones lingüísticos. La representación utilizada en esta base es un conjunto de reglas con el formato siguiente:

$$\text{Condiciones} - \text{del} - \text{sensor} \rightarrow \text{Patrón} - \text{de} - \text{descripción}$$

Es decir, reglas que incluyen en el antecedente condiciones sobre las características del sensor y, en el consecuente, el patrón utilizar. Un ejemplo de regla puede ser:

$$\begin{aligned} & \text{TipoSensor}(s, \text{caudal}), \text{Latitud}(s, x), \text{Longitud}(s, y) \rightarrow \\ & \text{Patrón}([\text{Referencia}(x_1, \text{río}, \text{coordenadas}(x, y)), \\ & \text{Referencia}(x_2, \text{población}, \text{coordenadas}(x, y))], \\ & [\text{Relación}(\text{en}, x_1, x_2)]) \end{aligned}$$

Las reglas pueden estar ordenadas de mayor a menor prioridad, de forma que el proceso de inferencia que selecciona los patrones, selecciona antes las primeras reglas y, por tanto, en primer lugar los patrones mayor prioridad.

Nótese que este conocimiento puede estar expresado de forma muy general y, por tanto, ser reutilizable por amplias familias de sensores. Se trata de conocimiento cuyo coste de adquisición es bajo y, también, la frecuencia de modificación es baja dado que normalmente, los tipos de sensores no suelen cambiar demasiado en una instalación. Al mismo tiempo, esta representación da la flexibilidad suficiente para expresar descripciones muy específicas y particulares en dominios que así lo requieran. Las reglas proporcionan flexibilidad para representar los criterios utilizados para crear descripciones de sensores por los expertos. Las reglas ayudan a generar la descripción, teniendo en cuenta la magnitud física medida, algunas características geográficas y otros criterios establecidos por expertos en el dominio.

Criterios-de-selección-Open-Data. Esta base de conocimiento específica que indica qué fuentes abiertas se utilizan y qué criterios de selección de dichas fuentes se aplican. La base de conocimiento relaciona para cada característica (población, río, área-montañosa, lugar-histórico, embalse, etc.), el conjunto de funciones de acceso a los datos que pueden extraer valores de dicha característica, junto con los parámetros necesarios para realizar el acceso. El formato para esta base es el siguiente:

$$\text{Característica}(x_1, x_2, \dots) \rightarrow \text{FunciónDeConsulta}(x_1, x_2, \dots, y_1, y_2, \dots)$$

Por ejemplo:

$$\begin{aligned} &\text{Característica}(\text{río}, \text{Coordenadas}(x, y)) \rightarrow \\ &\text{FunciónDeConsulta}(\text{OpenStreetMap}, \text{Coordenadas}(x, y), \\ &\text{RIVER}, \text{STREAM}, \text{CANAL}, \text{DRAIN}, \text{WATER}, \text{RIVERBANK}) \end{aligned}$$

Recursos-léxico-lingüísticos. Base de conocimiento que incorpora colecciones de recursos léxicos lingüísticos susceptibles de ser empleados en la construcción de las descripciones lingüísticas. Por ejemplo, este tipo de recursos permite transformar las relaciones utilizadas en los patrones (por ejemplo, la relación *Relación(en, ...)*) en el texto en lenguaje natural que la describe (por ejemplo, “a su paso por”).

5.6. Modelo de Inferencias

En esta sección se describe el modelo de inferencias que incluye las siguientes cinco etapas de razonamiento:

1. *Preprocesar-datos*
2. *Componer-descripción*
3. *Generar-hipótesis-de-descripción*
4. *Seleccionar-fuente-de-datos*
5. *Buscar-en-fuente-de-datos*

En los próximos apartados se describen los detalles de cada una de ellas.

5.6.1. Inferencia Preprocesar-datos

La inferencia *Preprocesar-datos* tiene como objetivo transformar los datos brutos de entrada en un conjunto de datos uniforme y válido para realizar el proceso de enriquecimiento (la especificación *CommonKADS* de la inferencia puede verse en los anexos). De forma general, la etapa de preproceso puede incluir acciones como las siguientes para uniformizar los datos de entrada:

1. Filtrar los datos
2. Preparar los datos
3. Validar los datos
4. Adaptar los datos

Por ejemplo, se pueden realizar acciones como: comprobar que los datos de medidas de sensores están en un intervalo válido, comprobar duplicidad de sensores, comprobar que los sensores sólo pertenecen a una subred excluyente, etc. En la Figura 5.9 se muestra el *mapping* a nivel de dominio de la inferencia. Como se puede observar, esta inferencia utiliza el conocimiento sobre clases de sensores como modelo de validación de los datos.

Figura 5.9: *Mapping* con el dominio de la inferencia *Preprocesar-datos*.

5.6.2. Inferencia Componer-descripción

La inferencia Componer-descripción tiene como fin integrar partes de la descripción lingüística, obtenidas por diferentes etapas de razonamiento previas. En concreto, integra el patrón lingüístico y los componentes de la misma (referencias geográficas). De la combinación se obtiene la descripción lingüística en el formato final.

Algoritmo 5.3 Inferencia *Componer-descripción*.

```

INFERENCE Componer-descripción ;
  ROLES:
    INPUT: Plan-de-la-descripción ;
    OUTPUT: Descripción-de-un-sensor ;
    STATIC: Modelo-léxico ;
  SPECIFICATION:
    "Combina los resultados obtenidos en inferencias
     previas obteniendo la descripción lingüística en el
     formato apropiado." ;
END INFERENCE Componer-descripción ;

KNOWLEDGE-ROLE Plan-de-la-descripción ;
  TYPE: DYNAMIC ;
  DOMAIN-MAPPING: Patrón-lingüístico ,
                  Referencias-espacio/temporales ;
END KNOWLEDGE-ROLE Plan-de-la-descripción ;

KNOWLEDGE-ROLE Descripción-de-un-sensor ;
  TYPE: DYNAMIC ;
  DOMAIN-MAPPING: Descripción-Lingüística ;
END KNOWLEDGE-ROLE Descripción-de-un-sensor ;

KNOWLEDGE-ROLE Modelo-léxico ;
  TYPE: STATIC ;
  DOMAIN-MAPPING: Recursos-léxico-lingüísticos ;
END KNOWLEDGE-ROLE Modelo-léxico ;

```

Como ejemplo de funcionamiento de la inferencia en el dominio hidrológico de la red de sensores SAIH, para un patrón como el siguiente:

$$\begin{aligned}
& TipoSensor(s, caudal), Latitud(s, x), Longitud(s, y) \\
& \rightarrow Patrón([Referencia(x_1, río, coordenadas(x, y)), \\
& \quad Referencia(x_2, población, coordenadas(x, y))], \\
& \quad [Relación([x_1, "en", x_2]])])
\end{aligned}$$

Donde el sensor s es de tipo *caudal* y dicho sensor tiene la característica *río* con valor x_1 y la característica *población* con valor x_2 , la descripción lingüística que se compondría para el sensor s sería la concatenación de las cadenas: x_1 , "en", x_2 . Supongamos que x_1 es *Río Ebro* y x_2 corresponde a *Reinosa*, la Plantilla contendría la lista ["*Río Ebro*", "en", "*Reinosa*"]. El resultado de la ejecución de la inferencia *Componer-descripción* sería la descripción textual "*Río Ebro en Reinosa*".

En los anexos se muestra la especificación textual *CommonKADS* para la inferencia *Componer-descripción*. En la Figura 5.10 se muestra el *mapping* a nivel de dominio de la inferencia *Componer-descripción*:

Figura 5.10: *Mapping* con el dominio de la inferencia *Componer-descripción*.

5.6.3. Inferencia Generar-hipótesis-de-descripción

El objetivo del paso de inferencia *Generar-hipótesis-de-descripción* es generar un patrón geoespacial que indica cómo se describe la información sobre el sensor. Para generar patrones de descripciones, esta tarea hace uso de la base de conocimiento de selección de patrones lingüísticos. El proceso de generación de patrones de descripciones para un sensor es un proceso iterativo. Así, el proceso genera inicialmente un patrón de descripción, el más prioritario de acuerdo con algún esquema de preferencia. A continuación, si se solicita

de nuevo, el proceso puede generar otro patrón siguiente en prioridad y así sucesivamente, hasta completar todos los patrones posibles.

En la Figura 5.11 se muestra el *mapping* a nivel de dominio de la inferencia *Generar-hipótesis-de-descripción*. Como se observa en dicha figura, el modelo de generación de hipótesis utilizado corresponde a la base de conocimiento de selección de patrones lingüísticos.

Figura 5.11: *Mapping* con el dominio de la inferencia *Generar-hipótesis-de-descripción*.

5.6.4. Inferencia Seleccionar-fuente-de-datos

El objetivo de la inferencia Seleccionar-fuente-de-datos es determinar qué fuentes de datos son apropiadas para encontrar ciertas características requeridas. En la Figura 5.12 se muestra el *mapping* a nivel de dominio de la inferencia *Seleccionar-fuente-de-datos*.

En la Figura 5.12 se aprecia que los criterios de selección de *Open Data* sirven de estrategia de búsqueda en la selección de fuentes de datos. Dichos criterios se encuentran contenidos en la base de conocimiento *Estrategia-de-búsqueda*,

que define en que bases externas consultar para construir la descripción.

Figura 5.12: *Mapping* con el dominio de la inferencia *Seleccionar-fuente-de-datos*.

5.6.5. Inferencia Buscar-en-fuente-de-datos

El objetivo del paso de inferencia *Buscar-en-fuente-de-datos* es acceder a fuentes abiertas de datos para obtener información correspondiente a las características C_i (*features*) de las referencias. Se trata, por ejemplo, de características geográficas tales como las siguientes $\{población, río, área-montañosa, lugar-histórico, embalse, \dots\}$.

Para buscar las características geográficas se hace uso de la base de conocimiento criterios de selección *Open Data*. El proceso de obtención de los valores de características es un proceso de tentativo de búsqueda dado que las fuentes abiertas de datos no garantizan a priori que los valores estarán disponibles, por lo que el resultado de la consulta de una característica en una determinada fuente de datos puede no obtener resultado. Por tanto, el proceso opera de la siguiente forma. Dada una determinada característica, se aplican las reglas de la base de conocimiento en orden hasta que se obtiene un valor correcto para dicha característica. Si no se encuentra, el proceso termina con fracaso.

Esta solución es general y escalable para disponer de diferentes tipos de métodos de búsqueda. Los métodos de acceso y su selección quedan encapsulados en el paso de inferencia. La incorporación de nuevos métodos supone un esfuerzo bajo que requiere sólo la extensión de dicho paso de inferencia con las correspondientes reglas y funciones de acceso. En la Figura 5.13 se muestra el *mapping* a nivel de dominio de la inferencia *Buscar-en-fuente-de-datos*:

Figura 5.13: *Mapping* con el dominio de la inferencia *Buscar-en-fuente-de-datos*.

5.7. Ejemplo de razonamiento

En esta sección se muestra un ejemplo de razonamiento, que tiene como fin complementar la descripción general del método con un caso ilustrativo. La Figura 5.14 muestra la estructura de inferencia completa del método, con todas las inferencias y bases de conocimiento.

El ejemplo considerado pertenece al ámbito de la hidrología, más concretamente a la red de sensores del SAIH. Los datos de entrada al método contienen las coordenadas de los sensores así como el tipo de medida que registran, entre otra información adicional. Estos datos son los que se han definido como *Datos-en-bruto*.

Figura 5.14: Visión completa del método.

El primer paso del método consiste en la ejecución de la inferencia *Preprocesar-datos* sobre los mencionados *Datos-en-bruto*. Para ello se hace uso de la base de conocimiento *Modelo-de-validación*, en la que se encuentran definidas las acciones encargadas de adecuar los datos de entrada (filtrar , preparar, validar y adaptar los datos). De acuerdo con el *mapping* del dominio de inferencia *Preprocesar-datos*, se utiliza la información estática de entrada *Clases-de-*

sensores para tal fin. Como resultado de tal ejecución se obtienen los *Datos-corregidos*.

Una vez se tienen los datos en un formato apropiado para trabajar con ellos, comienza un proceso que se realiza para cada uno de los sensores de forma individual. Para ello se selecciona los *Datos-de-un-sensor* del primer sensor de los *Datos-corregidos*. Una vez se haya completado el proceso de este sensor, se pasará al segundo y así hasta terminar con todos.

Considérese que la información de un determinado sensor seleccionado (el sensor s28) es la siguiente (*Datos-de-un-sensor*):

- *TipoSensor*(s28, caudal)
- *Latitud*(s28, 42.34683990478520)
- *Longitud*(s28, -7.87428712844849)

A continuación se realiza el *paso Generar-hipótesis-de-descripción*. Dados los *Datos-de-un-sensor* y el conocimiento de la base *Modelo-de-generación*, se obtiene la *Estructura-de-la-hipótesis* y los *Componentes-requeridos-de-la-hipótesis*. De acuerdo con el *mapping* del dominio de inferencia *Generar-hipótesis-de-descripción* (ver Figura 5.11), se utiliza la información estática de entrada *Selección-de-patrones-lingüísticos* para tal fin. En concreto la regla que se utiliza en el caso del sensor considerado es la siguiente:

TipoSensor(s, caudal),
Latitud(s, x),
Longitud(s, y)
 $\rightarrow$ *Patrón*([*Referencia*(x₁, río, coordenadas(x, y)),
Referencia(x₂, población, coordenadas(x, y))],
[*Relación*(en, x₁, x₂)])

Por aplicación de dicha regla se obtiene el siguiente patrón lingüístico como *Estructura-de-la-hipótesis*:

Patrón(
[*Referencia*(x₁, río, coordenadas(-7,87428712844849, 42,34683990478520)),
Referencia(x₂, población, coordenadas(-7,87428712844849, 42,34683990478520))],
[*Relación*(en, x₁, x₂)])

Se obtienen también los *Componentes-requeridos-de-la-hipótesis*, que en este caso corresponden con las siguiente Características:

[*Referencia*(x_1 , río, *coordenadas*(−7,87428712844849, 42,34683990478520)),
Referencia(x_2 , población, *coordenadas*(−7,87428712844849, 42,34683990478520))]

El siguiente paso es la inferencia *Seleccionar-fuente-de-datos*, cuya entrada dinámica es la lista de *Componentes-requeridos-de-la-hipótesis* y con la información de la base de conocimiento *Estrategia-de-búsqueda* consistente en los *Criterios-de-selección-Open-Data*. La inferencia se ejecutará tantas veces como *Componentes-requeridos-de-la-hipótesis* sean requeridos, en este caso dos veces. Así mismo, para cada componente se ejecutará la inferencia *Buscar-en-fuente-de-datos* tantas veces como sea necesario hasta obtener un valor *Éxito* en la salida *Éxito-de-la-búsqueda*:

1. Para la *Característica* río se ejecuta la inferencia *Seleccionar-fuente-de-datos*, aplicando la siguiente regla:

Característica(río, *Coordenadas*(x , y)) →
FunciónDeConsulta(*OpenStreetMap*, *Coordenadas*(x , y),
 RIVER, STREAM, CANAL, DRAIN, WATER, RIVERBANK)

y obteniendo la siguiente *Consulta-en-fuente-de-datos*:

FunciónDeConsulta(*OpenStreetMap*,
Coordenadas(−7,87428712844849, 42,34683990478520),
 RIVER, STREAM, CANAL, DRAIN, WATER, RIVERBANK)

Con dicha consulta se procede a ejecutar la inferencia *Buscar-en-fuente-de-datos*, en *Open Street Maps* obteniendo resultados válidos, lo que generaría como salida de la inferencia *Éxito-de-la-búsqueda* el valor *Éxito*, así como la cadena de texto “río Miño” como valor para el *Componente*.

2. En el caso de población se ejecuta la inferencia *Seleccionar-fuente-de-datos* que aplica la siguiente regla:

Característica(población, *Coordenadas*(x , y)) $\rightarrow$
FunciónDeConsulta(*OpenStreetMap*, *Coordenadas*(x , y),
POPULATION, *CITY*, *TOWN*, *VILLAGE*, *MUNICIPALITY*)

obteniéndose la siguiente *Consulta-en-fuente-de-datos*:

FunciónDeConsulta(*OpenStreetMap*,
Coordenadas($-7,87428712844849$, $42,34683990478520$),
POPULATION, *CITY*, *TOWN*, *VILLAGE*, *MUNICIPALITY*)

Con dicha consulta se procede a ejecutar la inferencia *Buscar-en-fuente-de-datos*, en *Open Street Maps* la *Característica-espacio/tiempo PopulatedPlace*. En este caso se asume que no se obtienen resultados válidos, lo que generaría como salida de la inferencia *Éxito-de-la-búsqueda* el valor *Fracaso*. Al haber fallado en la búsqueda, se recurre a la regla:

Característica(población, *Coordenadas*(x , y)) $\rightarrow$
FunciónDeConsulta(*GeoNames*, *Coordenadas*(x , y),
ADMINISTRATIVE)

y se obtiene:

FunciónDeConsulta(*GeoNames*,
Coordenadas($-7,87428712844849$, $42,34683990478520$),
ADMINISTRATIVE)

Con dicha consulta se procede a ejecutar la inferencia *Buscar-en-fuente-de-datos*, en *GeoNames* la *Característica-espacio/tiempo PopulatedPlace* y se obtienen resultados válidos, lo que generaría como salida de la inferencia *Éxito-de-la-búsqueda* el valor *Éxito*, así como la cadena de texto “*Orense*” como valor para el *Componente*.

Podría darse el caso de que la búsqueda hubiera fallado nuevamente, continuando con las siguientes *Consulta-en-fuente-de-datos* hasta lograr el éxito. En caso de no alcanzarlo tras terminar con todas las consultas, se abortaría el proceso actual y se debería ejecutar de nuevo la inferencia *Generar-hipótesis-de-descripción* para utilizar otra *Estructura-de-la-hipótesis* con sus *Componentes-requeridos-de-la-hipótesis*, en caso

de haberla. En caso de no estar disponible dicha alternativa, no se construiría descripción para el sensor al ser imposible conseguir construir su descripción.

Una vez se tiene el *Plan-de-la-descripción*, el siguiente paso consiste en la ejecución de la inferencia *Componer-descripción*, cuya entrada dinámica es dicho *Plan-de-la-descripción*, consistente en la agrupación de la *Estructura-de-la-hipótesis* y de los *Componentes-encontrados* junto con la información de la base de conocimiento *Modelo-léxico* consistente en los *Recursos-léxico-lingüísticos*. En este caso, el recurso apropiado permite transformar la relación utilizada en el patrón de la *Estructura-de-la-hipótesis* (la relación *Relacion(en, x, y)* en el texto en lenguaje natural que la describe “*a su paso por*”).

Dado que se tienen los componentes $\{x : \text{río Miño}\}$ e $\{y : \text{Orense}\}$ junto con la relación *Relacion(en, x, y)*, el resultado de la ejecución de la inferencia *Componer-descripción* (que en este caso consistirá en concatenar los componentes con el recurso “*a su paso por*”) será la *Descripción-de-un-sensor* “*río Miño a su paso por Orense*” para este sensor.

Este proceso se repite para todos los sensores. Como último paso, se agrupan todas las descripciones obteniendo las Descripciones como resultado final y dando por terminada la ejecución.

La Figura 5.15 muestra un resumen gráfico de la ejecución del método correspondiente a este ejemplo. La ejecución se desarrolla en forma de árbol. En dicho árbol se aprecian los diferentes procesos del método y los intercambios de información que suceden entre ellos. De la misma forma, se puede observar el uso de las bases de conocimiento.

Figura 5.15: Ejemplo de generación de descripción.

6. Evaluación

En el presente capítulo se describe la evaluación de la propuesta realizada en este trabajo. La evaluación se centró en analizar tres características:

1. La *aplicabilidad* del método teórico general a un caso práctico de dimensión real con uso de bases de datos externas.
2. El *rendimiento* del método en la resolución del problema de generación de descripciones para enriquecimiento de datos sobre redes de sensores.
3. La *portabilidad* del método para ser aplicable a diferentes problemas y redes de sensores.

La evaluación se ha llevado a cabo, en primer lugar, con dos experimentos prácticos haciendo uso de una red de sensores (la red hidrológica SAIH):

1. *Experimento 1: Uso de una base de datos externa.* El objetivo de esta evaluación fue analizar la aplicabilidad y rendimiento del método a una red de sensores con una muestra de tamaño limitado. En este caso se utilizó únicamente una base de datos externa (*Geonames*). El experimento se realizó con el fin de llevar a cabo un primer análisis de la eficacia del método y dirigir un posterior planteamiento de una solución más completa.
2. *Experimento 2: Combinación de varias bases de datos externas.* Este experimento se realizó sobre una red de dimensión real y con uso de múltiples bases de datos externas. Este análisis permitió realizar un análisis más completo del rendimiento del método y de su aplicabilidad a una red de tamaño real.

Además de lo anterior, se realizó un análisis de la aplicabilidad del método a la generación de expresiones con referencias temporales con el fin de extraer conclusiones respecto a la generalidad y portabilidad del método propuesto. En los próximos apartados se detallan la forma en que se realizó este análisis y los resultados obtenidos en cada caso.

6.1. Experimento 1. Uso de una base de datos externa

En el experimento 1, el problema considerado ha sido generar descripciones textuales de datos sobre una red de sensores hidrológica, haciendo uso de una única base de datos geográfica externa [Molina and Sanchez-Soriano, 2011]. En el experimento se analizaron diferentes formas de uso de la base de datos considerada (*Geonames*), con el fin de enriquecer datos con descripciones correctas adaptadas a las características de la red hidrológica.

El fin último de estas descripciones es ayudar a construir descripciones textuales más complejas como parte de un sistema que explica diferentes situaciones hidrológicas en combinación con elementos gráficos (mapas, gráficos, animaciones, etc.) [Molina and Sanchez-Soriano, 2010].

En este experimento se ha utilizado la red de sensores medioambientales del SAIH que corresponde a un dominio real y con suficiente complejidad para la construcción de modelos y experimentación, que permite extraer conclusiones sobre la validez y utilidad práctica de las soluciones planteadas. Resulta adecuado por las siguientes razones:

- *Disponibilidad de medidas.* Actualmente existe un gran número de infraestructuras de medición de datos sobre comportamiento en tiempo real de áreas medioambientales. Ejemplos de estas infraestructuras son los sistemas información hidrológica y de calidad de las aguas que se han instalado en España principalmente a lo largo de las últimas décadas en las diferentes cuencas hidrográficas.
- *Diversidad de medidas.* La información registrada por este tipo de sistemas da lugar a volúmenes importantes de datos correspondientes a medidas cuantitativas de diferente naturaleza. Por ejemplo, datos sobre lluvia, caudal, nivel, presencia de contaminantes, grados de apertura de compuertas de embalses, datos meteorológicos (dirección del viento, temperatura, etc.). Además de los diferentes tipos de medidas, la información tiene asociación espacial y temporal.
- *Apoyo a toma de decisiones.* Los datos medidos sirven de base para la toma de decisiones según diferentes objetivos (previsión de avenidas, mejora de la gestión hídrica en escenarios de sequía, actuación ante la

presencia de contaminantes, actuación ante fallos de medición en los equipos, etc.) lo cual permite experimentar con diferentes modelos.

6.1.1. Aplicación del método

El problema considerado en este experimento es la generación de descripciones de texto para los sensores de la red hidrológica SAIH. Ejemplos de descripciones son: *el embalse de Ribarroja, Albalat, el río Ebro en Ascó* o para un conjunto de sensores, tales como *la mayoría de los embalses en la cuenca del Júcar*. Para ello se cuenta con la siguiente información de entrada para cada sensor:

1. El tipo de medida del sensor. Un valor de:
 - a) caudal
 - b) lluvia
 - c) nivel
 - d) volumen
2. Las coordenadas geográficas de la ubicación del sensor:
 - a) latitud
 - b) longitud

Para resolver el problema se utilizaron dos fuentes propietarias (no abiertas) de información geográfica:

1. *Archivos de datos SIG*. Ficheros de información hidrológica, para la representación de los ríos y de las cuencas hidrográficas españolas.
2. *Identificadores SAIH de sensores*. Corresponde a los códigos alfanuméricos asignados por expertos del SAIH. Normalmente son códigos técnicos por lo que se entienden únicamente por expertos (por ejemplo, E69_ZAHARA o M06_STE_ARR_VIL). Puesto que hay nueve centros de control SAIH independientes en España, cada centro de control utiliza diferentes criterios para identificar los sensores. Algunos identificadores incluyen lugares geográficos, por lo que esta información es potencialmente útil para el método.

Adicionalmente, se utilizó como base de datos externa el servidor de nombres *Geonames* cuyo funcionamiento se ha descrito en el capítulo sobre antecedentes, sobre el que se ensayaron diferentes formas de acceso.

El método general se aplicó a este problema de la siguiente manera. Para un sensor o grupo de sensores, se obtienen los valores de las características geográficas siguientes:

- *Río*. Si el sensor es de caudal, se determina el río en el que se ubica. Para ello, se utiliza un archivo de datos SIG vectorial con las líneas correspondientes a los segmentos de río y se aplica un procedimiento espacial para determinar en qué río encuentra cada sensor. El procedimiento encuentra el segmento de línea más cercano para cada ubicación de sensor devolviendo el nombre del río asociado a dicha línea.
- *Cuenca hidrográfica*. Se utiliza un archivo de datos SIG vectorial para buscar la cuenca hidrográfica, definidas como regiones espaciales. Se utiliza un procedimiento espacial sencillo para determinar la región espacial donde se encuentran las coordenadas de los sensores y devuelve el nombre de dicha región como el nombre de la cuenca.
- *Lugar*. Siguiendo la estrategia de la propuesta del método general, se utiliza el servidor *Geonames* para encontrar la referencia geográfica correspondiente a un lugar cercano a las coordenadas geográficas del sensor. No obstante, si puede extraerse a partir del identificador SAIH, se prefiere esta opción antes que el nombre proporcionado por *Geonames*, al suponer que los expertos SAIH proporcionan mejores nombres para los lugares en los que se ubican los sensores. El lugar se extrae del identificador SAIH (en los casos en que está presente) utilizando un procedimiento heurístico guiado por una base de conocimiento con patrones de texto.
- *Cuantificación*. La referencia geográfica para conjuntos de sensores se construye utilizando el área más específica que abarca a todos los sensores (río, cuenca o nación). La descripción se cuantifica con tres valores { *algunos*, *la mayoría*, *todos* } para expresar algunos (número de sensores por debajo del 50 % de la zona), la mayoría (mayor que el 50 %, pero inferior al 100 %), y todos (100 %).

<i>Sensor</i>	<i>Plantilla</i>	<i>Ejemplo de descripción</i>
<i>sensor de volumen/ nivel en embalse</i>	<i>embalse de <lugar></i>	<i>Embalse de Ribarroja</i>
<i>sensor de caudal/ nivel en río</i>	<i><río> en <lugar></i>	<i>Río Guadalquivir en Andújar</i>
<i>pluviómetro</i>	<i><lugar></i>	<i>Albalat</i>
<i>cjto. sensores volumen/ nivel en una cuenca</i>	<i><cuantificación> embalses en <cuenca></i>	<i>La mayoría de embalses en la cuenca del Júcar</i>
<i>cjto. de sensores caudal/ nivel en un río</i>	<i><cuantificación> secciones en <río></i>	<i>Algunas secciones en el río Ebro</i>
<i>cjto. de sensores caudal/ nivel en una cuenca</i>	<i><cuantificación> ríos en <cuenca></i>	<i>La mayoría de los ríos de la cuenca del Júcar</i>
<i>cjto. de pluviómetros en una cuenca</i>	<i><cuantificación> puntos en <cuenca></i>	<i>Algunos puntos en la cuenca del Júcar</i>

Tabla 6.1: Patrones lingüísticos para construir descripciones de sensores.

De acuerdo con el método general, las descripciones se construyen haciendo uso de patrones lingüísticos (en este caso en forma de plantillas). Pueden verse ejemplos de plantillas en la Tabla 6.1.

La solución aplicada en esta versión del método se basa en aprovechar los códigos que los expertos usan para identificar los sensores SAIH para identificar referencias geográficas. Ejemplos de dichos códigos son: E69_ZAHARA, AFORO RÍO GUADALHORCE (CÁRTAMA), M06_STE_ARR_VIL. Cuando no es posible obtener las referencias de los códigos de identificación, se puede hacer uso de otra fuente (como es *Geonames*) para intentar obtener las referencias.

No existe un criterio uniforme para escribir los identificadores y, normalmente, estos criterios son diferentes en cada centro de control SAIH (hay nueve centros de control en España).

La solución empleada consiste en utilizar un método heurístico para el procesamiento de texto. El procesado está basado en la coincidencia de patrones que extrae los nombres de los lugares de estos identificadores utilizando una base de conocimiento con los patrones de texto que representan los criterios seguidos por los expertos en los nueve centros de control. Se ha construido esta base de conocimiento mediante la observación de las reglas sintácticas seguidas por los expertos.

Por ejemplo, los expertos del centro de control del SAIH de la cuenca del Guadalquivir utilizan el identificador siguiente para un sensor: E69_ZAHARA. Este sensor corresponde a una estación de volumen situado en un embalse. La estrategia para los identificadores seguida por los expertos para estos sensores es escribir un código alfanumérico (E69), el subrayado de caracteres (" _") y, a continuación, el nombre del lugar (Zahara). Por lo tanto, el patrón en nuestra base de conocimientos incluye los siguientes criterios:

1. El tipo de sensor es una estación de volumen
2. El centro de control es el del Guadalquivir
3. El lugar geográfico es el texto después de guion bajo (_).

Otro ejemplo es el identificador del sensor AFORO RÍO GUADALHORCE (CÁRTAMA) desde el centro de control del SAIH de las cuencas del sur de España. Este nombre corresponde a una estación de caudal e incluye el nombre del río (Guadalhorce) y un lugar poblado (Cártama) entre paréntesis. Este criterio se utiliza para todas las estaciones de caudal en las cuencas del sur de España. Por lo tanto, el patrón de conocimiento incluye:

1. El tipo de sensor es una estación de flujo
2. El centro de control es el sur de España
3. El lugar geográfico es el texto entre paréntesis

Se determinó que con 19 era posible modelar la mayoría de las estrategias seguidas por los expertos en los nueve centros de control SAIH cuando se utiliza el nombre geográfico. Además de este procedimiento, se realizó un procedimiento de procesamiento de texto para generar el nombre final del lugar. Por ejemplo, el procedimiento cambia abreviaturas por palabras utilizando un diccionario con abreviaturas típicas (Fte. = Fuente, Sta. = Santa, Pte. = Puente, Af. = Aforo, etc.).

Esta solución funciona correctamente para la mayoría de identificadores SAIH. Sin embargo, hay casos en que no es posible extraer el nombre del identificador. Esto es, por ejemplo, porque los expertos utilizan sólo códigos alfanuméricos (por ejemplo, M06_STE_ARR_VIL en la cuenca del Guadalquivir

y C001L85PQUIN en la cuenca del Ebro), abreviaturas inusuales (AR31 G.MINCHONES VILL.V en la cuenca del Tajo), u otras excepciones a las reglas generales.

Cuando no es posible encajar ningún patrón con el identificador de SAIH, se utiliza el servidor de *Geonames* y las coordenadas geográficas del sensor. Se pide al servidor *Geonames* el nombre del lugar más cercano a la ubicación definida por las coordenadas geográficas. Por ejemplo, se ha elegido el nombre de lugar *Andújar* (un lugar poblado cerca de la ubicación del sensor) utilizando *Geonames* para un sensor con el identificador SAIH M10_GLQUVIR_AND en la cuenca del Guadalquivir.

En resumen, la aplicación del método general a este problema tiene las siguientes características:

- Se utiliza un conjunto de características geográficas como referencias de las descripciones lingüísticas (como son río, cuenca hidrográfica y lugar geográfico)
- Se utiliza como fuente externa la base de datos *Geonames*.
- Se manejan patrones lingüísticos en forma de plantillas de texto con espacios para las características geográfica.
- El procedimiento sigue una estrategia de generación y prueba aunque, en este caso, maneja una base de conocimiento propia para tratar de resolver las búsquedas localmente como paso previo antes de realizar búsquedas en *Geonames*.

El desarrollo de esta versión del método incluyó la validación y refino de la base de conocimiento propia observado al azar descripciones generadas durante las sesiones de trabajo en la aplicación Web VSAIH (usaba dichas descripciones) [Molina and Sanchez-Soriano, 2011, Molina and Sanchez-Soriano, 2010]. Esta forma de validación y puesta a punto permitió detectar errores (por ejemplo, nombres con extensión mayor de lo deseable, mayúsculas equivocadas o nombres mal expresados). Para subsanar estos errores, se calibraron los procedimientos de razonamiento espacial (por ejemplo, distancia mínima para ríos o la distancia máxima para lugares poblados) y se modificaron y ampliaron los contenidos de la base de conocimiento de patrones de texto.

6.1.2. Procedimiento de evaluación

Para la evaluación de la calidad del método en este experimento se aplicaron dos tipos de análisis:

1. *Evaluación de la calidad las descripciones lingüísticas generadas.*
2. *Evaluación de la calidad de lugares geográficos generados.*

Para evaluar la calidad de las descripciones lingüísticas generadas se obtuvo una muestra aleatoria de noventa descripciones que habían sido generadas automáticamente, diez para cada una de las nueve cuencas hidrográficas del SAIH. Para considerar una descripción correcta, se verificaron manualmente los siguientes requisitos para cada una de las descripciones:

- No hay errores sintácticos. Se consideran errores de este tipo:
 - Longitud del texto excesiva
 - Mayúsculas incorrectas
 - Abreviaturas
- La descripción es fácil de leer (está bien expresada)
- El lugar geográfico y el tipo de medida (lluvia, cauda, etc.) son correctos

Para evaluar la calidad del lugar geográfico generado se evaluó si el método selecciona nombres apropiados para los lugares geográficos a los que referencia en las descripciones. Se seleccionaron aleatoriamente cien descripciones de sensores que habían sido escritas por expertos hidrólogos del SAIH. Para cada sensor se comparó la descripción del experto con las descripciones generadas con diferentes estrategias:

1. *Geonames spot*. Se genera la descripción utilizando la característica geográfica *spot* de *Geonames* (Pueden ser puntos de referencia como monumentos históricos, centros comerciales, etc.)
2. *Geonames hydrographic*. Se genera la descripción utilizando la característica hidrográfica proporcionada por *Geonames*
3. *Geonames populated place*. Se genera la descripción utilizando la población más cercana a proporcionada por *Geonames*

4. *Identificadores SAIH*. Se utiliza la base de conocimiento de patrones de texto para extraer el lugar geográfico de los identificadores SAIH
5. *Identificadores SAIH. + Geonames populated place*. Combinación de las versiones 3 y 4 del algoritmo

Se considera que un resultado es correcto cuando la descripción del experto incluye la misma referencia que la descripción generada por el algoritmo (basándose en la suposición de que experto selecciona la referencia geográfica más apropiada en cada caso).

6.1.3. Resultados de la evaluación

Los resultados de la evaluación fueron los siguientes. En la evaluación de la calidad de las descripciones lingüísticas se obtuvo como resultado que: 85 de 90 descripciones fueron correctas (94,4 %) lo que supone una alta tasa de éxito. Para los casos erróneos (5,6 %), se detectaron los siguientes problemas:

1. Abreviaturas inusuales no incluidas en el modelo
2. Mayúsculas equivocadas (*río Turia O Guadalaviar en Tramacastilla*)
3. Descripción se ajusta a un patrón pero su información no es apropiada
4. Repeticiones de determinantes (*Embalse de del Limonero*)
5. Problemas con números romanos

<i>Evaluación</i>	<i>Algoritmo</i>	<i>Resultados correctos</i>
1	<i>Geonames spot</i>	7 %
2	<i>Geonames hydrographic</i>	8 %
3	<i>Geonames populated place</i>	58 %
4	Identificadores SAIH	75 %
5	Identificadores SAIH + <i>Geonames populated place</i>	89 %

Tabla 6.2: Resultados de la evaluación de la calidad del lugar geográfico generado con diferentes versiones del algoritmo.

En la Tabla 6.2 se muestran los resultados de la evaluación de la calidad del lugar geográfico generado haciendo uso de diferentes versiones del algoritmo.

Cuando se utiliza *Geonames spot* y *Geonames hydrographic* se obtienen valores bajos de acierto: 7% y 8% respectivamente (también se evaluaron las características proporcionadas por *Geonames* para vegetación, carreteras y ferrocarriles pero se obtuvo 0% en dichos casos).

Al usar *Geonames populated place* obtiene un valor del 58%. Los resultados también muestran que el algoritmo basado sólo en *identificadores SAIH* obtiene valores altos de acierto, con el 75%. El mejor valor (89%) se obtiene cuando se combina la extracción de lugares de los identificadores SAIH en combinación con *GeoNames populated place*.

Los resultados confirman que el uso de *Geonames* es útil como fuente de información externa para generar descripciones lingüísticas pero, en este caso, necesita ser combinado con otras fuentes (por ejemplo la base de conocimiento local con los identificadores SAIH) para aumentar la calidad de las descripciones generadas.

6.1.4. Conclusiones del experimento 1

Las conclusiones que se han obtenido de la evaluación del presente experimento han sido las siguientes:

- El método general es aplicable a este caso. Tal como se plantea en el método general, se maneja (1) una estrategia de generación y prueba para búsqueda de referencias geográficas (en este caso con dos fuentes principales de información, una local y otra externa) y (2) un conjunto de patrones lingüísticos con criterios de selección como soporte a la generación de descripciones (en este caso, los patrones se representan en forma de plantillas de texto).
- Se logra un rendimiento satisfactorio con el uso de una base de datos geográfica externa (como *Geonames*) combinado con una base de conocimiento propia. La evaluación se hizo manualmente con ayuda dos muestras de tamaño limitado (de 90 y 100 elementos).
- El manejo de la base de conocimiento propia resta generalidad y flexibilidad para su uso con otras redes. Dicha base ayuda a extraer términos por medio de patrones de texto y está adaptada a las características

propias de la red hidrológica SAIH y la forma en que los expertos construyen las descripciones.

6.2. Experimento 2. Uso de varias fuentes externas

El objetivo del segundo experimento fue construir descripciones textuales que expresen en lenguaje natural la ubicación geográfica de sensores hidrológicos (al igual que en el experimento 1) aunque, esta vez, con una red de sensores de mayor complejidad y sustituyendo el uso de la base de conocimiento local (que resta generalidad a la solución) por el acceso a múltiples bases de datos externas [Molina et al., 2015]. En este caso, también se aplicó a la red hidrológica de sensores SAIH.

6.2.1. Aplicación del método

En la Figura 6.1 puede verse un esquema del método aplicado a este caso concreto en el dominio geográfico de la red de sensores del SAIH.

Figura 6.1: Método aplicado al caso geográfico de la red SAIH.

La solución que se plantea a este problema hace uso de una base de datos SAIH que reúne información sobre la red de sensores entre la que se encuentran las coordenadas geográficas de los sensores, mediciones, identificadores, descripciones literales, tipo, unidades de medida, intervalos de medida, tolerancias, rangos de fiabilidad, etc.

Además se utilizan las siguientes cuatro bases de datos externas: *Geonames*, *Open Street Map*, *DBPedia* y *Panoramio* (descritas en el capítulo de antecedentes).

En este experimento, se aplica el método general considerando como entrada las coordenadas geográficas de cada sensor así como el tipo de magnitud física medida por el sensor (por ejemplo, nivel, lluvia, caudal, etc.). Como resultado, se generan descripciones lingüísticas sobre las ubicaciones geográficas de los sensores en lenguaje natural como salida, como por ejemplo: *el río Ebro en Ascó* o *el embalse de Ribarroja*.

Para representar patrones lingüísticos, se utiliza una particularización de la representación considerada por el método general, mediante plantillas de texto. Dichas plantillas definen los patrones de descripción en lenguaje natural y se seleccionan de acuerdo a las estrategias específicas formuladas mediante un conjunto de reglas de producción. El formato general de dichas reglas es el siguiente:

$$SensorType(x, y) \& C1 \& \dots \& Cn \rightarrow Template(x, y)$$

En el lado izquierdo las reglas incluyen el predicado $SensorType(x, y)$, es decir, el tipo del sensor x es y junto con un conjunto de condiciones C_i . Para formular las condiciones se utiliza un predicado lógico $Feature(x, y, z)$ donde para el sensor x , la característica geográfica y tiene el valor z . También se utilizan operadores de comparación (igual que, mayor que, menor que, etc.) para establecer condiciones para los valores de las características. En el lado derecho las reglas incluyen el predicado lógico $Template(x, y)$, cuyo significado es que la descripción textual del sensor x es y . La representación de y es una lista de cadenas de texto (por ejemplo ["río Ebro", "en", "Ascó"]. La

descripción de texto consiste en la concatenación de estas cadenas.

Para el caso particular correspondiente a este experimento, fue suficiente con 20 reglas con sus plantillas de texto. Dichas reglas, se obtuvieron de forma manual mediante el análisis de ejemplos de descripciones escritas por expertos hidrólogos. El tiempo empleado en la obtención de dichas plantillas apenas supuso un par de horas de trabajo.

Por ejemplo, una de las estrategias seguidas por los expertos es la siguiente: Si el sensor mide el caudal del río, la descripción debería hacer referencia al nombre del río pero si el sensor mide la lluvia, la referencia geográfica no debería incluir el nombre del río. Ejemplo de regla:

$$\begin{aligned} & \text{SensorType}(s, \text{WaterFlow}) \ \& \ \text{Feature}(s, \text{Río}, x) \ \& \\ & \text{Feature}(s, \text{PopulatedPlace}, y) \rightarrow \text{Template}(s, [x, \text{"en"}, y]) \end{aligned}$$

De acuerdo con la regla, si el sensor s es de tipo *WaterFlow* y dicho sensor tiene la característica *Río* con valor x y la característica *PopulatedPlace* con valor y , a continuación, la descripción de texto para el sensor s es la concatenación de las cadenas: x , "en", y (por ejemplo, la lista ["*Río Ebro*", "en", "*Ascó*"]). El siguiente conjunto de reglas permite obtener la descripción textual correspondiente a un pluviómetro:

$$\begin{aligned} & \text{SensorType}(s, \text{Pluviómetro}) \ \& \ \text{Feature}(s, \text{Pantano}, x) \rightarrow \text{Template}(s, [x]) \\ & \text{SensorType}(s, \text{Pluviómetro}) \ \& \ \text{Feature}(s, \text{HistoricPlace}, x) \rightarrow \text{Template}(s, [x]) \\ & \text{SensorType}(s, \text{Pluviómetro}) \ \& \ \text{Feature}(s, \text{Bosque}, x) \rightarrow \text{Template}(s, [x]) \\ & \text{SensorType}(s, \text{Pluviómetro}) \ \& \ \text{Feature}(s, \text{MountainArea}, x) \rightarrow \text{Template}(s, [x]) \\ & \text{SensorType}(s, \text{Pluviómetro}) \ \& \ \text{Feature}(s, \text{Municipio}, x) \rightarrow \text{Template}(s, [x]) \\ & \text{SensorType}(s, \text{Pluviómetro}) \ \& \ \text{Feature}(s, \text{PopulatedPlace}, x) \rightarrow \text{Template}(s, [x]) \end{aligned}$$

Dichas reglas ayudan a identificar la referencia geográfica más adecuada para la ubicación de un pluviómetro y expresan que la descripción de un pluviómetro se genera de la siguiente manera:

1. Si hay un pantano en la ubicación del sensor se usa primero el nombre de dicho pantano (función *Pantano*),
2. No hay pantano pero sí un lugar histórico en la ubicación del sensor. Se usa el nombre del lugar histórico (función *HistoricPlace*), etc.

Las reglas se procesan en orden como estrategia de control. Para un determinado tipo de sensor, en el caso de contener todos los atributos así como ajustarse a los atributos necesarios para pertenecer a una categoría de descripción, se podrá construir la descripción para dicho sensor. En el caso de los sensores de lluvia la clasificación se realizó de la siguiente forma:

- Lluvia para usuarios generales. Búsqueda en este orden:
 1. Embalse.
 2. Lugar histórico.
 3. Área boscosa.
 4. Área montañosa.
 5. Municipio.
 6. Pueblo.
 7. Ciudad.
 8. Caso por defecto: Área poblada de *Geonames*.
- Lluvia para usuarios regionales. Búsqueda en este orden:
 1. Pozo de agua.
 2. Fuente de agua.
- Lluvia para usuarios Hidrólogos.
 1. Construcción hidráulica
 2. Caudal de agua.
 3. Embalses.

Las características geográficas se obtienen dinámicamente a partir de las fuentes de datos geográficos externas, de la siguiente forma:

- *Open Street Map*: Permite obtener un conjunto de características de una ubicación en forma de etiquetas. Con este fin, se crea una envoltura rectangular alrededor de la ubicación geográfica del sensor para formular una consulta al API REST de *Open Street Map*. Como resultado de dicha consulta se obtienen todas las etiquetas de la citada

región para a continuación el método clasifique el accidente geográfico requerido (por ejemplo, *River*) utilizando para ello las etiquetas de la Tabla 6.3 (por ejemplo, río, arroyo, canal, etc.). Los límites de la región de búsqueda se definen por: longitud mínima x_1 , latitud mínima y_1 , longitud máxima x_2 y la latitud máxima y_2 . Para las coordenadas geográficas del sensor x (longitud), y (latitud), los parámetros del contorno se obtienen de la siguiente manera, donde k_1 es un parámetro ajustable:

$$x_1 = x - k_1, \quad y_1 = y - k_1, \quad x_2 = x + k_1, \quad y_2 = y + k_1$$

Para las pruebas realizadas en el presente trabajo se utilizó $k_1 = 0.01^\circ$ (grados) para pluviómetros y sensores de caudal y $k_1 = 0.018^\circ$ para nivel de embalses. Los valores k_1 se estimaron experimentalmente de acuerdo con los objetivos siguientes:

1. Obtener buenas referencias geográficas teniendo en cuenta las distancias aceptables entre el punto geográfico y referencias espaciales cercanas (por ejemplo, lugares poblados) así como el tamaño de ciertas características geográficas (por ejemplo, embalses)
2. Mantener el tamaño de la región de búsqueda lo más reducido posible para evitar una excesiva cantidad de información devuelta por *Open Street Map*.

Se realiza la consulta a OSM tras construir la URL con las coordenadas de la envoltura, de acuerdo con la especificación del API de OSM, obteniendo un archivo XML con toda la información geográfica almacenada en los servidores de OSM para la región solicitada. Esta información geográfica se proporciona de acuerdo a la especificación de OSM: nodos (*node*), formas (*way*) y relaciones (*relation*). Los resultados de la consulta se filtran analizando los elementos del fichero XML y, dependiendo del tipo al que pertenecen en OSM (*Node*, *Way* o *Relation*) se obtienen las características geográficas de cada uno. Se han tenido en cuenta 8 categorías (y subcategorías asociadas a dichas categorías) para atributos de OSM (ver Tabla 6.3)

<i>Categoría</i>	<i>Sub-categoría</i>
<i>WATERWAY</i>	<i>DAM, STREAM, RIVER, CANAL, DRAIN, RIVERBANK</i>
<i>HISTORIC</i>	
<i>NATURAL</i>	<i>PEAK, HEATH, WOOD, SCRUB, WATER, SPRING</i>
<i>MAN_MADE</i>	<i>WASTEWATERPLANT, WATERWORKS, RESERVOIRCOVERED, WATERTOWER, WATERWELL, DRINKINGWATER</i>
<i>LANDUSE</i>	<i>FOREST, RESERVOIR</i>
<i>IS_IN</i>	<i>COUNTRY, STATE, REGION, CITY, TOWN, VILLAGE, MUNICIPALITY</i>
<i>NAME</i>	
<i>TYPE</i>	
<i>POPULATION</i>	

Tabla 6.3: Categorías de atributos de OSM.

- *Geonames*: Se realizan búsquedas en el servicio de Geonames para obtener información geográfica cuando no se obtiene información en *Open Street Map* para las características *PopulatedPlace* y *MountainArea*. Para buscar el lugar poblado más cercano a la ubicación del sensor se utiliza la etiqueta ADMINISTRATIVE de *Geonames*. En dicho caso devuelve una lista de lugares siendo el primero aquel que se selecciona como el valor para la característica geográfica solicitada. En el caso de las zonas montañosas se utiliza la etiqueta HYPISOGRAPHIC.
- *DBpedia*: Se consulta para obtener el número de habitantes de lugares poblados. El método selecciona lugares cuando el número de habitantes N satisface $N > k_2$ (donde k_2 es un parámetro de calibración). Para las pruebas se utilizó el valor $k_2 = 0$ al considerar lugares poblados con cualquier número de habitantes.
- *Panoramio*: Se utiliza para averiguar si un determinado tipo de ubicación (por ejemplo, Lugar histórico o Bosque) es relevante. Cuando el número de fotografías N satisface $N > k_3$ (donde k_3 es un parámetro de calibración), se considera la característica geográfica como relevante. Para las pruebas se utilizó el valor $k_3 = 5$, que se ajustó empíricamente con pruebas parciales.

En resumen, la aplicación del método general al caso considerado en el experimento 2 tiene las siguientes características:

- Se utiliza un conjunto de características geográficas como referencias que forman parte de las descripciones lingüísticas.
- Se utilizan cuatro fuentes externas: *OSM*, *Geonames*, *DBPedia*, *Panoramio*.
- Se manejan patrones lingüísticos en forma de plantillas de texto con espacios para las características geográficas.
- El procedimiento sigue una estrategia de generación y prueba, buscando primero en unas fuentes (por ejemplo, *OSM*) y después en otras (por ejemplo, *Geonames*).

6.2.2. Procedimiento de evaluación

El objetivo de la evaluación consistió en medir la calidad de las descripciones generadas. Esta vez, en la evaluación se utilizó un conjunto de datos construido expresamente para este propósito. El conjunto de datos de evaluación se creó utilizando la información pública de la red SAIH (disponible en la Web del Ministerio de Medio Ambiente Español). Dicha información es generada y mantenida por los centros de control del SAIH en las principales cuencas hidrográficas, con el apoyo y coordinación del citado Ministerio. Incluye una descripción textual proporcionada por expertos de dominio para cada uno de los 1.811 sensores. Al ser 10 centros de control, las descripciones textuales fueron escritas al menos por 10 expertos diferentes.

Como métrica de evaluación de calidad se consideró la relación entre número de descripciones correctas y el número total de sensores. Se asume que una descripción es correcta cuando las referencias geográficas de la descripción generada son las mismas que las referencias geográficas de la descripción del conjunto de datos de evaluación. Este análisis se hizo de forma global, es decir considerando el total de sensores. También se realizó de forma separada para cada subred correspondiente a cada cuenca hidrográfica.

Con el fin de tener una base de referencia de comparación, los resultados obtenidos se contrastaron con un método que generó descripciones usando únicamente la fuente Geonames. Esta solución alternativa ha sido utilizada

por un gran número de aplicaciones para generar descripciones de ubicaciones geográficas y requiere un esfuerzo de implementación menor.

Figura 6.2: Distribución de sensores SAIH por centro de control.

Con más detalle, la solución basada en Geonames utilizada para generar descripciones para la localización de un sensor, cuyas coordenadas geográficas son (x, y) , fue la siguiente:

- *Pluviómetro*: La descripción textual para los pluviómetros es el nombre proporcionado por *Geonames* para (x, y) con la etiqueta ADMINISTRATIVE.
- *Sensor de nivel (pantano)*: La descripción textual de un sensor de nivel es el nombre proporcionado por *Geonames* para (x, y) con la etiqueta HYDROGRAPHIC.
- *Sensor de caudal*: La descripción textual de un sensor de caudal es la combinación de los nombres $n1$ y $n2$ para (x, y) correspondientes a las etiquetas HYDROGRAPHIC y ADMINISTRATIVE. La descripción se genera como " $n1$ en $n2$ ".

6.2.3. Resultados de la evaluación

Como resultado de la evaluación, se obtuvo una medida de calidad global de la aplicación del método propuesto de 0.52 (para el conjunto total de

sensores.) La calidad de las descripciones de la solución basada en *Geonames* (base de referencia) obtuvo una medida de 0.30. La evaluación también mostró diferencias significativas entre las zonas geográficas correspondientes a diferentes cuencas. Los mejores valores son para las áreas:

- cuenca del Júcar (0.66)
- cuenca del Guadiana (0,63)
- cuenca catalana (0,63)

El peor valor corresponde a la cuenca del Tajo (0.37). Dicho valor puede justificarse porque los expertos del dominio en ese ámbito crearon las descripciones siguiendo estrategias de nombrado totalmente diferentes y, en algunos casos aleatorias, en comparación con el resto áreas.

Resulta interesante ver los resultados de acuerdo a los diferentes tipos de medida (nivel, caudal y lluvia). Por ejemplo, para el acierto medio en nivel se obtiene un buen valor (0.89). Los tres mejores valores para las descripciones de nivel son para las siguientes áreas:

- cuenca del Júcar (1.00)
- cuenca del Guadalquivir (0,93)
- cuenca del Ebro (0,92)

En dichos casos el método obtiene muy buenos resultados cuando hay información disponible en las fuentes de datos geográficas externas.

El acierto medio en el caso de las descripciones de caudal es 0.58. En este caso, el método obtiene muy buenos resultados en tres áreas:

- cuenca del Duero (1.00)
- cuenca de Andalucía (1,00)
- cuenca del Júcar (0,88)

Un caso particular que debe tenerse en consideración corresponde al de la cuenca del Tajo, donde se obtiene un valor de acierto de 0.1. Se trata de un caso especial, en el que los expertos hidrólogos siguen una estrategia de nombrado que no coincide con el enfoque general.

Las descripciones para pluviómetros obtienen un acierto promedio de 0.44. Los tres mejores valores son para:

- cuenca de Cataluña (0,64)
- cuenca del Guadiana (0,61)
- cuenca del Júcar (0,57)

Las descripciones en este caso obtienen una precisión menor en comparación con los anteriores tipos de medida. Una explicación para dicha diferencia es que las descripciones de pluviómetros requieren un conocimiento geográfico más diverso en comparación con otros tipos de sensores y que esta información puede no estar siempre presente en las fuentes de datos geográficas abiertas externas.

Figura 6.3: Resultados de evaluación por centro de control.

Se han analizado las principales causas de generación de descripciones incorrectas. Se distinguen las siguientes situaciones (Figura 6.4):

- *Falta referencia geográfica.* La referencia geográfica requerida (población, embalse, río, montaña, monumento, etc.) no está presente en las bases de datos geográficas abiertas. Este problema sucede en el 20 % de los casos. La información que falta podría ser también causada porque la región de búsqueda seleccionada para un sensor es demasiado pequeña como para encontrar la referencia geográfica correcta.
- *Faltan atributos geográficos.* La referencia geográfica está presente pero no es posible obtener algunos de sus atributos (nombre del río, número de habitantes, etc.). Esto ocurre en el 3 % de los casos. Por ejemplo, los problemas pueden ser pobres en información (información incompleta), información excesivamente detallada, información errónea, o un lenguaje diferente.
- *Selección geográfica incorrecta.* La referencia geográfica correcta está presente en las fuentes de conocimiento pero el método selecciona erróneamente una referencia diferente. Esto ocurre en el 25 % de los casos.

En las dos primeras situaciones (23 % de los casos), los problemas son causados por defectos en los datos geográficos abiertos. Si se resolvieran dichos estos defectos, el método aumentaría su rendimiento hasta una medida del 0.75, lo que representa una mejora del 144 %. Estos defectos podrían ser resueltos con información más completa o proporcionando otros métodos de búsqueda además de la búsqueda por región rectangular.

En la tercera situación, los problemas son causados por estrategias incorrectas de asignación de nombres seguidas por el método para la generación de texto basada en plantillas (el 25 % de los casos). Para mejorar este problema, se requeriría una comprensión más profunda de cómo los expertos generan nombres para reducir este porcentaje. Sin embargo, un factor que también contribuye a este porcentaje es que se ha utilizado a diferentes expertos en el dominio de las diferentes áreas geográficas: Se han obtenido diferencias significativas entre los valores de las diferentes áreas geográficas debido a la selección geográfica equivocada (15 % en la cuenca del Júcar y 35 % en la Cuenca del Duero). De acuerdo con esto, la estandarización del proceso siguiendo estrategias de nomenclatura uniformes para todas las áreas podría aumentar el rendimiento del método hasta una precisión de 0.85 (por ejemplo, en el caso de la cuenca del Júcar, suponiendo que los problemas en las

áreas geográficas abiertas también sean resueltos).

Figura 6.4: Principales causas de descripciones incorrectas ((a) media y (b) mejor área geográfica).

6.2.4. Conclusiones del experimento 2

Como conclusiones principales del caso analizado como experimento 2 se pueden citar las siguientes:

- El método general resulta aplicable a este caso. Tal como se plantea en el método general, se maneja (1) una estrategia de generación y prueba para búsqueda de referencias geográficas (en este caso con cuatro fuentes de información externas) y (2) un conjunto de patrones lingüísticos con criterios de selección como soporte a la generación de descripciones (en este caso, los patrones se representan en forma de plantillas de texto, aunque con una representación más rica que la utilizada en el experimento 1).
- Se realiza una evaluación más amplia que en el experimento 1 (con una red diversa con aproximadamente 1800 sensores) mediante un procedimiento semi-automático y con una base de referencia de comparación. Se obtienen resultados satisfactorios en términos relativos, en cuanto a que la solución basada en el método propuesto supone una mejora del 73 % en comparación con la base de referencia.
- La medida obtenida en la evaluación global es 0.52 (en términos absolutos). Esta medida de calidad muestra además una dispersión significativa (geográfica y respecto a tipos de sensores). Este resultado muestra que cabe todavía introducir posibles mejoras hasta alcanzar la medida de calidad óptima (valor 1.00). En el análisis de los resultados

de la evaluación se identificaron las siguientes mejoras potenciales en las bases de datos externas que pueden mejorar la medida global del método (hasta un valor de 0.75):

- *Contenido uniforme.* La información de las bases de datos abiertas debe presentar uniformidad. Actualmente, ciertas áreas geográficas en *OpenStreetMap* presentan información más detallada que otras. En ciertos casos, la información importante para el método no está presente (por ejemplo, embalses, monumentos históricos, etc.). En la evaluación se encontraron diferencias significativas en el porcentaje de referencias geográficas que faltan: Confederación Hidrográfica del Segura (20 %) y las cuencas catalanas (10 %).
- *Etiquetas estándar.* Las estrategias de asignación de etiquetas para la información deben estandarizarse siguiendo unas prácticas comunes. En contraste con esto, algunos datos etiquetados de *OpenStreetMap* no siguen el estándar (por ejemplo, en Cataluña) y a veces se utilizan nombres excesivamente específicos en lugar de nombres más generales (por ejemplo, "presa" en lugar de "embalse"). La consecuencia de este problema es que el método no es capaz de encontrar alguna información solicitada, incluso aunque dicha información está presente en las bases de datos.
- *Contenido multi-lenguaje.* Los datos deberían cubrir los idiomas locales así como descripciones internacionales. Por ejemplo, en *Geonames* y en *OpenStreetMap* un problema importante es la presencia de diferentes idiomas.
- *Métodos de búsqueda.* Otra mejora estaría relacionada con métodos de búsqueda en las bases de datos geográficos abiertos. *OpenStreetMap* proporciona un método de búsqueda basado en regiones rectangulares. Dicha solución establece una restricción espacial rígida y puede devolver una gran cantidad de información inútil cuando la búsqueda está centrada sólo en algunas etiquetas. Sería útil disponer de un método de búsqueda más flexible basado en etiquetas que devuelva unos máximos establecidos con los datos asociados a las etiquetas específicas, en torno a una ubicación geográfica determinada.

- En comparación con el experimento 1, la solución en este caso no hace

6.3. Experimento 3. Referencias temporales

El diagrama clasifica las referencias en dos categorías principales: geográficas y temporales. Las geográficas se dividen en sin-repetición (PUNTUAL y REGIONAL) y con-repetición (cada ciudad del norte de España y las capitales de Europa). Las temporales se dividen en sin-repetición (PUNTUAL y REGIONAL) y con-repetición (Todos los viernes y El primer día de cada mes).

```

graph TD
    Referencias[referencias] --> Geograficas[geográficas]
    Referencias --> Temporales[temporales]
    Geograficas --> SinRep[sin-repetición]
    Geograficas --> ConRep[con-repetición]
    SinRep --> Puntual[PUNTUAL]
    SinRep --> Regional[REGIONAL]
    Puntual --> Nombre[nombre]
    Regional --> Poblacion[población]
    Regional --> Nacion[nación]
    Regional --> Provincia[provincia]
    Regional --> Continente[continente]
    Regional --> Cuenca[cuenca]
    ConRep --> CadaCiudad[cada ciudad del norte de España]
    ConRep --> LasCapitales[las capitales de Europa]
    Temporales --> SinRepTemp[sin-repetición]
    Temporales --> ConRepTemp[con-repetición]
    SinRepTemp --> PuntualTemp[PUNTUAL]
    SinRepTemp --> RegionalTemp[REGIONAL]
    PuntualTemp --> NombreTemp[nombre]
    RegionalTemp --> Segundo[segundo]
    RegionalTemp --> Minuto[minuto]
    RegionalTemp --> Dia[día]
    RegionalTemp --> Mes[mes]
    RegionalTemp --> Siglo[siglo]
    ConRepTemp --> TodosLosViernes[Todos los viernes]
    ConRepTemp --> ElPrimerDia[El primer día de cada mes]
  
```

El método propuesto maneja *referencias* (por ejemplo, geográficas: “río Júcar” o “Algemesi”) como elementos singulares con los se construyen descripciones lingüísticas (por ejemplo, “río Júcar a su paso por Algemesi”). La Figura 6.5 muestra una posible clasificación de referencias mostrando su dualidad espacio-tiempo. En principio, tanto si se trabaja con *referencias* geográficas como con temporales, se pueden dividir en dos categorías: *sin-repetición* y *con-repetición*. A su vez, la categoría sin-repetición puede descomponerse en: puntual y regional. En la Figura 6.5 se pueden ver ejemplos de cada uno.

118

ponde a sentencias tales como “todos los viernes”, “el primer día de cada mes”, etc. La información con repeticiones en el dominio geográfico podría ser, por ejemplo, “cada ciudad del norte de España”, “las capitales de Europa”, etc. Este tipo de expresiones no se aborda en el presente trabajo y puede ser objeto de investigación como línea de actividad futura.

6.3.1. Aplicación del método

El problema considerado en esta sección es la generación de descripciones lingüísticas con referencias temporales (en vez de geográficas) tales como las siguientes:

- “Retenciones registradas a la vuelta de vacaciones de Semana Santa”
- “Un año después del atentado de las Torres Gemelas de Nueva York”

Es decir, se trata de sentencias que hacen uso de referencias temporales conocidas y utilizadas por un gran número de personas (por lo que pueden considerarse *landmarks* temporales). Dichas descripciones pueden resultar muy convenientes en un contexto de enriquecimiento de datos sobre etiquetas temporales para ayudar a generar expresiones más naturales y cercanas a cómo son utilizadas por personas en determinados contextos. Este problema ha sido tratado parcialmente en algunos trabajos previos[Molina and Stent, 2009, Molina and Stent, 2010](ver Tabla 6.4).

Con el fin de analizar la aplicabilidad del método planteado en este trabajo a este problema, debe considerarse si se satisfacen los supuestos en los que se han basado el diseño del método. Dichos supuestos asumen que existe un volumen importante de conocimiento sobre referencias (temporales en este caso) y que dicho conocimiento está accesible en bases de datos abiertas.

Estos supuestos sí se cumplen, dado que las referencias temporales (tanto puntuales como de región) son abundantes y de diversa naturaleza como, por ejemplo, eventos o acontecimientos históricos (atentado de las Torres Gemelas), ontologías sobre calendarios (días, meses, etc.), fiestas religiosas (teniendo en cuenta diferentes confesiones religiosas en diferentes áreas geográficas), etc. Por otra parte, existen fuentes externas abiertas que podrían

proporcionar dicha información tales como: *DBPedia*, calendarios tipo *Google Calendar* y ontologías especializadas en calendarios.

Atributo	Valores
<i>Julian day</i>	<i>Number of days since 12:00 on January 1st of year -4712</i>
<i>Epoch day</i>	<i>Number of seconds since 00:00 on January 1, 1970.</i>
<i>Minute</i>	{1, ..., 59}
<i>Quarter of an hour</i>	{first, second, third, fourth}
<i>Half an hour</i>	{first, second}
<i>Hour</i>	{1, ..., 24}
<i>Part of day</i>	{morning, afternoon, evening, night}
<i>Third of day</i>	{early, middle, late}
<i>One-day holiday</i>	{Thanksgiving, Christmas, Columbus Day, ...}
<i>Day</i>	{1, ..., 31}
<i>Day of week</i>	{Monday, ..., Sunday}
<i>Few-day holiday</i>	{Christmas, Easter, ...}
<i>Part of week</i>	{weekday, weekend}
<i>Third of week</i>	{early, middle, late}
<i>Number of week</i>	{1, 2, ...}. <i>Number of week in the year.</i>
<i>Third of month</i>	{early, middle, late}
<i>Month</i>	{January, ..., December}
<i>Season</i>	{winter, spring, summer, fall}
<i>Quarter</i>	{first, second, third, fourth}
<i>Third of year</i>	{early, middle, late}
<i>Semester</i>	{first, second}
<i>Year</i>	{..., 1995, 1996, ...}
<i>Financial year</i>	{..., 1995, 1996, ...}
<i>Third of decade</i>	{early, middle, late}
<i>Decade</i>	{..., 1980's, 1990's, ...}
<i>Third of century</i>	{early, middle, late}
<i>Century</i>	{..., 18, 19, 20, ...}

Tabla 6.4: Atributos temporales usados por [Molina and Stent, 2010].

Además, es posible representar patrones lingüísticos con referencias temporales junto a criterios para su selección a partir de información de las características. Se trata de una representación para construcción de descripciones con alto grado de flexibilidad y generalidad (no comprometida con el dominio geográfico). Por ejemplo, para el caso de referencias temporales, se podría tener el siguiente patrón:

Patrón(
[Referencia(x_1 ,
festivo – *en* – *calendario* – *musulmán*,
fecha(16 – 11 – 2001)),
[Referencia(x_2 ,
principal – *acontecimiento* – *histórico* – *de* – *año*,
fecha(16 – 11 – 2001))],
[Relación(*en*, x_1 , x_2))]

Dicho patrón podría corresponder a la siguiente sentencia en lenguaje natural:

“El día del inicio de Ramadán en el año del atentado de las Torres Gemelas de Nueva York”

6.3.2. Conclusiones del experimento 3

Como conclusiones principales del análisis de aplicabilidad del método para generación de descripciones con referencias temporales se pueden mencionar las siguientes:

- El problema abordado por el método propuesto está concebido de forma general de forma que sí puede particularizarse a expresiones temporales, basándose en la dualidad espacio-tiempo de las referencias.
- Se cumplen también los supuestos considerados en cuanto al uso de un volumen importante y diverso de información que puede estar disponible en bases de datos externas.
- La representación del conocimiento planteada por el método (por ejemplo, en los patrones lingüísticos) no está comprometida con la naturaleza geográfica o temporal de las referencias, por lo que es también aplicable en ambos casos.

En consecuencia, es posible afirmar que el método propuesto puede aplicarse a la generación de referencias temporales. No obstante, queda pendiente realizar una evaluación práctica sobre un caso real con suficiente complejidad que determine el rendimiento del método en este caso con las posibles bases de datos externas a utilizar. Esta evaluación podría incluirse como trabajo futuro de continuación del presente trabajo.

Parte IV

CONCLUSIONES

7. Conclusiones

En este capítulo se resume el trabajo realizado en relación con los objetivos planteados, se destacan las aportaciones principales y se indican posibles líneas futuras de investigación.

7.1. Resumen

El presente trabajo se ha centrado en el problema de enriquecer los datos sobre redes de sensores con expresiones en lenguaje natural (por ejemplo, sobre localización) con el fin de facilitar la posterior generación de textos más complejos. El uso de descripciones en lenguaje natural facilita el acceso a los datos a una mayor diversidad de usuarios y, como consecuencia, permite aprovechar mejor las inversiones en redes de sensores. En el trabajo se ha considerado el uso de bases de datos abiertas para abordar la necesidad de disponer de un gran volumen y diversidad de conocimiento geográfico.

El trabajo incluyó la revisión del estado del arte sobre los temas relacionados con los objetivos generales. En concreto, se revisaron los avances recientes en las áreas de la curación de datos, bases de datos abiertas y de la generación de lenguaje natural, especialmente para el problema de generación de expresiones de referencias.

Se diseñó un método general para el problema considerado, basado en una estrategia de generación y prueba que incluye una forma de representación y uso del conocimiento heurístico con varias etapas de razonamiento para la construcción de descripciones lingüísticas de enriquecimiento de datos.

En la evaluación de la propuesta general se han manejado tres escenarios, dos de ellos para generación de referencias geográficas sobre redes de sensores complejas de dimensión real y otro para generación de referencias temporales. Los resultados de la evaluación han mostrado la validez de la propuesta general mostrando mejoras de rendimiento respecto a otros enfoques. Además, el análisis de los resultados ha permitido identificar y cuantificar el impacto previsible de diversas líneas de mejora en bases de datos abiertas.

7.2. Aportaciones

Como resultado del presente trabajo se pueden mencionar las siguientes contribuciones:

1. Identificación de un tipo de problema a resolver para mejorar el acceso a datos de redes de sensores a un mayor número de usuarios. El problema consiste en generar descripciones lingüísticas que realicen un enriquecimiento de datos sobre redes de sensores para facilitar la generación posterior de textos complejos relacionados con dichas redes.
2. Planteamiento de una solución basada en el uso de datos abiertos relacionados con información geográfica y temporal. La solución está concebida y formalizada de forma general, es decir, aplicable a diferentes tipos de redes de sensores de diferentes dominios.
3. Demostración de la aplicabilidad práctica de la propuesta teórica sobre redes de dimensión real para generar tanto expresiones de referencias espaciales o temporales.
4. Obtención de unos resultados de rendimiento que mejoran significativamente otros enfoques. Por ejemplo, se obtuvo una mejora de un 73 % en comparación con la generación de referencias mediante un servicio ampliamente utilizado como es *Geonames*.
5. Identificación de problemas a mejorar en bases de datos abiertas sobre dominios geográficos y cuantificación del impacto que las soluciones puede tener en la mejora del rendimiento del método propuesto. En concreto las mejoras propuestas son uniformizar los contenidos, el uso de etiquetas estándar, de contenidos multilenguaje así como la extensión de los métodos de búsqueda (por ejemplo, de forma global se cuantificó que dichas soluciones pueden incrementar el rendimiento del método propuesto hasta un 144 %).
6. Construcción de un juego de datos de evaluación (correspondiente a unos 1.800 sensores) sobre una red de sensores hidrológica²⁰. Dicho juego de datos se ha publicado y puesto a disposición de la comunidad

¹⁹El juego de datos se encuentra publicado en <http://www.mdpi.com/1424-8220/15/7/16009/s1> (último acceso 03/11/2015).

investigadora con el fin de servir de herramienta de comparación entre la eficacia de otros métodos similares.

7.3. Líneas futuras

Algunas posibles líneas de trabajo futuras que pueden plantearse como continuación del presente trabajo son las siguientes:

- Trabajar en expresiones construidas a partir de información con repeticiones, tanto en el dominio geográfico como en el dominio temporal dado que no se aborda en el presente trabajo y puede ser objeto de investigación como línea de actividad futura.
- Realizar mejoras en las fuentes de datos externas, de acuerdo con las propuestas realizadas en este trabajo. Dichas fuentes de datos pueden incorporar en concreto: (1) mejoras en los métodos de búsqueda y (2) procesos automáticos que ayuden a observar problemas de calidad junto con sugerencias de mejora.
- Estudiar la incorporación de ayudas a la construcción de las bases de conocimiento del método con nuevos avances sobre tecnologías semánticas (por ejemplo, ontologías) sobre redes de sensores.
- Integrar la solución final propuesta con sistemas de presentación en lenguaje natural y realizar evaluaciones conjuntas del rendimiento y su utilidad práctica (este aspecto se llevó a cabo parcialmente con el experimento 1).
- Aplicar el método a nuevos dominios y otras redes de sensores extensas incluyendo además la generación de expresiones temporales.

Referencias

- [Aguilar et al., 2015] Aguiar, R. L., Benhabiles, N., Pfeiffer, T., Rodriguez, P., Viswanathan, H., Wang, J., and Zang, H. (2015). Big data, iot, buzz words for academia or reality for industry? In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, MobiCom '15, pages 550–551, New York, NY, USA. ACM.
- [Aloysius, 2001] Aloysius, J.-R. (2001). National disaster management information systems & networks: An Asian overview. *GDIN*.
- [Appelt, 1985] Appelt, D. (1985). Planning english referring expressions. *Artificial Intelligence*, (26):1–33.
- [Appelt and Kronfeld, 1987] Appelt, D. and Kronfeld, A. (1987). A computational model of referring. *10th International Joint Conference on Artificial Intelligence*, pages 640–647.
- [Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data*. Springer.
- [Ballatore and Bertolotto, 2013] Ballatore, A. and Bertolotto, M. (2013). A survey of volunteered open geo-knowledge bases in the semantic web. pages 93–120.
- [Bizer et al., 2009] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *Int. J. Semant. Web Inf. Syst.*, 5:1–22.
- [Botts et al., 2008] Botts, M., Percivall, G., Reed, C., and Davidson, J. (2008). Ogc® sensor web enablement: Overview and high level architecture. In *GeoSensor networks*, pages 175–190. Springer.
- [Calder et al., 2010] Calder, M., Morris, R. A., and Peri, F. (2010). Machine reasoning about anomalous sensor data. *Ecological Informatics*, 5(1):9–18.
- [Cervone, 2008] Cervone, F. (2008). Itil: a framework for managing digital library services. *OCLC Systems & Services: International digital library perspectives*, 24(2):87–90.

- [Clabo and Ramos-Vielba, 2015] Clabo, N. and Ramos-Vielba, I. (2015). Re-utilización de datos abiertos en la administración pública en españa y uso de licencias-tipo/the re-use of open data in the spanish public administration and the application of public copyright licenses. *Revista Espanola de Documentacion Cientifica*, 38(3):1.
- [Compton et al., 2012] Compton, M., Barnaghi, P., Bermudez, L., Garca-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., Huang, V., Janowicz, K., Kelsey, W. D., Phuoc, D. L., Lefort, L., Leggieri, M., Neuhaus, H., Nikolov, A., Page, K., Passant, A., Sheth, A., and Taylor, K. (2012). The {SSN} ontology of the {W3C} semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:25 – 32.
- [Compton et al., 2009a] Compton, M., Henson, C. A., Lefort, L., Neuhaus, H., and Sheth, A. P. (2009a). A survey of the semantic specification of sensors.
- [Compton et al., 2009b] Compton, M., Neuhaus, H., Taylor, K., and Tran, K.-N. (2009b). Reasoning about sensors and compositions. In *SSN*, pages 33–48. Citeseer.
- [Dale et al., 2005] Dale, R., Geldof, S., and Prost, J.-P. (2005). Using natural language generation in automatic route description. *Journal of Research and Practice in Information Technology*, 37.
- [Dale and Reiter, 1995] Dale, R. and Reiter, E. (1995). Computational interpretations of the grice maxims in the generation of referring expressions. *Cognitive Science*, 18.
- [de Lerma et al., 2008] de Lerma, M. S., Abeign, C. I., and Power, P. A. (2008). Sistemas de referencia e identificadores geogrficos (iso 19111, iso 19112, iso 6709). *Mapping*, (123):40–47.
- [DGA, 2009] DGA (2009). Direccin geneal del agua. el programa saih: Descripcin y funcionalidad. el presente y el futuro del sistema.
- [Dubey et al., 2015] Dubey, H., Yang, J., Constant, N., Amiri, A. M., Yang, Q., and Makodiya, K. (2015). Fog data: Enhancing telehealth big data through fog computing. In *Proceedings of the ASE BigData & SocialInformatics 2015*, page 14. ACM.

- [Ebert et al., 1996] Ebert, C., Glatz, D., Jansche, m., Meyer-Klabunde, R., and Porzel, R. (1996). From conceptualization to formulation in generating spatial description. *Proceedings of the First European Workshop on Cognitive Modeling*, pages 235–241.
- [Eid et al., 2006] Eid, M., Liscano, R., and El Saddik, A. (2006). A novel ontology for sensor networks data. In *Computational Intelligence for Measurement Systems and Applications, Proceedings of 2006 IEEE International Conference on*, pages 75–79. IEEE.
- [Eid et al., 2007] Eid, M., Liscano, R., and El Saddik, A. (2007). A universal ontology for sensor networks data. In *Computational Intelligence for Measurement Systems and Applications, 2007. CIMSIA 2007. IEEE International Conference on*, pages 59–62. IEEE.
- [Garbarino and Echeverria, 2010] Garbarino, J. and Echeverria, A. (2010). *Protocolos para redes inalámbricas de sensores*. PhD thesis, Universidad de Buenos Aires.
- [Garden, 2002] Garden, C. (2002). Generating minimal definite descriptions. In *40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 96–103, Philadelphia, PA, USA.
- [Goodchild, 2007] Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.
- [Goodwin and Russomanno, 2006] Goodwin, C. and Russomanno, D. J. (2006). An ontology-based sensor network prototype environment. In *Proceedings of the Fifth International Conference on Information Processing in Sensor Networks*, pages 1–2.
- [Goodwin et al., 2007] Goodwin, J. C., Russomanno, D. J., and Qualls, J. (2007). Survey of semantic extensions to uddi: Implications for sensor services. In *SWWS*, pages 16–22.
- [Grice, 1975] Grice, H.-P. (1975). Logic and conversation. *P. Cole and J. Morgan, editors, Syntax and Semantics, Speech Acts*. Academic Press, 3.
- [Gubbi et al., 2013] Gubbi, J., Buyya, R., Marusic, S., and Palaniswami, M. (2013). Internet of things (iot): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7):1645 – 1660.

- [Haklay et al., 2010] Haklay, M. et al. (2010). How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning. B, Planning & design*, 37(4):682.
- [Herold et al., 2005] Herold, S., Swada, M., and Wellar, B. (2005). Integrating geographic information systems, spatial databases and the internet: A framework for disaster management. *Annual Canadian Institute of Geomatics Conference. University of Ottawa*.
- [Herzog and Buchmann, 2009] Herzog, A. and Buchmann, A. (2009). Predefined classification for mixed mode environments. *Technische Universität Darmstadt*.
- [Herzog et al., 2008] Herzog, A., Jacobi, D., and Buchmann, A. (2008). A3me-an agent-based middleware approach for mixed mode environments. In *Mobile Ubiquitous Computing, Systems, Services and Technologies, 2008. UBICOMM'08. The Second International Conference on*, pages 191–196. IEEE.
- [Higgins, 2008] Higgins, S. (2008). The dcc curation lifecycle model. *International Journal of Digital Curation*, 3(1):134–140.
- [Holger and Compton, 2009] Holger, N. and Compton, M. (2009). The semantic sensor network ontology: A generic language to describe sensor assets. In *12th AGILE International Conference on Geographic Information Science, Workshop on Challenges in Geospatial Data Harmonisation, Hannover, Germany*.
- [Hu et al., 2007] Hu, Y., Wu, Z., and Guo, M. (2007). Ontology driven adaptive data processing in wireless sensor networks. In *Proceedings of the 2nd international conference on Scalable information systems*, page 46. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).
- [Hunter et al., 2008] Hunter, J., Gatt, A., Portet, F., Reiter, E., and Sripatha, G. (2008). Using natural language generation technology to improve information flows in intensive care units. In *5th Conference on Prestigious Applications of Intelligent Systems*, Patras, Greece.

- [Jiang and Thill, 2015] Jiang, B. and Thill, J.-C. (2015). Volunteered geographic information: Towards the establishment of a new paradigm. *Computers, Environment and Urban Systems*, 53:1 – 3. Special Issue on Volunteered Geographic Information.
- [Kelleher and Kruijff, 2006] Kelleher, J. and Kruijff, G. (2006). Incremental generation of spatial referring expressions in situated dialog. *Proceedings of the 21st International Conference on Computational Linguistics (COLING) and 44th Annual Meeting of the Association for Computational Linguistics (ACL), Sydney, Australia*.
- [Kim and Myung, 2015] Kim, K. and Myung, H. (2015). Sensor node for remote monitoring of waterborne disease-causing bacteria. *Sensors*, 15(5):10569–10579.
- [Krahmer et al., 2003] Krahmer, E., Erk, S., and Verleg, A. (2003). Graph-based generation of referring expressions. *Comput. Linguist.*, (29):53–72.
- [Krahmer and Theune, 2002] Krahmer, E. and Theune, M. (2002). Efficient context-sensitive generation of referring expressions. *K. van Deemter & R. Kibble (Eds.). Information sharing: Givenness and newness in language processing. CSLI Publications, Stanford*.
- [Kronfeld, 1990] Kronfeld, A. (1990). Reference and computation: An essay in applied philosophy of language. *Cambridge University Press*.
- [Kviesis et al., 2015] Kviesis, A., Zacepins, A., Durgun, M., and Tekin, S. (2015). Application of wireless sensor networks in precision apiculture. *Engineering for rural development*.
- [Lester and Bares, 1998] Lester, J. C. and Bares, W. H. (1998). Natural language generation journeys to interactive 3d worlds.
- [Lourenço et al., 2015] Lourenço, A., Alves, A. P., Carreiras, C., Duarte, R. P., and Fred, A. (2015). Cardiowheel: Ecg biometrics on the steering wheel. In *Machine Learning and Knowledge Discovery in Databases*, pages 267–270. Springer.
- [Machado et al., 2015] Machado, A., Veloso, A., Lobo-Pappa, G., and Meira, W. (2015). Data enrichment technical note. W3C Editor’s Draft.

- [Mashhadi et al., 2015] Mashhadi, A., Quattrone, G., and Capra, L. (2015). The impact of society on volunteered geographic information: The case of openstreetmap. In *OpenStreetMap in GIScience*, pages 125–141. Springer.
- [Mezquida and Martínez, 2009] Mezquida, D. A. and Martínez, J. L. (2009). Short communication.: Platform for bee-hives monitoring based on sound analysis. a perpetual warehouse for swarm s daily activity. *Spanish journal of agricultural research*, (4):824–828.
- [Molina, 2012] Molina, M. (2012). Simulating data journalism to communicate hydrological information from sensor networks. In Springer, editor, *IBERAMIA 2012*, pages 722–731, Berlin, Germany.
- [Molina and Bayarri, 2011] Molina, M. and Bayarri, S. (2011). A multinational sdi-based system to facilitate disaster risk management in the andean community. *Computers & Geosciences*, 37(9):1501–1510.
- [Molina and Flores, 2012] Molina, M. and Flores, V. (2012). Generating multimedia presentations that summarize the behavior of dynamic systems using a model-based approach. *Expert Syst. Appl.*, (39):2759–2770.
- [Molina and Sanchez-Soriano, 2010] Molina, M. and Sanchez-Soriano, J. (2010). Modeling sensor knowledge of a national hydrologic information system. *IADIS International Conference - Applied Computing 2011*, 978-989-8425-31-7.
- [Molina and Sanchez-Soriano, 2011] Molina, M. and Sanchez-Soriano, J. (2011). Generating text descriptions for geographically distributed sensors. *Proceedings of the IADIS International Conference on WWW/Internet Interaction. Rio de Janeiro*, 978-989-8533-06-7.
- [Molina et al., 2015] Molina, M., Sanchez-Soriano, J., and Corcho, O. (2015). Using open geographic data to generate natural language descriptions for hydrological sensor networks. *Sensors*, 15(7):16009.
- [Molina and Stent, 2009] Molina, M. and Stent, A. (2009). Generating descriptions that summarize geospatial and temporal data. In *Tools with Artificial Intelligence, 2009. ICTAI’09. 21st International Conference on*, pages 485–492. IEEE.

- [Molina and Stent, 2010] Molina, M. and Stent, A. (2010). A knowledge-based method for generating summaries of spatial movement in geographic areas. *International Journal on Artificial Intelligence Tools*, 19(04):393–415.
- [Moulin and Kettani, 1999] Moulin, B. and Kettani, D. (1999). Route generation and description using the notions of object’s influence area and spatial conceptual map. *Spatial Cognition and Computation*, 1(3):227–259.
- [Probst, 2006] Probst, F. (2006). Ontological analysis of observations and measurements. In *Geographic information science*, pages 304–320. Springer.
- [Qanbari et al., 2015] Qanbari, S., Behinaein, N., Rahimzadeh, R., and Dustdar, S. (2015). Gatica: Linked sensed data enrichment and analytics middleware for iot gateways. In *Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on*, pages 38–43.
- [Reiter and Dale, 1992] Reiter, E. and Dale, R. (1992). A fast algorithm for the generation of referring expressions. *Information Sharing: Givenness and Newness in Language Processing*.
- [Reiter and Dale, 1995] Reiter, E. and Dale, R. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognit. Sci.*, (18):233–263.
- [Reiter and Dale, 2009] Reiter, E. and Dale, R. (2009). Building natural language generation systems. *Cambridge University Press*.
- [Reiter et al., 2000] Reiter, E., Dale, R., and Feng, Z. (2000). *Building natural language generation systems*, volume 33. MIT Press.
- [Reiter et al., 2005] Reiter, E., Sripada, S., Hunter, J., Yu, J., and Davy, I. (2005). Choosing words in computer-generated weather forecasts. pages 137–169.
- [Rettig et al., 2015] Rettig, A. J., Khanna, S., and Beck, R. A. (2015). Open source rest services for environmental sensor networking. *Applied Geography*, 60:294–300.

- [Ritze and Eckert, 2014] Ritze, D. and Eckert, K. (2014). Data enrichment in discovery systems using linked data. In Spiliopoulou, M., Schmidt-Thieme, L., and Janning, R., editors, *Data Analysis, Machine Learning and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 455–462. Springer International Publishing.
- [Robin and Botts, 2006] Robin, A. and Botts, M. (2006). Creation of specific sensorml process models (draft). *University of Alabama in Huntsville: Huntsville, AL, USA*.
- [Roth and Frank, 2009] Roth, M. and Frank, A. (2009). A nlg-based application for walking direction. In *ACL-IJCNLP 2009 Software Demonstrations*, Suntec, Singapore.
- [Russomanno et al., 2005a] Russomanno, D. J., Kothari, C., and Thomas, O. (2005a). Sensor ontologies: from shallow to deep models. In *System Theory, 2005. SSST'05. Proceedings of the Thirty-Seventh Southeastern Symposium on*, pages 107–112. IEEE.
- [Russomanno et al., 2005b] Russomanno, D. J., Kothari, C. R., and Thomas, O. A. (2005b). Building a sensor ontology: A practical approach leveraging iso and ogc models. In *IC-AI*, pages 637–643.
- [Sánchez Alonso et al., 2008] Sánchez Alonso, S., Sicilia Urbán, M. Á., and de Rato Leguina, G. (2008). Sobre la interoperabilidad semántica en las descripciones archivísticas digitales. *Revista española de documentación científica*, 31(1):11–38.
- [Sarango and Fabricio, 2014] Sarango, C. and Fabricio, E. (2014). Diseño e implementación del prototipo de un sistema domótico para la medición del consumo de agua potable a través de internet y correo electrónico. Master’s thesis, Universidad Nacional de Loja.
- [Schade et al., 2015] Schade, S., Granell, C., and Perego, A. (2015). Coupling public sector information and public-funded research data in europe: A vision of an open data ecosystem. *Information and Communication Technologies in Public Administration: Innovations from Developed Countries*, 195:275.

- [Schreiber et al., 2000] Schreiber, G., Akkermans, H., Anjewierden, A., De-Hoog, R., Shadbolt, N.-R., Van-de Velde, W., and Wielinga, B.-J. (2000). *Knowledge engineering and management: the CommonKADS methodology*. MIT press.
- [Schuldes et al., 2011] Schuldes, S., Boland, K., Roth, M., Strube, M., Krömer, S., and Frank, A. (2011). Modeling spatial knowledge for generating verbal and visual route directions. *Knowledge-Based and Intelligent Information and Engineering Systems Lecture Notes in Computer Science*.
- [Steiniger and Bocher, 2009] Steiniger, S. and Bocher, E. (2009). An overview on current free and open source desktop GIS developments. *International Journal of Geographical Information Science*, 23(10):1345–1370.
- [Stevenson et al., 2009] Stevenson, G., Knox, S., Dobson, S., and Nixon, P. (2009). Ontonym: a collection of upper ontologies for developing pervasive systems. In *Proceedings of the 1st Workshop on Context, Information and Ontologies*, page 9. ACM.
- [Thomas and Sripada, 2008] Thomas, K. E. and Sripada, S. (2008). What’s in a message?: interpreting geo-referenced data for the visually-impaired. *Proceedings of the Fifth International Natural Language Generation Conference*.
- [Turner et al., 2009] Turner, R., Sripada, S., and Reiter, E. (2009). Generating approximate geographic descriptions. *Proceedings of the 12th ENLG - European Workshop on Natural Language Generation, Athens, Greece*.
- [Turner et al., 2006] Turner, R., Sripada, S., Reiter, E., and Davy, I. (2006). Generating spatio-temporal descriptions in pollen forecasts. *EACL06 Companion Volume*.
- [Turner et al., 2008] Turner, R., Sripada, S., Reiter, E., and Davy, I. (2008). Using spatial reference frames to generate grounded textual summaries of georeferenced data. *Proceedings of INLG08*.
- [Underbrink et al., 2008] Underbrink, A., Witt, K., Stanley, J., and Mandl, D. (2008). Autonomous mission operations for sensor webs. In *AGU Fall Meeting Abstracts*, volume 1, page 05.

- [Upton et al., 2015] Upton, V., Ryan, M., O'Donoghue, C., and Dhubhain, A. N. (2015). Combining conventional and volunteered geographic information to identify and model forest recreational resources. *Applied Geography*, 60:69 – 76.
- [Varges, 2005] Varges, S. (2005). Spatial descriptions as referring expressions in the maptask domain. *ENLG-05*.
- [Vatant and Wick, 2006] Vatant, B. and Wick, M. (2006). Geonames ontology (2006). *Online at [http://www.geonames.org/ontology/N.Freire/Facilitating Access and Reuse of Research Materials: The Case of the European Library](http://www.geonames.org/ontology/N.Freire/Facilitating%20Access%20and%20Reuse%20of%20Research%20Materials:The%20Case%20of%20the%20European%20Library)*, 76.
- [Viethenand and Dale, 2008] Viethenand and Dale, R. (2008). The use of spatial relations in referring expressions. *Proceedings of the 5th INLG - International Natural Language Generation Conference*.
- [Vilches-Blázquez et al., 2010] Vilches-Blázquez, L., Villazón-Terrazas, B., Saquicela, V., de León, A., Corcho, O., and Gómez-Pérez, A. (2010). Geo-linked data and inspire through an application case. In *18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 446–449, San Jose, CA, USA.
- [Yu and Liu, 2015] Yu, L. and Liu, Y. (2015). Using linked data in a heterogeneous sensor web: challenges, experiments and lessons learned. *International Journal of Digital Earth*, 8(1):17–37.

Nomenclatura

BBDD Bases de Datos

CSW Catalog Service for the Web

ETL Extract, Transform and Load: Extracción, Transformación y Carga

GER Generación de Expresiones de Referencia

GIS Geographic Information Systems (sistema de información geográfica)

GLN Generación de Lenguaje Natural

GN GeoNames

GPS Global Positioning System (sistema de posicionamiento global)

IDE Infraestructura de Datos Espaciales

IG Información Geográfica

IoT Internet of Things (Internet de las cosas)

LOD Linked Open Data (datos enlazados abiertos)

NWP Numerical Weather Prediction (modelos numéricos de predicción meteorológica)

OGC Open Geospatial Consortium (consorcio de datos geoespaciales abiertos)

OSM Open Street Maps

OWL Web Ontology Language (lenguaje de marcado para ontologías)

QPF Quantitative precipitation forecast (predicciones cuantitativas de precipitación)

RISP Reutilización de la Información del Sector Público

SAIH Sistema Automático de Información Hidrológica de España

SOA Service Oriented Architecture (arquitectura orientada a servicios)

SWE Sensor Web Enablement

TI Tecnologías de la Información

URI Uniform Resource Identifier (identificador de recursos uniforme)

UTM Universal Transverse Mercator

VGI Volunteer Geographic Information (información geográfica voluntaria)

WMS Web Map Service

XML eXtensible Markup Language (lenguaje de marcado extensible)

Parte V

ANEXOS

A. Especificación CommonKADS

En este anexo se presenta la especificación *CommonKADS* de los componentes del modelo de conocimiento para el método general propuesto. Dicha especificación es general (independiente de dominios específicos) y expresa los algoritmos utilizados que sirven de complemento y refino a los modelos gráficos presentados en el capítulo que describe la propuesta del método.

De acuerdo con la metodología *CommonKADS* se describen tres tipos de componentes (tareas, métodos e inferencias) haciendo uso del lenguaje propuesto por la metodología [Schreiber et al., 2000]. Los próximos apartados muestra la especificación de cada componente en dicho orden.

A.1. Tarea *Enriquecer-datos-sobre-sensores*

Algoritmo A.1 Tarea *Enriquecer-datos-sobre-sensores*.

TASK Enriquecer-datos-sobre-sensores;

GOAL:

"Enriquecer fuentes de datos con descripciones
lingüísticas mediante bases de datos geográficas";

ROLES:

INPUT:

Datos-en-bruto: "Datos de redes de sensores";

OUTPUT:

Descripciones: "Descripciones lingüísticas para
datos de redes de sensores";

SPEC:

"Las descripciones se obtienen de un proceso de
búsqueda y generación de descripciones lingüísticas
con acceso a fuentes de información externas. Las
descripciones obtenidas son términos lingüísticos
(nombres de referencias geográficas, partes de
frases completas) que caracterizan los datos de
redes de sensores";

END TASK Enriquecer-datos-sobre-sensores;

A.2. Tarea *Planificar-contenido-de-la-descripción*

Algoritmo A.2 Tarea *Planificar-contenido-de-la-descripción*.

TASK Planificar-contenido-de-la-descripción;
GOAL:
 "Planificar la generación de descripciones
 lingüísticas mediante bases de datos geográficas";
ROLES:
INPUT:
 Datos-de-un-sensor: "Datos en formato unificado
 de un único sensor";
OUTPUT:
 Estructura-de-la-hipótesis: "Patrón lingüístico
 en forma de plantilla";
 Componentes-encontrados: "Características obtenidas
 en las fuentes externas";
 Éxito-de-la-búsqueda: "Indicador de si ha sido
 posible o no encontrar los
 componentes";
SPEC:
 "La Estructura-de-la-hipótesis es una plantilla que
 describe un patrón lingüístico y los
 Componentes-encontrados son los elementos que forman
 parten de dicha plantilla.";
END TASK Planificar-contenido-de-la-descripción;

A.3. Método *Enriquecimiento-con-descripciones*

Algoritmo A.3 Método Enriquecimiento-con-descripciones.

```

TASK-METHOD Enriquecimiento-con-descripciones;
REALIZES: Enriquecer-datos-sobre-sensores;
DECOMPOSITION:
  INFERENCES:
    Preprocesar-datos, Componer-descripción;
ROLES:
  INTERMEDIATE:
    Datos-corregidos: "Datos de sensores en un
                      formato unificado";
    Datos-de-un-sensor: "Datos en formato unificado
                        de un único sensor";
    Plan-de-la-descripción: "Estrategia para construir
                            la descripción de un sensor";
    Descripción-de-un-sensor: "Descripción relativa
                              a un único sensor";
    Descripciones: "Agrupamiento de las descripciones de
                  los sensores";
CONTROL-STRUCTURE:
  Descripciones := [];
  Preprocesar-datos(
    Datos-en-bruto -> Datos-corregidos);
  FOR-EACH Datos-de-un-sensor IN Datos-corregidos DO
    Planificar-contenido-de-la-descripción(
      Datos-de-un-sensor -> Plan-de-la-descripción);
    Componer-descripción(Plan-de-la-descripción ->
      Descripción-de-un-sensor);
    Descripciones := Descripciones +
                    Descripción-de-un-sensor;
  END FOR-EACH
END TASK-METHOD Enriquecimiento-con-descripciones;

```

A.4. Método *Planificar-contenido-con-Open-Data*

Algoritmo A.4 Método Planificar-con-Open-Data.

TASK-METHOD Planificar-contenido-con-Open-Data;
REALIZES: Planificar-contenido-de-la-descripción;
DECOMPOSITION:
 INFERENCES:
 Generar-hipótesis-de-descripción ,
 Seleccionar-fuente-de-datos ,
 Buscar-en-fuente-de-datos;
ROLES:
 INTERMEDIATE:
 Estructura-de-la-hipótesis: "Patrón lingüístico que
 corresponde al caso del sensor tratado".
 Componentes-requeridos-de-la-hipótesis: "Componentes
 requeridos en la Estructura-de-la-hipótesis que
 hay que buscar en las fuentes de datos externas";
 Consulta-en-fuente-de-datos: "Consulta con los
 componentes requeridos para las fuentes de datos";
 Éxito-de-la-búsqueda: "Indicador del resultado de la
 consulta, si ha sido exitosa o ha fallado".
 Componentes-encontrados: "Los componentes que se han
 obtenido de las fuentes de datos".
CONTROL-STRUCTURE:
 Generar-hipótesis-de-descripción(
 Datos-de-un-sensor ->
 Estructura-de-la-hipótesis ,
 Componentes-requeridos-de-la-hipótesis);
 Componentes-encontrados = [];
 FOR-EACH C IN Componentes-requeridos-de-la-hipótesis DO
 REPEAT
 Seleccionar-fuente-de-datos(C ->
 Consulta-en-fuente-de-datos);
 Buscar-en-fuente-de-datos(C ->
 Éxito-de-la-búsqueda, Componente);
 Componentes-encontrados =
 Componentes-encontrados + Componente;
 UNTIL
 Éxito-de-la-búsqueda = false OR
 "no queden más fuentes de datos";
 END WHILE
END FOR-EACH
 Plan-de-la-descripción :=
 Estructura-de-la-hipótesis + Componentes-encontrados;
END TASK-METHOD Planificar-contenido-con-Open-Data;

A.5. Inferencia *Preprocesar-datos*

Algoritmo A.5 Inferencia *Preprocesar-datos*.

```
INFERENCE Preprocesar-datos;
  ROLES:
    INPUT: Datos-en-bruto;
    OUTPUT: Datos-corregidos;
    STATIC: Modelo-de-validación;
  SPECIFICATION:
    "Se realizan diferentes acciones o de preproceso
    que los datos en bruto de entrada sean uniformes
    (filtrar, preparar, validad y adaptar los datos)";
END INFERENCE Preprocesar-datos;

KNOWLEDGE-ROLE Datos-en-bruto;
  TYPE: DYNAMIC;
  DOMAIN-MAPPING: Red-de-sensores,
                  Localización-espacio-temporal;
END KNOWLEDGE-ROLE Datos-en-bruto;

KNOWLEDGE-ROLE Datos-corregidos;
  TYPE: DYNAMIC;
  DOMAIN-MAPPING: Red-de-sensores,
                  Localización-espacio-temporal;
END KNOWLEDGE-ROLE Datos-corregidos;

KNOWLEDGE-ROLE Modelo-de-validación;
  TYPE: STATIC;
  DOMAIN-MAPPING: Clases-de-sensores;
END KNOWLEDGE-ROLE Modelo-de-validación;
```

A.6. Inferencia *Componer-descripción*

Algoritmo A.6 Inferencia *Componer-descripción*.

```
INFERENCE Componer-descripción ;
  ROLES:
    INPUT: Plan-de-la-descripción ;
    OUTPUT: Descripción-de-un-sensor ;
    STATIC: Modelo-léxico ;
  SPECIFICATION:
    "Combina los resultados obtenidos en inferencias
     previas obteniendo la descripción lingüística en el
     formato apropiado." ;
END INFERENCE Componer-descripción ;

KNOWLEDGE-ROLE Plan-de-la-descripción ;
  TYPE: DYNAMIC ;
  DOMAIN-MAPPING: Patrón-lingüístico ,
                  Referencias-espacio/temporales ;
END KNOWLEDGE-ROLE Plan-de-la-descripción ;

KNOWLEDGE-ROLE Descripción-de-un-sensor ;
  TYPE: DYNAMIC ;
  DOMAIN-MAPPING: Descripción-Lingüística ;
END KNOWLEDGE-ROLE Descripción-de-un-sensor ;

KNOWLEDGE-ROLE Modelo-léxico ;
  TYPE: STATIC ;
  DOMAIN-MAPPING: Recursos-léxico-lingüísticos ;
END KNOWLEDGE-ROLE Modelo-léxico ;
```

A.7. Inferencia *Generar-hipótesis-de-descripción*

Algoritmo A.7 Inferencia *Generar-hipótesis-de-descripción*.

```
INFERENCE Generar-hipótesis-de-descripción ;
  ROLES:
    INPUT: Datos-de-un-sensor ;
    OUTPUT: Componentes-requeridos-de-la-hipótesis ,
            Estructura-de-la-hipótesis ;
    STATIC: Modelo-de-generación ;
  SPECIFICATION:
    "Generar una hipótesis de descripción geoespacial
     sobre cómo expresar la información del sensor.";
END INFERENCE Generar-hipótesis-de-descripción ;

KNOWLEDGE-ROLE Datos-de-un-sensor ;
  TYPE: DYNAMIC;
  DOMAIN-MAPPING: Red-de-sensores ,
                  Localización-espacio-temporal;
END KNOWLEDGE-ROLE Datos-de-un-sensor ;

KNOWLEDGE-ROLE Componentes-requeridos-de-la-hipótesis ;
  TYPE: DYNAMIC;
  DOMAIN-MAPPING: Características-espacio/tiempo;
END KNOWLEDGE-ROLE Componentes-requeridos-de-la-hipótesis ;

KNOWLEDGE-ROLE Estructura-de-la-hipótesis ;
  TYPE: DYNAMIC;
  DOMAIN-MAPPING: Patrón-lingüístico ;
END KNOWLEDGE-ROLE Estructura-de-la-hipótesis ;

KNOWLEDGE-ROLE Modelo-de-generación ;
  TYPE: STATIC;
  DOMAIN-MAPPING: Selección-de-patrones-lingüísticos ;
END KNOWLEDGE-ROLE Modelo-de-generación ;
```

A.8. Inferencia *Seleccionar-fuente-de-datos*

Algoritmo A.8 Inferencia Seleccionar-fuente-de-datos.

```
INFERENCE Seleccionar-fuente-de-datos ;
  ROLES:
    INPUT: Componentes-requeridos ;
    OUTPUT: Consulta-en-fuente-de-datos ;
    STATIC: Estrategia-de-búsqueda ;
  SPECIFICATION:
    "Selección de las fuentes de datos para
    obtener información de acuerdo a las
    estrategias de búsqueda." ;
END INFERENCE Seleccionar-fuente-de-datos ;

KNOWLEDGE-ROLE Componentes-requeridos ;
  TYPE: DYNAMIC ;
  DOMAIN-MAPPING: Características-espacio-tiempo ;
END KNOWLEDGE-ROLE Componentes-requeridos ;

KNOWLEDGE-ROLE Consulta-en-fuente-de-datos ;
  TYPE: DYNAMIC ;
  DOMAIN-MAPPING: Parámetros-de-búsqueda-en-Open-Data ;
END KNOWLEDGE-ROLE Consulta-en-fuente-de-datos ;

KNOWLEDGE-ROLE Estrategia-de-búsqueda ;
  TYPE: STATIC ;
  DOMAIN-MAPPING: Criterios-de-selección-Open-Data ;
END KNOWLEDGE-ROLE Estrategia-de-búsqueda ;
```

A.9. Inferencia *Buscar-en-fuente-de-datos*

Algoritmo A.9 Inferencia Buscar-en-fuente-de-datos.

```
INFERENCE Buscar-en-fuente-de-datos ;
  ROLES:
    INPUT: Consulta-en-fuente-de-datos ;
    OUTPUT: Componentes-encontrados ,
            Éxito-de-búsqueda ;
  SPECIFICATION:
    "Acceso a fuentes de datos para obtener información.";
END INFERENCE Buscar-en-fuente-de-datos ;

KNOWLEDGE-ROLE Consulta-en-fuente-de-datos ;
  TYPE: DYNAMIC;
  DOMAIN-MAPPING: Parámetros-de-búsqueda-en-Open-Data ;
END KNOWLEDGE-ROLE Consulta-en-fuente-de-datos ;

KNOWLEDGE-ROLE Componentes-encontrados ;
  TYPE: DYNAMIC;
  DOMAIN-MAPPING: Referencias-espacio/temporales ;
END KNOWLEDGE-ROLE Componentes-encontrados ;

KNOWLEDGE-ROLE Éxito-de-búsqueda ;
  TYPE: DYNAMIC;
  DOMAIN-MAPPING: Éxito/fracaso ;
END KNOWLEDGE-ROLE Éxito-de-búsqueda ;
```

B. Datos para evaluación

El presente anexo describe el conjunto de datos que fue construido en este trabajo para ser utilizado en el proceso de evaluación. El conjunto de datos contiene 1.811 sensores distribuidos en 10 áreas geográficas. Los datos se obtuvieron de las bases de datos que se pueden consultar a través de la Web pública del Ministerio de Medio Ambiente. Para la construcción de dicho conjunto de datos de evaluación fue necesario limpiar y filtrar los datos originales a fin de seleccionar la información útil para el problema abordado. Por ejemplo, se eliminaron los atributos no útiles (identificadores específicos o flags relacionados con el estado de funcionamiento). El conjunto de datos (publicado como material adicional en [Molina et al., 2015]²¹) es un archivo *.csv* donde cada línea corresponde a un sensor, y hay cinco atributos para cada sensor:

- *Latitud y longitud*: Coordenadas geográficas de la ubicación del sensor (sistema de referencia geodésico WSG84).
- *Medida*: Tipo de unidad física medida por el sensor, como caudal, lluvia y nivel.
- *Área*: El área geográfica del sensor. Descrita por el nombre de la cuenca principal del río (por ejemplo, cuenca del Ebro, cuenca del Duero, etc.).
- *Descripción*: Descripción textual completa escrita (en español) por expertos en el dominio para el punto geográfico donde se ubica el sensor (por ejemplo, el río Ebro en Ascó).

En las próximas páginas se muestran fragmentos del conjunto de datos.

²¹El dataset completo se encuentra publicado en <http://www.mdpi.com/1424-8220/15/7/16009/s1> (último acceso 03/11/2015).

B DATOS PARA EVALUACIÓN

```

Latitude;Longitude;Measure;Area;Description
41.26657104492190;1.63479590415955;Water flow;Catalonian basins;Castellet i la Gornal
42.42694091796880;3.12594079971313;water level;Catalonian basins;embalse de Portbou
42.42694091796880;3.12594079971313;Rainfall;Catalonian basins;embalse de Portbou
42.34429931640630;2.83564901351929;water level;Catalonian basins;embalse de Boadella
42.34429931640630;2.83564901351929;Rainfall;Catalonian basins;embalse de Boadella
42.31670379638670;2.91882729530334;Rainfall;Catalonian basins;Pont de Molins
42.30522918701170;3.00934958457947;Rainfall;Catalonian basins;Peralada
41.99152374267580;2.81980729103088;Water flow;Catalonian basins;río Ter en Girona
42.23857116699220;2.96902012825012;Rainfall;Catalonian basins;Santa Llogaia d'Àlguema
42.25567626953130;-2.92679023742676;Water flow;Catalonian basins;Castelló d'Empúries
...
36.88582992553710;-4.84897994995117;Water flow;Andalusian basins;río Turón en Ardales
36.76002502441410;-4.42910385131836;Rainfall;Andalusian basins;embalse del Limonero
36.81083679199220;-3.69523096084595;Rainfall;Andalusian basins;Cazulas
36.85853576660160;-3.49475622177124;water level;Andalusian basins;embalse de Rules
36.90473937988280;-4.11313533782958;Rainfall;Andalusian basins;Alcaucín
36.73283386230470;-4.46367311477661;Rainfall;Andalusian basins;depuradora del Atabal
36.56781005859380;-5.24455213546753;Rainfall;Andalusian basins;Jubrique
36.88582992553710;-4.84897994995117;Rainfall;Andalusian basins;Ardales
37.35001373291020;-2.51495289802551;Water flow;Andalusian basins;río Almanzora en Serón
36.51770782470700;-4.94653081893921;Rainfall;Andalusian basins;depuradora de Marbella
...
41.50585937500000;-3.09885096549988;Water flow;Duero basin;Gormaz
41.50585937500000;-3.09885096549988;Rainfall;Duero basin;Gormaz
41.64498519897460;-3.57276439666748;Rainfall;Duero basin;Vadocondes
41.62887191772460;-4.36311531066895;Rainfall;Duero basin;Quintanilla de Onésimo
41.50595092773440;-4.92784929275513;Rainfall;Duero basin;San Miguel del Pino
441.51664733886720;-5.41107559204102;Rainfall;Duero basin;Toro
49.49980163574220;-5.79678201675415;Rainfall;Duero basin;Zamora
41.76358032226560;-4.19574689865112;Rainfall;Duero basin;Torre de Esgueva
42.04283142089840;-3.27185106277466;Rainfall;Duero basin;Salas de los Infantes
42.05508804321290;-3.50755262374878;Rainfall;Duero basin;Covarrubias
...
42.68877029418950;-2.95428586006165;Water flow;Ebro basin;río Ebro en Miranda
42.18071746826170;-1.69385600090027;Water flow;Ebro basin;río Ebro en Castejón
42.37524032592770;-1.94773066043854;Water flow;Ebro basin;río Ega en Andosilla
42.31606292724610;-1.80004918575287;Water flow;Ebro basin;río Arga en Funes
42.34614562988280;-1.64661157131195;Water flow;Ebro basin;río Aragón en Caparroso
42.34614562988280;-1.64661157131195;Rainfall;Ebro basin;Caparroso
41.29877090454100;-1.96604382991791;Water flow;Ebro basin;río Jalón en Cetina
41.19888687133790;-1.79114449024200;Water flow;Ebro basin;río Piedra en Nuevalos
41.65900039672850;-0.87953668832779;Water flow;Ebro basin;río Ebro en Zaragoza
42.17224121093750;-0.75167781114578;Water flow;Ebro basin;río Gállego en Ardisa
...
37.39892578125000;-3.72151589393616;Rainfall;Guadalquivir basin;embalse Colomera
37.39892578125000;-3.72151589393616;water level;Guadalquivir basin;embalse Colomera
37.27720642089840;-3.68198013305664;Rainfall;Guadalquivir basin;embalse Cubillas
37.27720642089840;-3.68198013305664;water level;Guadalquivir basin;embalse Cubillas
37.66992950439450;-6.13210630416870;Rainfall;Guadalquivir basin;Cala
37.69474029541020;-3.81778740882874;Rainfall;Guadalquivir basin;Villares
38.55087661743160;-3.29027843475342;Rainfall;Guadalquivir basin;Castel
37.75998306274410;-3.75113010406494;Rainfall;Guadalquivir basin;Jontoya
37.76260757446290;-2.90276265144348;Rainfall;Guadalquivir basin;embalse La Bolera
37.76260757446290;-2.90276265144348;water level;Guadalquivir basin;embalse La Bolera
...

```

B DATOS PARA EVALUACIÓN

Latitude; Longitude; Measure; Area; Description

```

39.36283493041990;-4.25751829147339;Rainfall;Guadiana basin;embalse Torre Abrahan
39.10334396362300;-4.58258485794067;Rainfall;Guadiana basin;Puebla de Don Rodrigo
38.54568481445310;-5.01039075851440;Rainfall;Guadiana basin;embalse de la Colada
38.71068572998050;-4.93442678451538;Rainfall;Guadiana basin;Guadalmaz
39.24313735961910;-5.73975563049316;Rainfall;Guadiana basin;embalse Alcollarín
38.88984680175780;-6.18858861923218;Rainfall;Guadiana basin;Valverde de Mérida
38.96512222290040;-6.92736673355103;Water flow;Guadiana basin;río Gévora con Guadiana
38.86904525756840;-6.89716958999634;Rainfall;Guadiana basin;Badajoz
38.82834243774410;-6.95840167999268;Water flow;Guadiana basin;embalse Calamón
37.95225906372070;-7.21112012863159;Rainfall;Guadiana basin;embalse de Chanza

39.68527221679690;-0.47882825136185;Rainfall;Jucar basin;Portaceli
39.95308303833010;-2.06985998153687;Rainfall;Jucar basin;Talayuelo
40.67184829711910;0.23433916270733;Rainfall;Jucar basin;Ulldecona
40.58526611328130;0.00215020333417;Rainfall;Jucar basin;Vallibona
39.47591781616210;-0.58072513341904;Water flow;Jucar basin;Rambla Pollo
39.72780227661130;-1.09880340099335;water level;Jucar basin;embalse de Benageber
38.81465911865230;-0.35326674580574;water level;Jucar basin;embalse de Beniarres
39.26358032226560;-1.09513926506042;Water flow;Jucar basin;Cofrentes
38.83413696289060;-0.19578149914742;Water flow;Jucar basin;Gallinera
39.56562423706050;-2.11062097549438;water level;Jucar basin;embalse de Alarcón

42.53473281860350;-6.95513963699341;Rainfall;Northern basins;Oencia
42.16755294799800;-6.89488410949707;Water flow;Northern basins;Porto
42.16755294799800;-6.89488410949707;Rainfall;Northern basins;Porto
42.34683990478520;-7.87428712844849;Water flow;Northern basins;río Miño en Orense
42.37385177612300;-8.11734867095947;Rainfall;Northern basins;Leiro
42.29133224487300;-8.14065551757813;Water flow;Northern basins;Ribadavia
42.29133224487300;-8.14065551757813;Rainfall;Northern basins;Ribadavia
42.63006591796880;-7.71106672286987;water level;Northern basins;embalse de Belesar
42.63006591796880;-7.71106672286987;Rainfall;Northern basins;embalse de Belesar
42.46666336059570;-7.72201728820801;water level;Northern basins;embalse de los Peares

38.18712615966800;-0.98313564062119;Rainfall;Segura basin;depuradora de Torrealta
37.90378570556640;-0.79578793048859;Rainfall;Segura basin;Cartagena
38.13017654418950;-1.30245447158813;Water flow;Segura basin;río Segura en Archena
37.68824386596670;-1.99499607086182;Rainfall;Segura basin;Venta del Castillo
38.23917388916020;-1.55495226383209;Rainfall;Segura basin;Almadenes
38.23917388916020;-1.55495226383209;Water flow;Segura basin;río Segura en Almadenes
38.09373855590820;-0.66071718931198;Water flow;Segura basin;río Segura en Guardamar
38.11001968383790;-2.03693270683289;Rainfall;Segura basin;Casas de Navarro
38.04980850219730;-1.00497210025787;Rainfall;Segura basin;Beniel
38.04980850219730;-1.00497210025787;Water flow;Segura basin;río Segura en Beniel

40.60717010498050;-1.77487635612488;Rainfall;Tajo basin;Fuente del Hocinillo
40.48529434204100;-6.00540208816528;Water flow;Tajo basin;río Alagon en Garcibuey
40.11316299438480;-5.94442319869995;Rainfall;Tajo basin;río Ambroz en el Villar
40.75443267822270;-3.99327898025513;Rainfall;Tajo basin;Barranca Navacerrada
40.62753295898440;-2.08663725852966;Rainfall;Tajo basin;Valsalobre
41.02186203002920;-3.70113587379456;Rainfall;Tajo basin;Villavieja
40.36899948120120;-4.83087730407715;Rainfall;Tajo basin;Casas de los Veneros
40.54559326171880;-4.58475255966187;Rainfall;Tajo basin;Santa Cruz de Pinares
40.09896469116210;-5.84587335586548;Rainfall;Tajo basin;Piornal
40.41117858886720;-5.76708364486694;Rainfall;Tajo basin;Valdesangil

```