Descubrimiento automático de *mappings* en un caso de uso real con altas exigencias de certeza

J. A. Ramos[1], M. Fernández-López[2], A. Gómez-Pérez[1]

[1] Ontology Engineering Group – Universidad Politécnica de Madrid
Avda. Montepríncipe s/n – 28660 – Boadilla del Monte, Madrid, Spain
{jarg, asun}@fi.upm.es

[2] Escuela Politécnica Superior. Universidad San Pablo CEU
Ctra. de Boadilla del Monte km 5,300
28668 - Boadilla del Monte, Madrid, Spain
mfernandez.eps@ceu.es

Resumen. Los sistemas de integración de información resuelven las diferencias entre las fuentes, en la mayoría de los casos, mediante la creación de *mappings*, puentes semánticos entre los elementos de las fuentes. Hasta ahora se han propuesto comparadores para generar un conjunto de *mappings* para cada par de elementos de las fuentes a integrar, y se han realizado estudios experimentales con ellos. El valor añadido del presente trabajo frente a los trabajos experimentales anteriores es que se ha llevado a cabo en un caso real embebido en una aplicación real (en el dominio geográfico) con altas exigencias de certeza.

Palabras clave: web semántica, gestión del conocimiento, mappings, comparador de elementos, información geográfica.

1 Introducción

El clásico problema de la heterogeneidad se ha visto parcialmente resuelto desde hace tiempo con un progresivo aumento de metodologías y herramientas de integración de información, tanto a nivel de esquemas como a nivel de datos [1; 2; 3]. Estos sistemas en su mayoría representan las semejanzas entre las fuentes de información mediante el establecimiento de *mappings*, puentes semánticos entre los elementos de las fuentes. Uno de los retos actuales es el descubrimiento de estos *mappings*, un proceso del que dependerá la calidad de la integración total del sistema.

Dentro del proceso de descubrimiento de *mappings* [4], el principal desarrollo hasta el momento se ha producido en la fase de generación de *mappings*, donde se descubren las relaciones entre los elementos de las fuentes. El resto de las fases están dedicadas fundamentalmente a la depuración y agregación de *mappings*. En la fase de generación es donde ejecutan los comparadores, que calculan la relación (similitud, distancia semántica, generalización/especialización, etc.) entre cada par de elementos y un grado de certeza en que tal relación exista realmente. Se pueden encontrar herramientas de descubrimiento automático de *mappings* (véanse las referencias del

portal Ontology Matching[1]) y un conjunto de ellas anualmente compite dentro de la iniciativa denominada OAEI[2]. Cada herramienta utiliza un conjunto de comparadores y una combinación de comparadores para obtener los mejores resultados.

Los comparadores se basan en múltiples formas de cálculo de similitud: técnicas de comparación de etiquetas (comparación de cadena como distancia de edición [5], cálculo de distancia "semántica" en recursos externos [6], comparación con técnicas de lenguaje natural [7], etc.), comparación de estructuras (cuando las fuentes son estructuradas, como ontologías o tesauros [8]), comparación con técnicas ad hoc del dominio, etc.

Las evaluaciones que se hacen de estos comparadores, tanto en la citada OAEI como en otros trabajos (por ejemplo, [9] y [10]), se realizan con conjuntos reales (algunos son conjuntos públicos de portales webs, por ejemplo). Sin embargo, no se consideran las características de una aplicación real con altas exigencias de certeza en un dominio bastante específico y, aunque elaborado por expertos, no estandarizado.

En este artículo se presenta un caso real de descubrimiento de *mappings*. Como quedará demostrado, las características del caso limitan la eficacia de los comparadores más habituales y serán las técnicas ad hoc para el dominio las que supongan un incremento en el número de *mappings* descubiertos.

En el epígrafe 2 se muestra una serie de definiciones previas. En el 3, se describe el caso de uso y las características que lo hacen interesante. En el epígrafe 4 se detalla el análisis del descubrimiento: comparadores útiles y comparadores ad hoc a desarrollar. En el epígrafe 5 se muestran los resultados de la evaluación de los comparadores del dominio. En la sección 6 se presentan las conclusiones y el trabajo futuro.

2 Definiciones previas

Definición 1. Un *mapping* es una 5-upla ⟨`id`, e_1, e_2, R, c⟩, donde `id` es un identificador, e_1 es el elemento origen, e_2 es el elemento destino, R es la relación que existe entre ellos (subclase de, equivalencia, etc.), y c es el grado de certeza que se tiene sobre el *mapping*. La no existencia de mapping entre los elementos e_1 y e_2 se representa como ⟨`id`, e_1, e_2, ∅, c⟩.

Ejemplo 1. ⟨123, `Aeroclub`, `Aeródromo`, ≤, 0,9⟩ significa que `Areoclub` es subclase de `Aeródromo` con un grado de certeza de 0,9.

Definición 2. Un *comparador* es una función que, a partir de dos elementos e_1 y e_2, devuelve un *mapping* entre e_1 y e_2.

Definición 3. Un *comparador estricto de cadena* es una función que, a partir de un par de elementos e_1 y e_2 etiquetados con las cadenas t_1 y t_2, devuelve ⟨i,e_1,e_2,≡,1⟩ si t_1 es exactamente igual que t_2, y ⟨i,e_1,e_2,∅,1⟩ en otro caso. A lo largo de este artículo, este comparador será identificado con el número **1**.

Ejemplo 1. Si t_1="`cauce artificial`" y t_2="`cauce artificial`", devuelve ⟨i,e_1,e_2,≡,1⟩.

[1] http://www.ontologymatching.org/
[2] http://oaei.ontologymatching.org/

Ejemplo 2. Si t_1="`Cauce artificial`" y t_2="`cauce artificial`" (es decir, hay una diferencia en una mayúscula), devuelve $\langle i,e_1,e_2,\varnothing,1\rangle$.

Definición 3. Un *normalizador* de cadena es una función que, a partir de una palabra, obtiene otra donde se han cambiado mayúsculas por minúsculas, se han eliminado guiones, etc.

Definición 4. Dado un normalizador de cadena σ, un *comparador laxo de cadena* es una función μ_2 que, a partir de un par de elementos e_1 y e_2 etiquetados con las cadenas t_1 y t_2, devuelve $\langle i,e_1,e_2,\equiv,1\rangle$ si $\sigma(t_1)=\sigma(t_2)$, y $\langle i,e_1,e_2,\varnothing,1\rangle$ en otro caso. A lo largo de este artículo, este comparador será identificado con el número **2**.

Ejemplo 3. Si t_1="`autovía`" y t_2="`Autovía-`" y la función de normalización de cadena convierte toda la cadena a minúsculas, elimina guiones, etc., entonces el comparador devuelve $\langle i,e_1,e_2,\equiv,1\rangle$.

Definición 5. Dado un normalizador de cadena σ, un *comparador de inclusión de cadenas* es una función μ_3 que, a partir de un par de elementos e_1 y e_2 etiquetados con las cadenas t_1 y t_2, devuelve $\langle i,e_1,e_2,\geq,0,7\rangle$ si $\sigma(t_1)$ es una subcadena de $\sigma(t_2)$, y nada en otro caso. A lo largo de este artículo, este comparador será identificado con el número **3**.

Ejemplo 4. Si t_1="`autovía`" y t_2="`Autovía en construcción`", entonces el comparador devuelve $\langle i,e_1,e_2,\geq,0,7\rangle$.

2 Descripción del caso de uso

El Grupo de Ingeniería Ontológica (Ontology Engineering Group[3]) está colaborando en los últimos años con el Instituto Geográfico Nacional (IGN[4]), dependiente del Ministerio de Fomento. Esta colaboración tiene como objeto trabajar en varias líneas de investigación (generación de ontologías a partir de bases de datos, integración de servicios de cartografía, etc.) dentro del dominio de la información geográfica y geoespacial. La principal línea de investigación es la integración de fuentes de información geográfica. Esta línea de investigación es de especial relevancia, como demuestra INSPIRE[5] (Iniciativa de la Comisión Europea) cuyo funcionamiento se recoge en la Directiva 2007/2/CE del Parlamento Europeo y del Consejo que tiene como objetivo la creación de una Infraestructura de Datos Espaciales en Europa. La Directiva establece los objetivos, y los Estados miembros tendrán dos años desde su publicación para ajustar sus respectivas legislaciones y procedimientos administrativos nacionales.

El escenario que se plantea para llevar a cabo esta integración aglutina una gran cantidad de características y lo convierten en un caso muy rico para llevar a cabo la experimentación del presente trabajo. Así, en este escenario se dan multitud de tipos

[3] http://www.oeg-upm.net/
[4] http://www.ign.es/
[5] http://inspire.jrc.ec.europa.eu/

de heterogeneidad: entre representaciones del conocimiento diferentes, entre modelos, entre variantes léxicas de un término, etc. (se puede consultar más sobre heterogeneidad semántica propia de la información geo-espacial en [11; 12; 13; 14]).

El IGN es el productor oficial de información geográfica en el ámbito nacional. Tiene cuatro bases de datos que se corresponden con cuatro escalas diferentes: Nomenclátor Conciso (NC) (1:1.000.000), Nomenclátor Geográfico Nacional (NGN) (1:50.000), Base Cartográfica Numérica (BCN200) (1:200.000) y Base Cartográfica Numérica (BCN25) (1:25.000). Estas bases de datos son mantenidas separadamente y presentan gran heterogeneidad en diferentes aspectos.

El sistema para la integración de estas fuentes de modo que se establezca un vocabulario único se realizó mediante la construcción de la ontología pública PhenomenOntology[6], que se alinea, mediante *mappings*, con cada una de las fuentes a través del conjunto de fenómenos geográficos de cada una de ellas (más detalles en [15]). Durante la explotación del sistema integrado, las consultas se realizan utilizando el vocabulario definido en la ontología, y las respuestas incluyen enlaces a las fuentes del IGN. Tales enlaces deben ser exactos. Es decir, no caben respuestas aproximadas.

A la hora de evaluar comparadores, las características más relevantes del caso de uso son las que se muestran a continuación:

1. Hay fuentes no estructuradas y otras con estructura representada en los identificadores de los elementos. Este último caso se da en aquellas fuentes que cumplen la propiedad de que un elemento es subclase de otro si y sólo si el código del segundo es una subcadena del primero.
2. Los términos de las fuentes son muy específicos del dominio geográfico.
3. Las fuentes han sido elaboradas y revisadas por expertos.
4. La terminología utilizada no se considera como estándar.
5. Dado que los enlaces han de ser exactos, los umbrales de certeza en el filtrado de *mappings* serán cercanos a 1 (0,9 y 0,95 como valores candidatos mínimos según los expertos del dominio consultados).
6. Las fuentes están en español.

3 Selección de comparadores para el caso de uso

La selección de comparadores pasa por un análisis sencillo de las características del dominio.

Hay etiquetas representativas, por tanto, el comparador 1 y el comparador 2 son aplicables.

Las etiquetas de las fuentes están revisadas. Por tanto, los comparadores que se basan en cálculo de distancias de edición (principalmente orientados a salvar errores ortográficos y palabras parecidas para un mismo concepto) van a aportar al conjunto de *mappings*, principalmente, *mappings* inválidos (por tener grados de certezas medios y bajos). Aunque se han realizado experimentos para validar esta afirmación, no se presentan aquí por motivos de espacio.

[6] http://mayor2.dia.fi.upm.es/oeg-upm/files/phenomontology/Phenom4.0.1.owl

Las etiquetas de los términos del caso de uso no se encuentran normalmente en los diccionarios electrónicos generalistas, como de otros dominios que sí que aparecen. Para demostrar esta afirmación, se han consultado en la última versión que se dispone de EuroWordNet [16] los términos de las fuentes a mapear:
- NC: 22 términos buscados, 14 encontrados (63%).
- NGN: 52 términos buscados, 8 encontrado (15%).
- BCN25: 441 términos buscados, 35 encontrados (8%).
- BCN200: 511 términos buscados, 65 encontrados (13%).

En total, de 1.026 términos buscados se han encontrado 122 (12%). Con tan baja tasa de aparición de los términos, se puede deducir que los resultados de las comparaciones basadas en el uso de recursos generalistas externos no van a suponer aportes importantes. Esto afecta tanto a los comparadores que realizan consultas semánticas (basadas en los *synset* de EWN) como a los que establecen relaciones a partir de las posiciones relativas dentro del conjunto ordenado de conceptos de EWN.

No existen diccionarios electrónicos del dominio (tesauros u ontologías) públicos y en español (de hecho, es intención del IGN que PhenomenOntology sea el vocabulario de referencia para las diferentes agencias geográficas, llenando ese vacío). Sin embargo, el IGN sí dispone de dos diccionarios, dando definiciones a muchos de los términos de dos de las fuentes del caso de uso: de NC con 22 definiciones y de BCN25 con 366 definiciones. Las definiciones no están realizadas por terminólogos, sino por expertos del dominio. Al disponer de estos recursos, se decidió el desarrollo de un comparador basado en información semántica extraída de estos diccionarios.

Tras el análisis del conjunto de definiciones de dominio en lenguaje natural que el IGN ha proporcionado, se observó que se pueden descubrir *mappings* mediante el comparador que se define a continuación:

Definición 6. Un *comparador de primera palabra* es aquél que devuelve $\langle i,e_1,e_2,\leq,0,9\rangle$ si el primer sustantivo que aparece en la definición de e_1 es e_2, y $\langle i,e_1,e_2,\varnothing,1\rangle$ en otro caso. A lo largo de este artículo, este comparador será identificado con el número **4**.

Ejemplo 5. Dada la definición `Aeroclub: Aeródromo para uso particular de socios afiliados`, el *mapping* generado es $\langle id$, `Aeroclub`, `Aeródromo`, $\leq$, $0,9\rangle$.

Hay otras técnicas posibles basadas en definiciones, pero se han dejado para líneas futuras de experimentación.

En las bases cartográficas numéricas (BCN) existe, como información de cada fenómeno, un código numérico que funciona de identificador único y global para todos los catálogos. Se usará esta información para obtener *mappings* de sinonimia e hiponimia. Así, se identifica un *mapping* de equivalencia entre `064401 Vías de estación de FFCC.Vía de servicio` (BCN25) y `064401 FFCC.VIA_DE_SERVICIO` (BCN200), y un *mapping* de *subclase estricta de* entre `064401 FFCC.VIA_DE_SERVICIO` (BCN200) y `06 (Comunicaciones)` (BTN25). Debido a las restricciones de espacio, no se proporcionará una definición semi-formal de este comparador. Se identificará con el número **5**.

También se pueden aprovechar los valores de las propiedades enumeradas. Un ejemplo del descubrimiento siguiendo este enfoque sería el que se muestra a continuación: sean el concepto `muro` de la ontología *o* y sea el término `muro en ruinas` de la fuente *s*. Si en `muro` existe la propiedad `estado`, cuyos valores predefinidos son `en construcción`, `en uso` y `en ruinas`, entonces se genera un *mapping*.

⟨*id*, `muro` ∩ ∃ `estado.{en_ruinas}`, `muro_en_ruinas`, ≡, 0,91⟩.

También por razones de espacio, tampoco se proporcionará una definición semi-formal para este comparador. Se identificará como número **6**.

Estas tres técnicas presentadas, asignan unos valores de certeza fruto de una evaluación empírica sobre su precisión en un pequeño conjunto de datos.

4 Evaluación

Para la evaluación en el descubrimiento automático de *mappings* los investigadores normalmente usan dos medidas: precisión y exhaustividad [17; 18; 19]. Ambas son medidas basadas en conjuntos y son calculadas teniendo en cuenta como base un conjunto de resultados correctos. Este conjunto, denominado conjunto de referencia o *gold-standard*, normalmente se construye de manera manual y es completo y correcto. Así, las herramientas de descubrimiento a evaluar producen un conjunto de *mappings* (un alineamiento) a partir de dos fuentes y el resultado es comparado con los *mappings* del conjunto de referencia entre esas dos mismas fuentes. Para cada combinación de fuentes y cada conjunto de relaciones existe un conjunto de referencia diferente. Esta comparación genera el valor de la precisión (*mappings* descubiertos que aparecen en el conjunto de referencia respecto al total de *mappings* descubiertos) y la exhaustividad (*mappings* descubiertos que aparecen en el conjunto de referencia respecto al total de *mappings* en el conjunto de referencia). Dada la gran cantidad de potenciales *mappings* (millones en algunas parejas de fuentes aquí utilizadas), la selección de *mappings* de referencia se ha realizado mediante muestreo utilizando sólo algunas de las ramas de cada taxonomía. Para este caso concreto, el muestreo no se ha realizado estrictamente, sino que se ha tomado el subconjunto del que se disponía de conjuntos de referencia suficientes como para poder realizar la evaluación.

Como grupo de control se utilizaron los comparadores 1, 2 y 3 ejecutándolos secuencialmente. Así, se evaluará la variación (aumento o disminución) de las medidas respecto a las obtenidas con el grupo de control.

Evaluación del valor añadido del comparador 4: primera palabra de la definición. Se diferencian aquí dos escenarios: sólo relaciones de equivalencia (véase la Tabla 1) (lo más comúnmente descubierto por el resto de herramientas), y relaciones de equivalencia y subsunción (subclase de y superclase de) (véase la Tabla **2**). Se puede apreciar que, en el escenario 1, **la precisión se mantiene** y en todos los casos **aumenta la exhaustividad**, mientras que, en el escenario 2, **tanto la precisión como la exhaustividad aumentan** en todos los casos.

Tabla 1. Comparador 4, sólo equivalencia

Fuente 1	**Fuente 2**	**Nº *mappings***	**Precisión**	**Exhaus-tividad**	**Compa-radores**	**Rela-ciones**
NC	Phe4.01	1	1,00	0,50	*1, 2, 3*	=
Phe4.01	NC	2	**1,00**	**1,00**	*1, 2, 3, 4*	=
NGN	Phe4.01	1	1,00	0,33	*1, 2, 3*	=
Phe4.01	NGN	3	**1,00**	**1,00**	*1, 2, 3, 4*	=
Phe4.01	BCN200	14	1,00	0,70	*1, 2, 3*	=
Phe4.01	BCN200	15	**1,00**	**0,75**	*1, 2, 3, 4*	=
Phe4.01	BCN25	17	1,00	0,36	*1, 2, 3*	=
Phe4.01	BCN25	23	**1,00**	**0,39**	*1, 2, 3, 4*	=

Tabla 2. Comparador 4, equivalencia y subsunciones

Fuente 1	**Fuente 2**	**Nº *mappings***	**Precisión**	**Exhaus-tividad**	**Compa-Radores**	**Rela-ciones**
Phe4.01	NC	2	0,50	0,50	*1, 2, 3*	*=, sc y SC*
Phe4.01	NC	3	**0,67**	**1,00**	*1, 2, 3, 4*	*=, sc y SC*
Phe4.01	NGN	6	0,50	0,43	*1, 2, 3*	*=, sc y SC*
Phe4.01	NGN	8	**0,63**	**0,71**	*1, 2, 3, 4*	*=, sc y SC*
Phe4.01	BCN200	83	0,18	0,23	*1, 2, 3*	*=, sc y SC*
Phe4.01	BCN200	84	**0,19**	**0,25**	*1, 2, 3, 4*	*=, sc y SC*
Phe4.01	BCN25	53	0,34	0,32	*1, 2, 3*	*=, sc y SC*
Phe4.01	BCN25	61	**0,39**	**0,42**	*1, 2, 3, 4*	*=, sc y SC*

Evaluación del valor añadido del comparador 5: basado en códigos identificadores. En la Tabla **3** se puede apreciar que **la precisión se mantiene** y **aumenta la exhaustividad** de manera significativa.

Tabla 3. Comparador 5

Fuente 1	**Fuente 2**	**Nº *mappings***	**Precisión**	**Exhaus-tividad**	**Compa-radores**	**Rela-ciones**
BCN200	BCN25	9	1,00	0,17	*1, 2, 3*	=
BCN200	BCN25	52	**1,00**	**0,96**	*1, 2, 3, 5*	=

Evaluación del valor añadido del comparador 6: basado en atributos con tipos enumerados. En la Tabla 4 se puede apreciar que la precisión y la exhaustividad se mantienen en los casos de los nomenclátores (NC y NGN), en el caso de BCN200 disminuyen ambas medidas y en el caso de BCN25 se mantiene la precisión y disminuye la exhaustividad.

Tabla 4. Comparador 6

Fuente 1	**Fuente 2**	**Nº *mappings***	**Precisión**	**Exhaus-tividad**	**Compa-radores**	**Rela-ciones**
NC	Phe4.01	1	1,00	0,50	*1, 2, 3*	=
Phe4.01	NC	1	**1,00**	**0,50**	*1, 2, 3, 6*	=
NGN	Phe4.01	1	1,00	0,33	*1, 2, 3*	=
Phe4.01	NGN	1	**1,00**	**0,33**	*1, 2, 3, 6*	=
Phe4.01	BCN200	14	1,00	0,70	*1, 2, 3*	=
Phe4.01	BCN200	36	**0,39**	**0,13**	*1, 2, 3, 6*	=
Phe4.01	BCN25	17	1,00	0,36	*1, 2, 3*	=
Phe4.01	BCN25	25	**1,00**	**0,29**	*1, 2, 3, 6*	=

Tabla 5. Comparador 6. Detallado

Fuente 1	**Fuente 2**	**Nº *map.***	**Refe-rencia**	**Acier-tos**	**Fallos**	**Ausen-cias**	**Preci-sión**	**Exhaus-tividad**	**Compa-radores**
Phe4.01	BCN200	14	20	14	0	6	1,00	0,70	*1, 2, 3*
Phe4.01	BCN200	36	105	14	22	91	0,39	0,13	*1, 2, 3, 6*
Phe4.01	BCN25	17	47	17	0	30	1,00	0,36	*1, 2, 3*
Phe4.01	BCN25	25	85	25	0	60	1,00	0,29	*1, 2, 3, 6*

Así pues el comparador 6 parece que no supone una mejora en cuanto a las técnicas no específicas del dominio. Para analizar más la razón de esta disminución, se van a detallar algunos datos de la evaluación y ejemplos concretos (véase la Tabla 5). Como se puede comprobar, el número de *mappings* de los conjuntos de referencia se incrementa drásticamente al pasar de elementos simples a elementos simples y conjuntos de elementos simples (de 20 a 105 en un caso y de 47 a 85 en el otro).

En el caso de BCN25 todos los *mappings* que descubre el comparador son correctos. El comparador descubre 8 nuevos *mappings* de los 38 nuevos *mappings* que podría descubrir. El comparador es muy preciso en este caso, pero poco exhaustivo y por ello baja el global de la exhaustividad de todos los comparadores. Por lo tanto, el comparador es preciso, pero menos exhaustivo que los del grupo de control.

En el caso de BCN200 todos los *mappings* que descubre el comparador (22) son incorrectos. La causa no es un error en el descubrimiento de los valores de los atributos en los nombres, sino la incompletitud del conjunto de valores identificados. Esto es debido a que en los nombres de BCN200 se usan abreviaturas para uno de los atributos y el comparador no las identifica. Un ejemplo de esto es el siguiente *mapping* descubierto:

⟨*id52*, `FFCC ELEC  ESTRECHO DOBLE`,
`FFCC` ∩ ∃ `ANCHO-DE-VIA.{ESTRECHO}` ∩
∃ `NUMERO-DE-VIAS.{DOBLE}`, ≡, 0,92⟩

Como se puede ver, el comparador identifica los valores ESTRECHO del atributo ANCHO DE VÍA y DOBLE del atributo NUMERO DE VÍAS, del concepto Ferrocarril convencional, que tiene como etiqueta FFCC. Sin embargo, el *mapping* no es correcto, puesto que faltaría por identificar en el conjunto de elementos simples el valor ELECTRIFICADO del atributo ELECTRIFICACIÓN. Este valor no se identifica porque en los nombres del catálogo BCN200 se usa ELEC para ELECTRIFICADO y el comparador no tiene información sobre posibles abreviaturas de valores. Igualmente, el texto NO ELEC no es identificado como el valor NO ELECTRIFICADO. Todos los *mappings* son incorrectos por este motivo.

5 Conclusiones y líneas futuras

En el presente artículo se ha presentado una evaluación de comparadores para un caso de uso real para una aplicación con altas exigencias de certeza.

Se ha podido comprobar que hay una serie de grupos de comparadores ya clásicos en la investigación sobre descubrimiento de *mappings* entre fuentes heterogéneas, como pueden ser los basados en diccionarios de términos comunes (como es el caso de EWN), o los basados en distancias de edición, que aportan un valor añadido muy escaso, o incluso son contraproducentes en este contexto.

Los comparadores que han permitido el descubrimiento de la mayor parte de los *mappings* han sido los basados en semejanzas entre cadenas (combinados con normalizaciones) y aquellos ad hoc para el dominio. En este último grupo están aquellos que utilizan un análisis sencillo de definiciones en lenguaje natural, los que aprovechan la información estructural que hay en los identificadores de los términos, y los basados en atributos de tipos enumerados.

Entre las líneas futuras está el uso de otras técnicas posibles basadas en definiciones. Por ejemplo, si se comparan los adjetivos de los sustantivos con los que comienzan las definiciones, se pueden extraer atributos que discriminan entre unas clases y otras. Este análisis podría ayudar posteriormente al descubrimiento de *mappings* mediante la comparación de atributos discriminantes.

También como línea futura se ha establecido el extender la experimentación a otros dominios diferentes del geográfico.

6 Agradecimientos

Este trabajo ha sido financiado parcialmente por el proyecto USP BS PPC05/2010 y por el proyecto “BabeLData: multingüismo en ontologías y Linked Data” del Ministerio de Ciencia e Innovación (TIN2010-17550).

7 Referencias

1. Batinti C, Lenzerini M, Navathe SB (1986) A Comparative Analysis of Methodologies for Database Schema Integration. ACM Computing Surveys, Vol. 18, No. 4, December 1986.
2. Lenzerini M (2002) Data integration: a theoretical perspective. /PODS '02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems/, pp. 233-246, New York, NY, USA. ACM Press
3. Beneventano D, Bergamaschi S (2004) The MOMIS Methodology for Integrating Hetrogeneous Data Sources. IFIP Congress Topical Sessions, 2004, pp. 19-24.
4. Euzenat J, Shvaiko P (2007) Ontology Matching. Springer Verlag, 2007.
5. Hahn U, Chater N (1997) Concepts and similarity. Knowledge, Concepts and Categories. Cambridge,Massachusetts. Physhology Press/MIT Press, 1997.
6. Jain P, Hitzler P, Sheth AP, Verma K, Yeh PZ (2010) Ontology Alignment for Linked Open Data. 9th International Semantic Web Conference (ISWC2010). Shanghai (China), 2010. LNCS 6496. Pp: 401-416.
7. Locoro A, Mascardi V, Scapolla AM (2010) NLP and Ontology Matching: A Successful Combination for Trialogical Learning. ICAART-2010. Pp: 253-258.
8. Niepert M, Meilicke C, Stuckenschmidt H (2010) A Probabilistic-Logical Framework for Ontology Matching. In Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI), Atlanta, Georgia, AAAI Press, 2010.
9. Cohen WW, Ravikumar P, Fienberg SE (2003) A Comparison of String Distance Metrics for Name-Matching Tasks. IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03). 2003. Pp: 73-78.
10. David J, Euzenat J (2008) Comparison between ontology distances. 7th International Semantic Web Conference (ISWC2008). LNCS 5318. Springer, 2008. Pp: 245–260.
11. Kuhn W (2002) Modeling the semantics of geographic categories through conceptual integration. Proceedings of GIScience, Boulder, Colorado, USA. 2002.
12. Mark DM, Smith B, Tversky B (1999) Ontology and Geographic Objects: An Empirical Study of Cognitive Categorization. Spatial Information Theory, Freksa, Chr., Mark, D.M., (eds.). Lecture Notes in Computer Science 1661, Springer, Berlin, New York, 199. Pp: 283-298.
13. Harvey F, Kuhn W, Pundt H, Bishr Y, Riedemann C (1999) Semantic interoperability: A central issue for sharing geographic information. The Annals of Regional Science, vol. 33, no. 2, 1999. Pp: 213-232.
14. Pundt H, Bishr Y (2002) Domain Ontologies for Data Sharing – An Example from Environmental Monitoring Using Field GIS. Computer & Geosciences, vol. 28. no. 1, 2002. Pp: 95-102.
15. Gómez-Pérez A, Ramos JA, Rodríguez-Pascual AF, Vilches-Blázquez LM 2008, 'The IGN-E case: Integrating through a hidden ontology', The 13th International Symposium on Spatial Data Handling (SDH 2008), June 23rd - 25th, 2008. Montpellier, France. Pages: 417-135. ISBN: 978-3-540-68565-4.
16. Vossen P (2004) EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. Semi-special issue on multilingual databases, IJL 17/2.
17. Do H, Melnik S, Rahm E (2002) Comparison of Schema Matching Evaluations. 2nd Int. Workshop on Web Databases (German Informatics Society), 2002. Pp: 221-237.
18. Ehrig M, Euzenat J (2005) Relaxed precision and recall for ontology matching. K-Cap 2005 workshop on Integrating ontology. Banff (Canada). 2005. Pp: 25–32.
19. Euzenat J (2007) Semantic Precision and Recall for Ontology Alignment Evaluation. IJCAI 2007. Pp: 348-353.