

# Anomalías en ontologías provisionales

Joaquín Borrego Díaz<sup>1</sup>, Antonia M<sup>a</sup> Chávez González<sup>2</sup>

<sup>1</sup>Dpto. Ciencias de la Computación e Inteligencia Artificial Universidad de Sevilla, Sevilla, España. tchavez@us.es

<sup>2</sup>Dpto. Ciencias de la Computación e Inteligencia Artificial Universidad de Sevilla, Sevilla, España. jborrego@us.es

## Resumen

*Los formalismos de representación del conocimiento proporcionan estructuras para organizar el conocimiento, pero no mecanismos para compartirlo. Una ontología en sí es el resultado de formalizar una conceptualización compartida acerca de un dominio de conocimiento. Sin embargo, las ontologías están en continuo cambio, es difícil considerarlas “terminadas” o cerradas, pues es necesario acceder al conocimiento que tienen como base, el conocimiento al que representan y, en general, dicho conocimiento es susceptible de evolucionar, haciéndose más preciso y definido. Es ese carácter dinámico el que hace necesaria la revisión de forma permanente, (siendo posible, además, que dichas revisiones no siempre las realice una misma persona). El objetivo de este trabajo es la formalización lógica de algunas anomalías de tipo léxico-lógico.*

**Palabras clave:** Inconsistencia, Ontología, Representación del conocimiento, Revisión ontológica.

## Abstract

*Formalisms for knowledge representation provide structures to manage knowledge but not to share it. Ontology is a formalization of a sharing conceptualization about a specific domain. However, ontologies are not static, but evolve over time. It is difficult to consider them as closed or finished because the piece of knowledge they represent is evolving making itself more precise, and changes in the conceptualization require modifications of the ontology. This dynamic property of knowledge makes necessary revisions (that can be realized by different users). At the same time, ontology contents can be reused by applications which, sometimes, make knowledge approachless. The aim of the paper is to formalize some logical-lexical anomalies.*

**Keywords:** Inconsistency, Knowledge representation, Ontology, Ontology revision.

## 1 Introducción

Antes de utilizar cualquier aplicación (de análisis o manipulación) sobre una base de datos, es necesario corregir o, en su caso, eliminar los posibles datos ilícitos. Tenemos que disponer de una clasificación de los tipos de datos ilícitos para proceder de una forma u otra a su corrección, o bien de una métrica para evaluar la calidad de los mismos. Sin ellas no es posible medir el grado de confianza en los resultados que se deriven del uso de una aplicación concreta.

En (Kim, 2003), los autores presentan una clasificación muy completa e, inicialmente, definen los datos ilícitos como aquellos de los que el usuario (o la aplicación en cuestión) obtiene resultados erróneos, o bien no es posible derivar un resultado, debido a problemas inherentes a los datos en sí. Las fuentes de datos ilícitos incluyen por un lado entradas y salidas erróneas y, por otro, errores en la transmisión y el procesamiento de los mismos.

Una ontología en sí es el resultado de formalizar una conceptualización compartida acerca de un dominio de conocimiento. Sin embargo, las ontologías están en continuo cambio, es difícil considerarlas “terminadas” o cerradas, pues es necesario acceder al conocimiento que tienen como base, el conocimiento al que representan y, en general, dicho conocimiento es susceptible de evolucionar, haciéndose más preciso y definido. Es ese carácter dinámico el que hace necesario la revisión de forma permanente, (siendo posible, además, que dichas revisiones no siempre las realice una misma persona). Una fuente importante de revisión ontológica proviene de las anomalías.

Las anomalías se deben a diversas razones, entre las que destacamos las siguientes<sup>1</sup>:

- El conjunto de datos es inconsistente con el dominio de conocimiento, debido a las inconsistencias formales producidas por datos incorrectos.
- La base de datos no está nunca completa, esto es, el usuario puede seguir introduciendo datos.
- Es posible que la componente intensional de la base de datos, la ontología, no refleje las creencias del usuario y, por tanto, sean necesarias nuevas actualizaciones. Esto ocurrirá frecuentemente en la WS.
- Es posible que algún criterio considerado en el ciclo de limpieza lleve a tomar una decisión que haga inconsistente la BC. Este tipo de errores, que pueden aparecer al gestionar la ontología visualmente.
- Los sistemas de integración de datos necesitan una reformulación o una interpretación lógica, ya que proporcionan acceso simultáneo a múltiples fuentes de información. Si, por ejemplo, la transformación no respeta criterios lógicos, podemos transformar una BC consistente en una inconsistente.
- Un algoritmo pregunta-respuesta, si resulta ser incompleto, puede no generar testigos de alguna restricción de integridad de carácter existencial. Un caso extremo aparece cuando lo que se usa es un demostrador automático de teoremas (DAT) para obtener respuestas (Alonso-Jiménez, 2002).

---

<sup>1</sup> Hemos destacado las que atañen al razonamiento automático con BC en DL al utilizar sistemas de razonamiento automático (SRA).

- La inconsistencia puede producirse en la especificación del lenguaje de la ontología (Fikes, 2002).
- El método elegido para el mantenimiento de la consistencia puede no ser lo suficientemente sólido y fallar con cierto tipo de actualizaciones.
- Los sistemas de minería de datos pueden producir salidas que no satisfacen las restricciones de integridad o los requerimientos de la ontología.
- A veces, la expresividad se ve limitada y no permite usar nuevas características en la representación usual del conocimiento. Esto implica que algunas definiciones sean incorrectas. Es un problema de pobreza del lenguaje de representación.
- Por último, una ontología pobre (poco detallada o poco cuidada) asociada a datos consistentes puede producir inconsistencias. Esto puede ocurrir en las fases iniciales de la vida de la ontología.

El objetivo de este trabajo es la formalización lógica de algunas anomalías de tipo léxico-lógico, siguiendo las ideas expuestas en (Borrego, 2006).

### 1.1 Incompletitud y anomalías

La incompletitud de una BC se debe entender de dos formas distintas: la incompletitud lógica (con respecto a cierto tipo de preguntas) y la incompletitud debida a la falta de conceptos o roles (incompletitud de carácter expresivo). En el problema de la verificación de BC en DL aparece, por supuesto, el problema de la inconsistencia lógica.

Existe cierto tipo de anomalías debido a la pobreza expresiva, las causadas por la carencia, desde el punto de vista del usuario, del perfil exacto de un determinado concepto. Cuando esto ocurre, el usuario trabaja con creencias sobre tales conceptos que aún no aparecen de forma explícita en la BC. A los conceptos de este tipo, existentes en las ontologías provisionales, los denominaremos *nociones*. La existencia de nociones en una ontología, provoca, además, que no sea posible distinguir dos conceptos amparados bajo la misma noción, es decir, que la ontología sea *gruesa*.

## 2 Clasificación de Anomalías en BC consistentes

El cálculo de conexión de regiones (RCC) es una teoría axiomatizada de primer orden para la representación de las relaciones entre entidades espaciales desde el punto de vista topológico (Cohn, 1997). La utilizaremos como herramienta del razonamiento espacial cualitativo para representar espacialmente (mediante regiones espaciales) el conocimiento, los conceptos de la ontología en cuestión. Interesa trabajar con un conjunto de relaciones base. La relación primaria es la conexión  $C(x; y)$ , que se interpreta como “*las clausuras de  $x$  e  $y$  se cortan*”. Las ocho relaciones básicas DC (desconexión), EC (conexión externa), PO (solapamiento parcial), EQ (igualdad), TPP (parte propia tangencial), NTPP (parte propia no tangencial), TPPi (inversa de TPP), NTPPi (inversa de NTPP) forman un conjunto exhaustivo de relaciones disjuntas dos a dos (JEPD) y constituyen el cálculo relacional RCC8, que ha sido estudiado en profundidad por J. Renz y B. Nebel en (Renz, 1999).

Suponemos conocida en todo momento la relación de RCC existente entre dos regiones cualesquiera. Sin embargo, hay que considerar que este hecho no es siempre cierto: en ocasiones, dicha relación no es explícita, lo cual no significa que el desconocimiento sea total, sino que no sabemos qué relación mantienen dos regiones de entre un conjunto determinado de posibilidades. Por otra parte, la extensión del conjunto de relaciones debe estar soportada

por una interpretación que extienda una clase de los modelos de la teoría inicial, para manejar dichas extensiones con un soporte semántico.

## 2.1 Introduciendo una relación de indefinición disjunta con RCC8

Para estudiar la posibilidad de expresar una relación de “indefinición”, introduciremos en RCC una nueva relación  $I(x, y)$  que expresaremos “*la relación entre  $x$  e  $y$  es desconocida (hasta cierto punto)*”. Además pretendemos que dicha relación sea disjunta con RCC8, pues, efectivamente, la indeterminación debe ser un hecho disjunto con la determinación exacta que proporcionan las relaciones de RCC8. De esta forma, puede utilizarse como parte de un cálculo que extienda a RCC8 (para que sea un sistema JEPD).

**Teorema. 2.1.1** *Existen ocho y sólo ocho extensiones  $E_{RCC}$ -conservativas del retículo RCC con una relación de indefinición que, junto con RCC8, forma un JEPD.*

Llamaremos  $I_{i, i=\{1, 2, \dots, 8\}}$  a las ocho relaciones de indefinición posibles que dan lugar a las ocho extensiones del retículo de relaciones de RCC. Podemos encontrar una representación de esas ocho extensiones en (Borrego, 2005b). A continuación analizamos y realizamos una clasificación de las anomalías utilizando las relaciones de indefinición insertables en la ontología RCC y su interpretación mediante regiones tipo “huevo-yema” (Cohn96), (Borr05a). Dicha interpretación consiste en representar y razonar con regiones vagas. Así, una región vaga, con frontera indeterminada, (que representará un concepto vagamente definido) estará representada mediante un par de subregiones concéntricas que sí poseen fronteras determinadas. Estas fronteras limitan la indeterminación de la región vaga  $A$  a la que ambas regiones representan. Así podremos expresar el mismo tipo de ideas sobre dos regiones vagas (cada una representada por dos subregiones) que sobre regiones con frontera definida, por ejemplo, que una contiene a otra, que se solapan, que son disjuntas, etc, mediante las relaciones de RCC. De esta forma, una región vaga representa el espacio ocupado (el área abarcada) por una de las dos subregiones, las dos a la vez, o bien por una de ellas y no la otra. De las dos subregiones concéntricas, llamamos *yema* a la interior, y es la que denotamos  $\tilde{a}$ .

Véase (Figura 1), donde se muestran las 46 posibles posiciones relativas de dos regiones vagas. El huevo  $a$  y su yema  $\tilde{a}$  estarán representados por la figura de línea intermitente y  $b$  y su yema por la de línea punteada. La yema representará el concepto, y el huevo, la noción correspondiente.

Las bases de conocimiento (BC) en Lógicas de Descripción (DL) pueden verse afectadas por las anomalías clásicas de cualquier BC. Al ser una BC en DL, un par  $\langle \mathcal{T}, \mathcal{A} \rangle$ , donde  $\mathcal{A}$  es un conjunto de hechos (la componente *extensional*), y  $\mathcal{T}$  un conjunto de relaciones entre los conceptos (la componente *intensional*), las anomalías pueden producirse en cada una de esas componentes de manera individual o bien ser anomalías de carácter mixto.

**Teorema 2.2.2** *Dada una base de conocimiento  $\Sigma$ , a l clasificar las posibles posiciones relativas de dos conceptos  $C1, C2$  (representados respectivamente por las regiones delimitadas por la línea intermitente y la punteada), y sus nociones  $\sigma(C1), \sigma(C2)$ , obtenemos la tabla<sup>2</sup> de la Figura 2.*

<sup>2</sup> El guión entre los números 2-6 y 9-11 indica la sucesión de posiciones relativas comprendida entre ambos. Es decir, 2-6 representa el conjunto de posiciones  $\{2, 3, 4, 5, 6\}$  de la Figura 1.

Por ejemplo, el conjunto de figuras {22, 35} se corresponde con el caso en que los conceptos se solapan parcialmente y las nociones están una contenida en la otra. Expresamos este hecho como sigue:

$$PO_{\Sigma}(C1, C2) \wedge PP_{\Sigma}(\sigma(C1), \sigma(C2))$$

A la vista de la tabla de la Figura 2, debemos hacer las siguientes consideraciones:

- La columna encabezada por la relación  $PP_{\Sigma}$  corresponde con la interpretación de  $I1$ . Por tanto, la caracterización de  $I1$  es  $I1 \equiv PP_{\Sigma} \wedge C^{\sigma}_{\Sigma}$
- Las columnas  $PP_{\Sigma}$  y  $EQ_{\Sigma}$  corresponden a la interpretación con pulsación de  $I2$ . Caracterización de  $I2$ :
  - $I2 \equiv (PP_{\Sigma} \vee EQ_{\Sigma}) \wedge C^{\sigma}_{\Sigma}$
- Análogamente, la columna  $PPi_{\Sigma}$  coincide con la interpretación de  $I3$  y ésta, junto con la columna  $EQ_{\Sigma}$ , resulta ser la interpretación de  $I4$ . Las caracterizaciones de éstas son, pues:  $I3 \equiv PP_{\Sigma} \wedge C^{\sigma}_{\Sigma}$ ,  $I4 \equiv (PPi_{\Sigma} \vee EQ_{\Sigma}) \wedge C^{\sigma}_{\Sigma}$ .
- La interpretación de la relación  $I5$  coincide con el contenido de las columnas  $PP_{\Sigma}$ ,  $PPi_{\Sigma}$ ,  $EQ_{\Sigma}$  y  $PO_{\Sigma}$ . Por tanto,
  - $I5 \equiv (PPi_{\Sigma} \vee EQ_{\Sigma} \vee PP_{\Sigma} \vee PO_{\Sigma}) \wedge C^{\sigma}_{\Sigma}$
- La columna  $DR_{\Sigma}$  corresponde a la interpretación de  $I7$ . Luego  $I7 \equiv DR_{\Sigma} \wedge C^{\sigma}_{\Sigma}$ .
- Por último, la tabla completa (unión de todas las posibles posiciones entre los dos conceptos y sus nociones) se corresponde con la unión de  $I5$  y  $I7$ , es decir, la interpretación con pulsación de  $I8$ , que es la relación de *indeterminación total*. No disponemos de una relación entre los dos conceptos.
- Las casillas en blanco representan posiciones imposibles, como el caso en que, estando dos conceptos conectados, las respectivas nociones sean disjuntas (discretas). Esto no es posible ya que siempre se tiene  $C \sqsubseteq \sigma(C)$ , por la propia definición de noción. Corresponden, por tanto, a anomalías de tipo inconsistencia (ver sección 4).
- Las casillas de la primera diagonal contienen las posiciones que intuitivamente no presentan un tipo de anomalía. Este es el caso de 1, 46, 24, 25, etc. En estos casos se conserva la relación entre los conceptos.
- Las casillas restantes albergan las anomalías producidas al no conservarse la relación original entre conceptos. En las figuras 33 y 36, la relación  $PP$  se invierte con las nociones. En 39 y 43, los conceptos coinciden y sus nociones se solapan (o viceversa). Además hay casos menos aceptables en general, como el de 7, 12 y 28, en los que ambos conceptos son disjuntos (discretos), mientras sus nociones pueden estar contenidas una en otra (o al revés, como en 8, 13, 34), e incluso coincidir, como en 42.

### 3 Anomalías de tipo léxico-lógico

En el estudio de la semántica del lenguaje, el signo lingüístico tiene dos caras: *significante* y *significado*. El significante es la “parte material” del signo, el significado es la imagen mental que sugiere el significante. En el contexto en el que trabajamos en este capítulo, el hecho  $\Sigma \models A \equiv B$ , que expresa que dos conceptos  $A$  y  $B$  son equivalentes, se interpreta como “la base de conocimiento no es capaz de distinguir” los conceptos  $A$  y  $B$  entre sí, es decir, son dos conceptos con el mismo significante. De manera análoga, cuando ocurre que dos conceptos tienen la misma pulsación, es decir,  $\Sigma \models \sigma(A) \equiv \sigma(B)$ , se entenderá que

tienen el mismo significado, o bien que sus definiciones coinciden. Para aplicar la clasificación clásica, identificaremos “mismo significante” con  $EQ_{\Sigma}(A, B)$ .

Desde esta perspectiva podemos estudiar las relaciones entre significante y significado, como son la monosemia, sinonimia y la polisemia. Estas dos últimas pertenecen al grupo de anomalías de tipo léxico-lógico. Por otro lado, podemos estudiar también las relaciones entre los significados, concretamente, la antonimia.

### 3.1 Sinonimia lógica

La sinonimia entre dos conceptos  $A$  y  $B$  de la base de conocimiento  $\Sigma$ , ocurre cuando los conceptos no son equivalentes, pero tienen el mismo significado:  $\Sigma \models A \neq B, \Sigma \models \sigma(A) \equiv \sigma(B)$ . Por tanto, los conceptos en cuestión verifican  $\neg EQ_{\Sigma} \wedge EQ^{\sigma}_{\Sigma}$ . Con esta interpretación, no se distingue RCC8, ya que no se consideran las fronteras, de manera que tendremos  $\neg EQ_{\Sigma} \rightarrow PO_{\Sigma} \vee DR_{\Sigma} \vee PP_{\Sigma} \vee PPI_{\Sigma}$ . De todo ello, la sinonimia viene descrita por la expresión  $(PO_{\Sigma} \vee DR_{\Sigma} \vee PP_{\Sigma} \vee PPI_{\Sigma}) \wedge EQ^{\sigma}_{\Sigma}$ . Expresión que da lugar a cuatro casos que se corresponden con las posiciones desde la 42 a la 45 de la Figura 1.

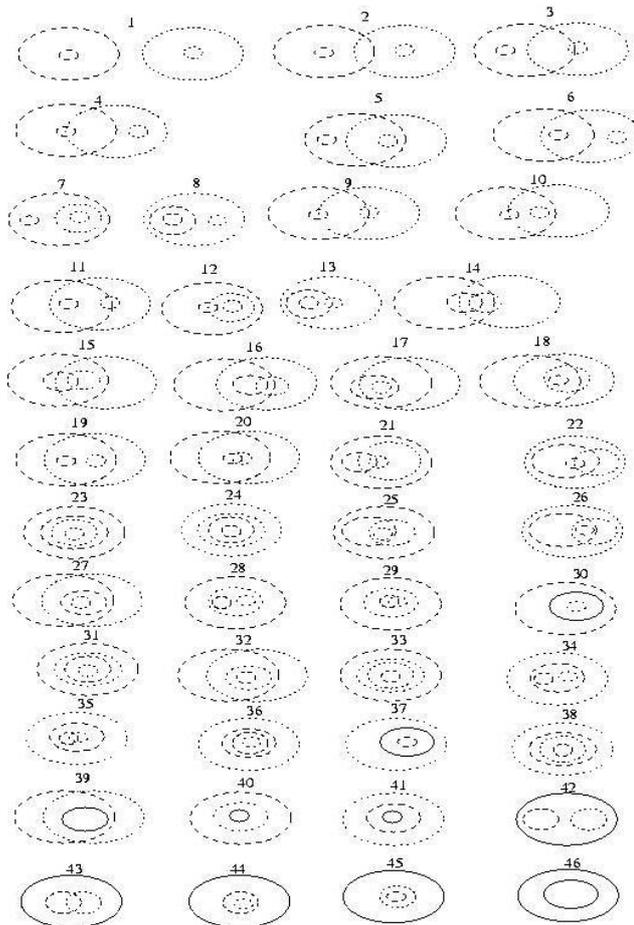


Fig. 1. Todas las posibles posiciones relativas de dos regiones vagas.

		$C_1, C_2$				
		$PPi_{\Sigma}$	$EQ_{\Sigma}$	$PR_{\Sigma}$	$PO_{\Sigma}$	$DR_{\Sigma}$
$\sigma(C_1), \sigma(C_2)$	$PPi_{\Sigma}$	23, 25, 30, 31	41	33	21, 29	7, 12, 28
	$EQ_{\Sigma}$	45	46	44	43	42
	$PR_{\Sigma}$	36	40	24, 26, 37, 38	22, 35	8, 13, 34
	$PO_{\Sigma}$	17, 27	39	18, 32	14, 15, 16, 20	2-6, 9-11, 19
	$DR_{\Sigma}$	-	-	-	-	1
		$I_A$	$I_B$		$\overline{PO}$	$I_7$
		$I_4$		$I_1$		
		$I_A$				
INTERPRETACIÓN CON PULSACIÓN						

Fig. 2. Tabla de posiciones relativas

### 3.2 Polisemia lógica

La polisemia es fenómeno contrario al que acabamos de exponer: conceptos equivalentes (indistinguibles) en la base de conocimiento, pero con definiciones (pulsaciones) no equivalentes:  $\Sigma \models A \equiv B, \Sigma \models \sigma(A) \neq \sigma(B)$ . En este caso, los conceptos en cuestión verifican  $EQ_{\Sigma} \wedge \neg EQ_{\Sigma}^{\sigma}$ . La polisemia viene descrita por la expresión  $(PO_{\Sigma}^{\sigma} \vee DR_{\Sigma}^{\sigma} \vee PP_{\Sigma}^{\sigma} \vee PPi_{\Sigma}^{\sigma}) \wedge EQ_{\Sigma}$ . Las posiciones de la Figura 1 que reflejan esta anomalía son 39, 40 y 41. Nótese que no es posible el caso  $DR_{\Sigma}^{\sigma} \wedge EQ_{\Sigma}$ .

### 3.3 Monosemia lógica

La monosemia no es un caso de anomalía, por el contrario, este fenómeno expresa el caso ideal: el concepto coincide con su definición en la base de conocimiento, es decir,  $\Sigma \models A \equiv \sigma(A)$ .

## 4 Anomalías por inconsistencia de tipo léxico-lógico

A continuación describimos anomalías de tipo léxico-lógico que son inaceptables, pues son signo de inconsistencia.

### 4.1 Antonimia lógica

Las relaciones entre los significados de los conceptos pueden ser de inclusión y de oposición. En sentido general, se entiende por antonimia el hecho de que dos conceptos tengan significados contrarios. Esta realidad puede matizarse, dando lugar a tres casos distintos de oposición: complementariedad (fiel/infiel), reciprocidad (entrega/recepción) y antonimia propiamente dicha, que considera entre los significados una gradación (frío/caliente). Esta última es una anomalía por inconsistencia que expresamos de la manera

siguiente:  $\Sigma \models A \sqcap B \not\equiv \perp$ ,  $\Sigma \models \sigma(A) \sqcap \sigma(B) \equiv \perp$ . Es obvio que la conjunción de ambas circunstancias lleva a la inconsistencia de la base de conocimiento  $\Sigma$ , pues  $\Sigma \models A \equiv \sigma(A)$ .

#### 4.2 Anomalías detectables por composición de relaciones mereotopológicas o de proyecciones reticulares

En apartados anteriores se han estudiado anomalías por pares de conceptos. Otra opción es realizar la aproximación mediante composición de proyecciones reticulares. La idea es que resulta más sencillo hacer inferencias espaciales en el modelo (la representación gráfica) que en la teoría. Veamos el siguiente ejemplo en el que consideremos la base de conocimiento  $\Sigma$ :

$T = \{\text{Niño} \sqsubseteq \text{Hombre},$

Abuelo  $\sqsubseteq$  Hombre  $\sqcap$  Padre,

Padre  $\sqsubseteq$   $\exists$ TieneHijo.(Niño  $\sqcup$  Niña)  $\sqcap$  Niño}

$A = \{\text{Padre}(\text{Pedro}), \text{Hombre}(\text{Pedro}), \text{Hombre}(\text{Juan}), \neg \text{Niño}(\text{Juan}), \neg \text{Niño}(\text{Pedro})\}$

Se tiene que:

$$PP_{\Sigma}(\text{Niño}, \text{Hombre}) \wedge PO_{\Sigma}(\text{Hombre}, \text{Padre})$$

Si  $\Sigma$  fuese consistente, entonces

$$DR_{\Sigma}(\text{Niño}, \text{Padre}) \vee PO_{\Sigma}(\text{Niño}, \text{Padre}) \vee PP_{\Sigma}(\text{Niño}, \text{Padre})$$

En este caso,  $\Sigma \models PP(\text{Padre}, \text{Niño})$ , lo cual es una contradicción. Luego  $\Sigma$  es inconsistente.

### Bibliografía citada

- ALONSO-JIMÉNEZ, J. A. [et al.]. A methodology for the computer-aided cleaning of complex Knowledge databases. En: *IECON 2002: actas*. IEEE Press 2002, p. 1806-1811.
- BORREGO-DÍAZ, J., CHÁVEZ-GONZÁLEZ, A. M. Extension of ontologies assisted by automated reasoning systems. En: *International Conference on Computer Aided Systems Theory. Lecture Notes in Computer Science (10<sup>th</sup>: 2005): actas*, p. 247-253.
- BORREGO-DÍAZ, J., CHÁVEZ-GONZÁLEZ, A. M. Controlling ontology extension by uncertain concepts through cognitive entropy [recurso electrónico]. En: *International Semantic Web Conference, Workshop on Uncertainty Reasoning for the Semantic Web (4<sup>th</sup>: 2005): actas* <<http://CEUR-WS.org/Vol-173/>> [Consultado: 15 dic.2006]
- BORREGO-DÍAZ, J., CHÁVEZ-GONZÁLEZ, A. M. Visual ontology cleaning: cognitive principles and applicabilit. En: *European Semantic Web Congress (ESWC 2006) (3<sup>rd</sup>: 2006): actas*, p. 317-331.
- COHN, A. G. [et al.]. Representing and reasoning with qualitative spatial relations about regions. En: STOCK, O. (ed.) *Temporal and Spatial Reasoning*. Kluwer, 1997. p 97-134.
- COHN, A. G.; GOTTS, N. M. The 'Egg-Yolk' Representation of regions with indeterminate boundaries. En: BURROUGH, P.; FRANK, A. M. (eds.). *Proc. GISDATA Specialist Meeting on Geographical Objects with Undetermined Boundaries*. GISDATA Series, vol. 3, Taylor and Francis, 1996. p. 171-187.

- FIKES, R.; MCGUINNESS, D. L.; WALDINGER, R. *A First-Order logic semantics for semantic Web markup languages*. Tech. Rep. KSL-01-01. Knowledge Systems Laboratory, Stanford University, 2002.
- KIM, W. [et al.]. A Taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 2003, vol. 7, p. 81-99.
- RENZ, J.; NEBEL, B. On the complexity of qualitative spatial reasoning: a maximal tractable fragment of the region connection calculus. *Artificial Intelligence*, 1999, vol. 108, n. 1-2, p. 69 -123.