Using a hybrid approach for the development of an ontology in the hydrographical domain

F. J. López-Pellicer [a,1], L. M. Vilches-Blázquez [b], J. Nogueras-Iso [a,1], O. Corcho [c], M. A. Bernabé [d], A. F. Rodríguez [b]

[a] *Computer Science and Systems Engineering Department, University of Zaragoza, Zaragoza (Spain)*

[b] *National Geographic Institute, Madrid (Spain)*

[c] *School of Computer Science, University of Manchester,Manchester(UK)*

[d] *Technical University of Madrid, Madrid (Spain)*

Abstract

This work presents a hybrid approach for domain ontology development, which merges top-down and bottom-up techniques. In the top-down approach the concepts in the ontology are derived from an analysis and study of relevant information sources about the domain (e.g., hydrographic features). In the bottom-up approach the concepts in the ontology are the result of applying formal methods on a analysis of the data instances on the repositories (e.g., repositories containing hydrographical features).

Key words: hydrography, urban ontologies, ontological engineering

Email addresses: fjlopez@unizar.es (F. J. López-Pellicer), lmvilches@fomento.es (L. M. Vilches-Blázquez), jnog@unizar.es (J. Nogueras-Iso), Oscar.Corcho@manchester.ac.uk (O. Corcho), ma.bernabe@upm.es (M. A. Bernabé), afrodriguez@fomento.es (A. F. Rodríguez).

[1] This work has been partially supported by the Spanish Ministry of Education and Science through the project TIN2006-00779 from "the National Plan for Scientific Research, Development and Technology Innovation".

1 Introduction

This work presents a hybrid approach for domain ontology development, which merges top-down and bottom-up techniques. In the top-down approach the concepts in the ontology are derived from an analysis and study of relevant information sources about the domain (e.g., hydrographic features). In the bottom-up approach there is an analysis of application domain repositories (e.g., repositories containing hydrographical features). The results of this analysis are applied to generate dynamically the ontology.

The purpose of applying this hybrid approach is to provide a pragmatic aspect which might help to verify the appropriateness and feasibility of the theoretical domain ontology proposed in top-down approaches with the application ontology obtained in the bottom-up approach. Additionally, the merging of top-down and bottom-up approaches facilitates the mapping between the domain ontology and a particular repository, a task which is usually required for projects related to data harmonization of heterogeneous repositories. This hybrid approach represents a novel way of developing ontologies, which has not been usually applied in the literature of ontological engineering until now. However, we think that it can provide important benefits in contexts that require the harmonization and conversion of heterogeneous data repositories.

Additionally, this work describes as a use case the applicability of this methodology in the context of the Hydrography and Urban Civil Engineering domains. Hydrography and related phenomena represent an essential part of reality in our cities as a consequence of the water supply needs they all have. This is going to characterize some aspects of city planning owing to the presence of water infrastructures and to the addition of certain hydrographic features in urban landscapes (Vilches-Blázquez et al., 2007). Even natural features such as rivers, when crossing urban environments, have their boundaries shaped by people and can be considered as artificial objects (Fonseca et al., 2000).

The Spanish National Geographic Institute (IGN), the organizational body leading the development of the Spanish Spatial Data Infrastructure (IDEE), is defining a hydrographic domain ontology to establish mappings between the IGN feature catalogues and others managed at local, national, regional, and European level. IGN has begun to build a domain ontology of hydrographic features, which is called "hydrOntology", whose purpose is to serve as a harmonization framework among Spanish cartographic producers. For the development of "hydrOntology" we have followed the proposed hybrid approach.

The rest of this paper is organized as follows. Next section describes the hybrid approach methodology, describing the activities involved in this methodology

and the techniques applied for each activity. Then section 3 shows the applicability of this approach for the development of this methodology in the hydrography domain. Finally, this paper ends with some concluding remarks and proposals for further research.

2 Hybrid approach methodology for the development of a domain ontology

The methodology for the hybrid approach proposed consists of the following activities:

- Development of a draft version of the ontology following a top-down approach.
- Development of a draft version of the ontology following a bottom-up approach.
- Comparison of ontologies. The objective of this activity is to find a consensus between top-down and bottom-up approaches.

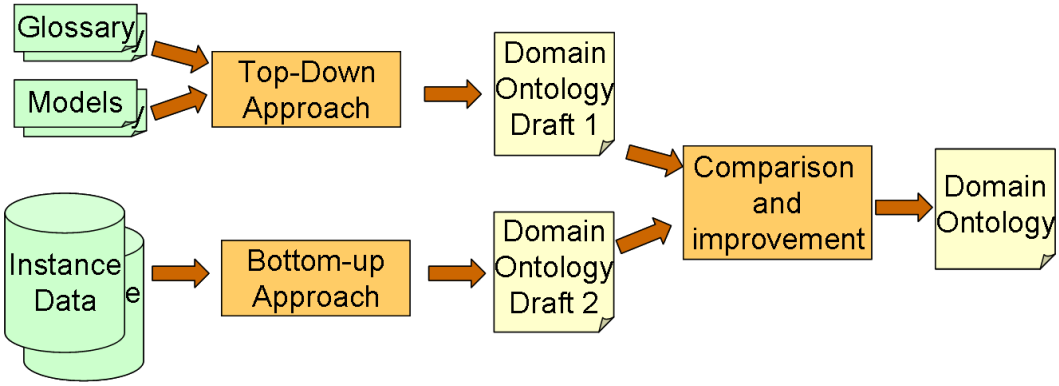

Figure 1. Hybrid approach methodology

Figure 1 displays the process proposed for this hybrid approach methodology. The following subsections describe in more detail the top-down and bottom-up approaches.

2.1 Top-down ontology

For the top-down approach, we propose the use of METHONTOLOGY, a widely-used methodology for building ontologies. METHONTOLOGY emphasises the reuse of existing domain and upper-level ontologies and proposes to use, for formalisation purposes, a set of intermediate representations that can be later transformed automatically into different formal languages. Therefore this methodology is suitable for developing ontologies at the knowledge level.

Moreover, it takes into account the main activities identified by the IEEE software development process (IEEE, 1996) and other knowledge engineering methodologies.

METHONTOLOGY has been used by different groups to build ontologies in different knowledge domains, such as Chemistry, Science, Knowledge Management, e-Commerce, etc. A detailed description of the methodology of this ontology building can be found in (Gómez-Pérez et al., 2003). Figure 2 shows the ontology building tasks suggested in the METHONTOLOGY framework (Corcho et al., 2005).

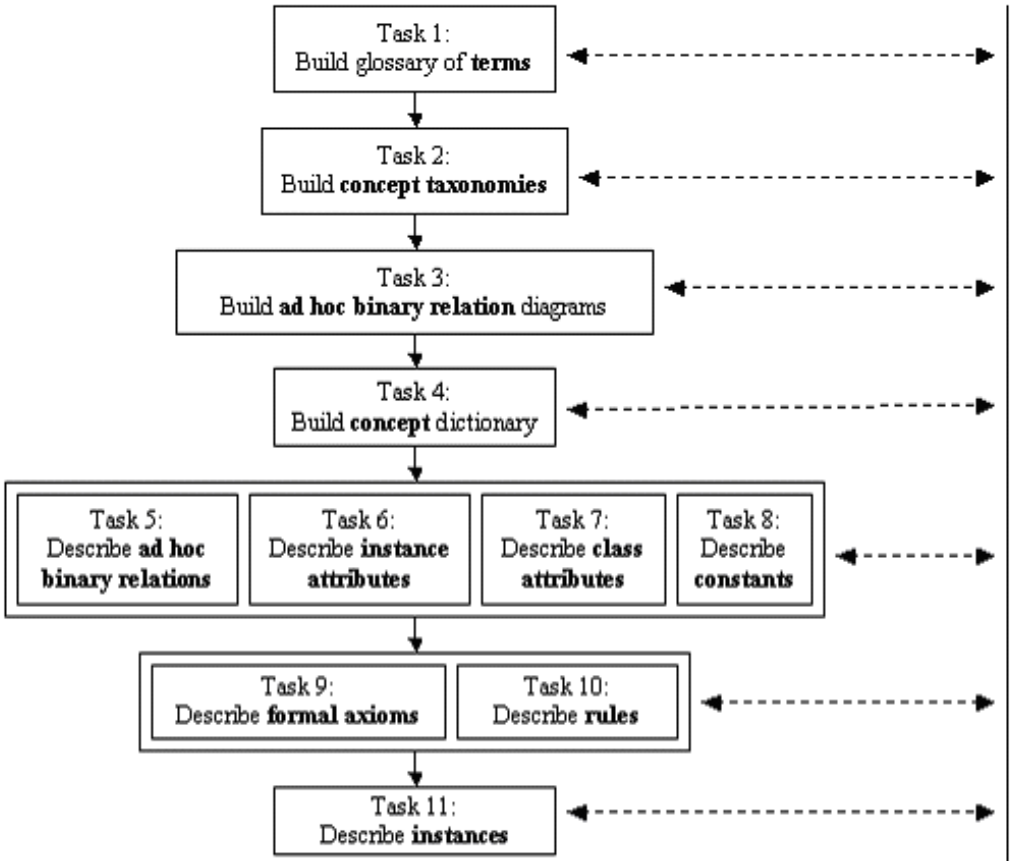

Figure 2. Tasks of the conceptualization activity according to METHONTOLOGY (Gómez-Pérez et al., 2003)

The figure 2 emphasizes the ontology components (concepts, attributes, relations, constants, formal axioms, rules and instances) built inside each task. Also, this figure illustrates the steps this methodology proposes for creating such components during the conceptualization activity. This is not a sequential modelling process, though some order must be followed to ensure the consistency and completeness of the represented knowledge(Corcho et al., 2005).

METHOTOLOGY proposes a tasks set for capturing a domain knowledge(Gómez-Pérez et al., 2003). Theses ones can be divided into three groups of tasks.

The first group would be steering to enclosure and structure the domain by means of tasks 1 to 4 (see figure 2).

- Task 1: To build the glossary of terms that identifies the set of terms to be included on the ontology, their natural language definition, and their synonyms and acronyms.
- Task 2: To build concept taxonomies to classify concepts. The output of this task could be one or more taxonomies where concepts are classified.
- Task 3: To build ad doc binary relations diagrams to identify ad hoc relationships between concepts of the ontology and with concepts of other ontologies.
- Task 4: To build the concept dictionary, which mainly includes the concept instances for each concept, their instance and class attributes, and their ad hoc relations.

The second group of tasks, from 5 to 7, would help to document the acquired knowledge from the previous tasks.

- Task 5: To describe in detail each ad hoc binary relation that appears on the ad hoc binary relation diagram and on the concept dictionary. The result of this task is the ad-hoc binary relation table.
- Task 6: To describe in detail each instance attribute that appears on the concept dictionary. The result of this task is the table where instance attributes are described.
- Task 7: To describe in detail each class attribute that appears on the concept dictionary. The result of this task is the table where class attributes are described.

Finally, METHONTOLOGY proposes others tasks, from 8 to 11, to complete a domain knowledge.

- Task 8: To describe in detail each constant and to produce a constant table. Constants specify information related to the domain of knowledge, the always take the same value, and are normally used in formulas.
- Once that concepts, taxonomies, attributes and relations have been defined, METHONTOLOGY proposes to describe formal axioms (task 9) and rules (task 10) that are used for constraint checking and for inferring values for attributes. Optionally, information about ontologies should be introduced (task 11).

It is important to mention that different domain ontologies may have different knowledge representation needs, so this methodology suggests that the previous set of tasks should be reduced or extended as needed.

2.2 Bottom-up ontology

For the development of a domain ontology following a bottom-up approach we propose the applicability of Formal Concept Analysis (FCA) techniques (Ganter and Wille, 1999; Stumme and Maedche, 2001) to output a hierarchy of concepts from the feature instances contained in the repositories used as data sources (See figure 3).

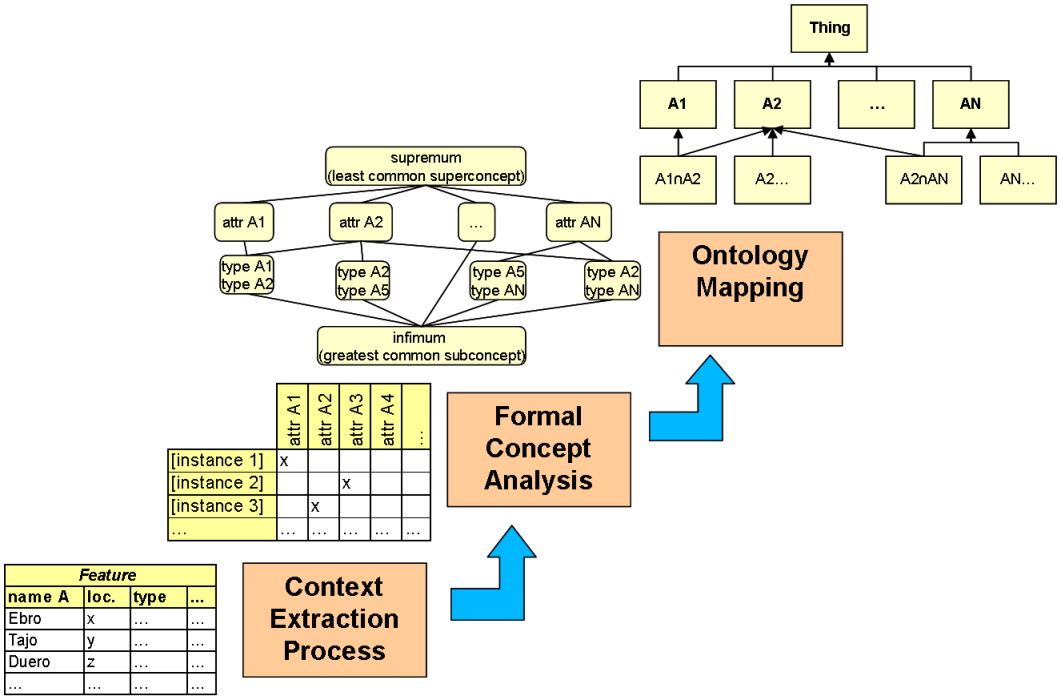

Figure 3. Bottom-up process

The basis of FCA is the definition of a *formal context* ($\mathbb{K}$), which consists in a triple (G, M, I) where G is a set of *objects* and M is a set of *attributes*. I (*incidence*) represents the binary relation between "objects" and "attributes" with only two possible values, *present* or *absent*.

There are two *closure operators* that link G and M within a *formal context* $\mathbb{K}$:

$$A \subseteq G, A' = \{m \in M | \forall g \in A, (g, m) \in I\} \tag{1}$$

$$B \subseteq M, B' = \{g \in G | \forall m \in B, (g, m) \in I\} \tag{2}$$

A' can be understood as the maximum set of attributes common to the objects in A and B' as the maximum set of objects which have in common the attributes in B. Given these definitions, the pair (A, B) is called a *formal*

concept if and only if:

$$A \subseteq G, B \subseteq M, A' = B \wedge A = B' \tag{3}$$

In other words, (A, B) is called a *formal concept* if and only if the maximum set of attributes shared by the objects in A is B and, on the other hand, A is the maximum set of objects which share the attributes in B. A is called the concept *extent* and B the concept *intent*. The set of all the *formal concepts* of the *formal context* is partially ordered by the order induced by the set inclusion:

$$(A_1, B_1) \leq_{\mathbb{K}} (A_2, B_2) \Longleftrightarrow A_1 \subseteq A_2 \ (\Longleftrightarrow B_2 \subseteq B_1) \tag{4}$$

Where the formal concept (A_1, B_1) is called *subconcept* of the formal concept (A_2, B_2), and (A_2, B_2) is called *superconcept* of the formal concept (A_1, B_1). Furthermore, the induced partial order is a complete *lattice*, known in this context as *concept lattice*.

Comparing FCA with respect to Object Orientation, *formal concepts* are equivalent to *classes*, and *superconcepts* and *subconcepts* relationship between concepts are equivalent to the *generalization* and *specialization* relationships.

FCA techniques have a direct application on repositories of data that consists of one single table. In this case, each row maps to an object and each column to a set of attributes. The incidence relation I derives from the contents of the table: for each row the presence or absence of a value or range of values in a column determines the presence or absence of one or several attributes. However the mapping between rows, columns and data values from the repository and objects and attributes of I is a non trivial task if the relational schema is denormalized.

Therefore, previous to the application of FCA techniques, the main issue is to obtain from the repositories a unified and homogenized view of the data as *objects* and *attributes*, the formal context required by FCA. As our purpose is to create an ontology draft, the selected data should contain thematic attributes. Data that best fit to this requirement is *hydrologic gazetteer data*. Among other thematic attributes, each gazetteer feature is described as belonging to a feature type and their name may contain valuable thematic data in the *generic name*.

Our approach is as follows:

(1) Select the gazetteer entries related to hydrography. Also prepare a set of common hydrographic names with their variants.

$GAZ \leftarrow$ Hydrographic gazetteer

$GEN \leftarrow$ Hydrographic names

(2) Set initially G as the set of features contained in the gazetteer.
$G \leftarrow \{g|g \in GAZ \cdot isFeature(g)\}$

(3) Set as M the set of feature types used in the gazetteer that belong to the hydrographic domain along with those generic hydrographic names which apear in the selected features.
$M \leftarrow \{t|t \in GAZ \cdot isFeatureType(t)\} \cup \{n|n \in GEN \cdot \exists g \in G, contain(g \rightarrow name, n)\}$

(4) Define I initially as the incidence relationship between features and generic names.
$I \leftarrow \{(g, m)|g \in G \cdot \forall m \in M \cdot isGeneric(m) \wedge contain(g \rightarrow name, m)\}$

(5) Remove from G features whose name does not contain a generic name.
$G \leftarrow G \setminus \{g|g \in G \cdot \nexists m \in M \cdot (g, m) \in I\}$

(6) Complete I with the incidence relationship between the remainder features and their feature types.
$I \leftarrow I \cup \{(g, m_1)|m_1 \in M \cdot \exists (g, m_2) \in I \cdot isFeatureType(m_1) \wedge g \rightarrow type = m_1\}$

Working with generic names is not an easy task. Gazetteers can contain *multilingual generic names* and *synonyms*. Another issue is the existence of *slight differences* between the generic name of a feature name and those in M. Finally there exists the possibility that the generic name and the feature type of a feature have *different semantics*. The multilingual generic names problem is solved with the use of a dictionary. The most promising approach to the matching problem is the use of robust string matching libraries, e.g. *SecondString*(Cohen et al., 2003). And the occasional different semantics problem is solved by counting duplicate rows in I and removing them, and hence the correspondent g, if their number is bellow a threshold.

Once obtained the incidence matrix, the concept lattice is generated using one of the several algorithms available, in our case *next closed set* (Ganter, 1987). This generated lattice identifies:

(1) Relevant feature types from their *extent*.
(2) New feature types derived from *formal concepts* that contain a generic as attribute.
(3) Feature types that are candidate to a disjoint-decomposition.

Thanks to the FCA technique and some minor adjustments, the original feature type taxonomy can be enriched in a way that helps the ontologist to understand better the domain.

3 Experiment: Applying the hybrid approach methodology to the hydrography domain

As mentioned in the introduction, IGN is defining a hydrographic domain ontology to establish mappings between their own feature catalogues and others managed at local, national and European level. This domain ontology is called "hydrOntology" and it has been developed following the hybrid 1approach described in previous section.

Following subsections describe the applicability of the hybrid approach methodology to the development of "hydrOntology".

3.1 Top-down ontology

In order to develop our ontology following the top-down approach, we have taken into account different knowledge models (feature catalogues of the National Geographic Institute of Spain, the Water Framework Directive, the Alexandria Digital Library, the UNESCO Thesaurus and quithe a lot of others), some integration problems of geographic information and several structuring criteria (Vilches-Blázquez et al., 2007).We have tried to cover most of existing GI sources in order to build a full domain ontology. For that reason, this ontology contains more than a hundred relevant concepts related to hydrography (e.g. river, reservoir, lake, channel, pipe, water tank, siphon and so on).

Figure 4 shows a "hydrOntology" model overview. It is divided into two levels; the upper level represents the most abstract features in the ontology and the lower level describes a set of well-known hydrographic features. The upper level contains the "Hydrographical Feature" concept, and other specialised concepts like "Inland Waters" and "Sea Waters". There is a different degree of specialisation in each of these concepts, since the current focus of this ontology is on "Inland Waters". According to the Water Framework Directive (European Parliament, 2000), these concepts are divided into "Superficial Waters" ("Transitional waters", "Stand Waters", "Flowing Waters" and "Sources" are subclasses of "Superficial Waters") and "Groundwaters". For each of these classes we have identified concepts in the lower level, where a detailed set of hydrographic features is provided.

Furthermore, in the "hydrOntology" development we have taken into account some concepts about feature capture that depend exclusively on different Spanish geographic regions. Among these features appear "ibón", "lavajo", "chortal", "bodón" and "lucio". These concepts are designated by their local

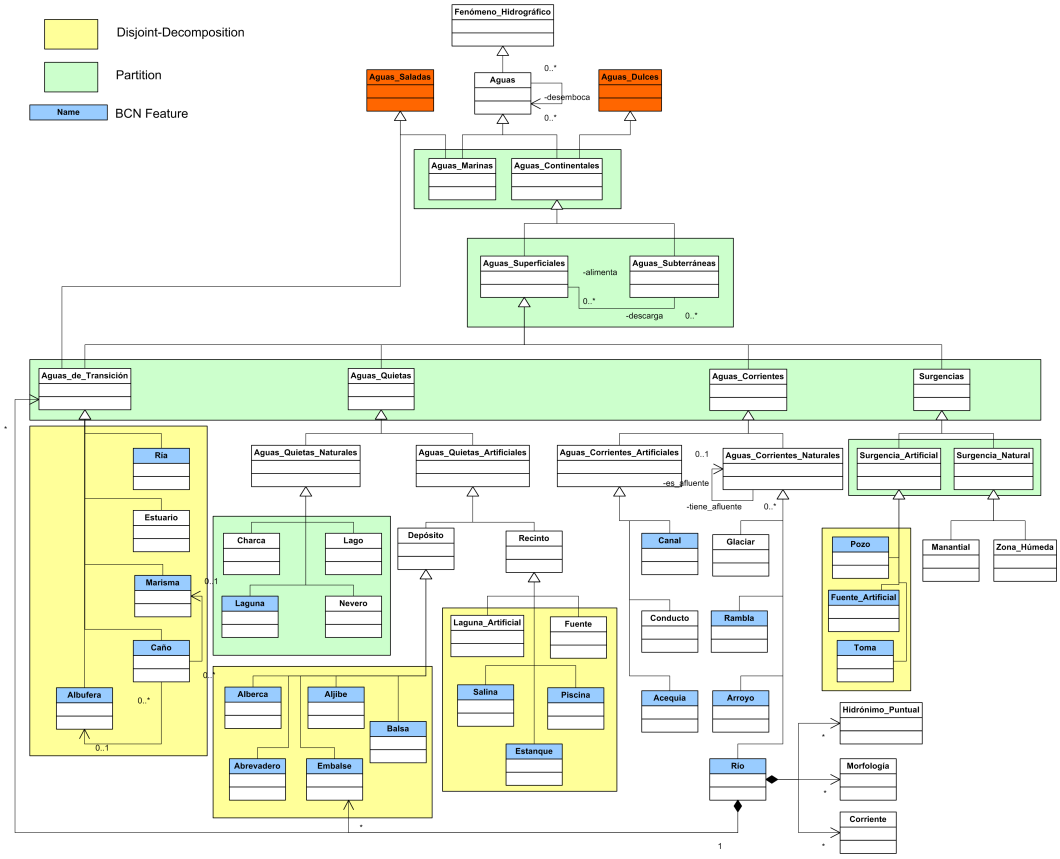

Figure 4. Top-down ontology

name and they are synonymous to the feature "Charca" [2].

Moreover, this figure shows some examples of the four taxonomic relations defined in the Frame Ontology (Farquhar et al., 1997) and the OKBC Ontology (Chaudhri et al., 1998), both used by METHONTOLOGY methodology (Vilches-Blázquez et al., 2007).

A concept C1 is a Subclass-Of another concept C2 if and only if every instance of C1 is also an instance of C2 (Corcho et al., 2005).

A Disjoint-Decomposition of a concept C is a set of subclasses of C that do not have common instances and do not cover C, that is, there can be instances of the concept C that are not instances of any of the concepts in the decomposition (Corcho et al., 2005). Some examples of this type of relationship are shown in figure 4.

An Exhaustive-Decomposition of a concept C is a set of subclasses of C that cover C and may have common instances and subclasses, that is, there cannot

[2] "Charca" is a small lake of shallow water. The above mentioned terms are Spanish local names.

be instances of the concept C that are not instances of at least one of the concepts in the decomposition (Corcho et al., 2005). Figure 4 shows an example of this type of relationship.

A Partition of a concept C is a set of subclasses of C that do not share common instances and that cover C, that is, there are not instances of C that are not instances of one of the concepts in the partition (Corcho et al., 2005). Some examples of a partition are shown in figure 4.

At the moment we are working on providing mappings of this ontology with other databases at several levels (from local to national level). Furthermore, we are planning to provide multilingual support for "hydrOntology" (English, French, Portuguese, Catalan, Basque, Galician languages) and to merge this ontology with other domain ontologies (e.g. Urban Civil Engineering).

3.2 Bottom-up ontology

For the bottom-up approach we are analyzing the repositories that have been used to build a gazetteer at the Spanish National Geographic Institute. In particular, we have focused on the part of the repositories used as source for the generation of the hydrographic names.

Figure 5 shows how the process has been applied to the feature repositories. As it can be observed, the *Thematic Analysis* module determines the feature type and the generic name of each feature. Both the feature type and the generic name is the thematic signature of a feature. Then the *Filter* select the distinct signatures which represent a significant number of features and create an incidence matrix whose rows are these signatures. Finally the *Lattice Builder* applies FCA and then transforms the formal concept lattice into OWL (Web Ontology Language) (Bechhofer et al., 2004), and RDF-language to express or encode ontologies.

Figure 6 shows part of the generated ontology. This ontology contains 51 concepts. They can be classified from their source as *IGN Feature Types* (suffixed with "IGN"), as *generic names* (suffixed with "GEN") and *maps* (prefixed with "MAP"). The most common concepts by far are "Corriente fluvial IGN" (stream of water) (71 % of instances) and their subclass "MAP Corriente fluvial arroyo", the map between "Corriente fluvial IGN" (stream of water) and "Arroyo GEN" (creek) (52 % of instances).

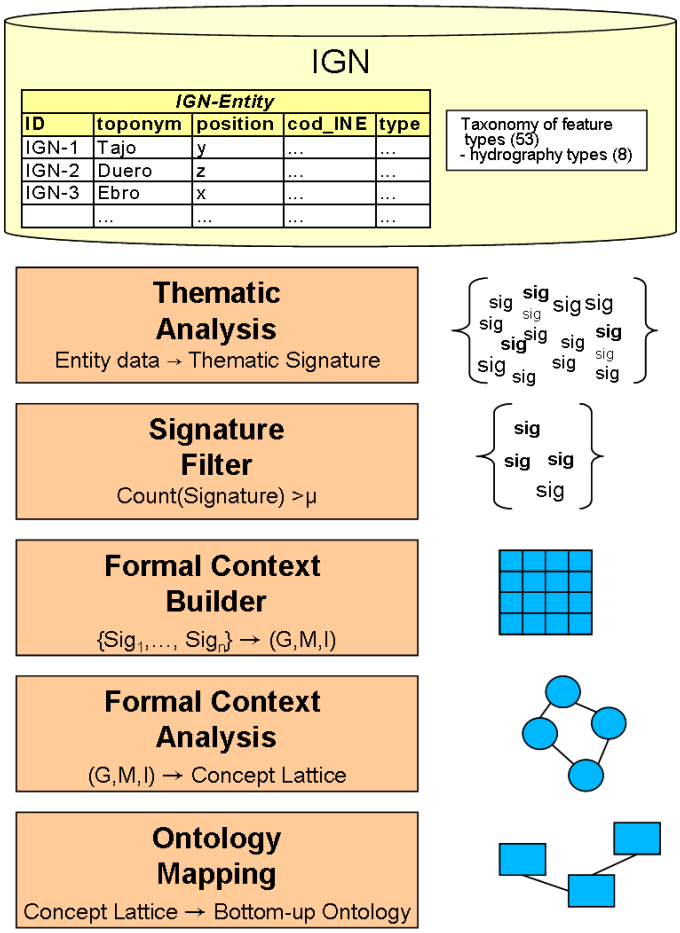

Figure 5. Building from the repository

3.3 Comparison of results for the improvement of hydroOntology

From the comparison between the results of the Formal Concept Analysis and "hydrOntology", some facts have been revealed:

(1) Equivalente terms between ontologies and data sources. Some IGN feature types have the same name as concepts described in "hydrOntology" and share the same semantics (e.g. "embalses", dams, "corriente fluvial", streams) or a broader one (e.g. "canales", artificial stream instead of the expected irrigation channel).
(2) Other few feature types have instances that belong to unrelated concepts rather than the lexically nearest concept in the ontology (e.g. "humedal", wetland). This might identify a missing relationship or attribute.
(3) Finally, other clues to identify of missing attributes are IGN feature types that describe features which only shares non thematic attributes such as position, shape or size. The best example is "Accidente hidrográfico" (hydrographic feature) that contains features described in "hydrOntology" as thermal features, springs and parts of rivers whose only shared

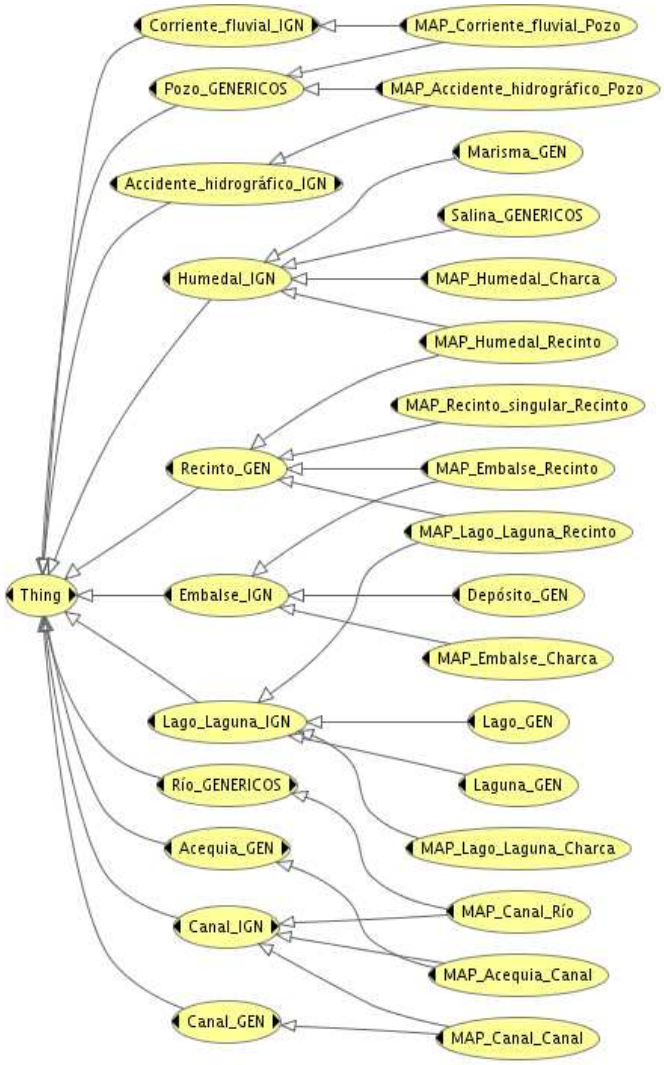

Figure 6. Bottom Up Ontology (part)

characteristic is their representation as a point.

The previous comparison has given the opportunity to obtain the necessary feedback to evaluate the feasibility of "hydrOntology" and enrich it. Some advices to improve the structure and concepts of "hydrOntology" could be the following:

(1) Simplification of some ontology concepts. Locally bound concepts with similar characteristics (e.g. small ponds such as "Bodón" or "Lavajo") should be merged or its description increased with a description of the geographic region where exclusively occurs.
(2) Locally bound concepts found in "hydrOntology" should be maintained only if they are relevant in size or number in an area (e.g., "Ría", which is a transitional water feature type only found in the north of Spain).
(3) Each concept should have not only multilingual support but also the

dialectal and local variants in each language.

4 Conclusions

This work has presented a hybrid approach for domain ontology development, which merges top-down and bottom-up techniques. Each technique produces ontologies which differs in their respective point of view. Top-down ontology draw the required/expected semantic of the data held in the repositories. Bottom-up ontology reveals the effective/possible semantics of the data held in the repositories. Comparing both ontologies provide useful information and feedback.

As regards the experiments in the hydrography domain, we can conclude that the ontology derived from FCA has provided insight on possible missing attributes and relationships in"hydrOntology" and advice on how to improve the multilingual support or to treat locally bound feature types. Future work will be oriented to find more automatic mechanisms for the comparison and merging of top-down and bottom-up approaches. For instance, we could merge both ontologies using tools such as PROMPT (Noy and Musen, 2000).

References

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., Stein, L. A., February 2004. OWL Web Ontology Language Reference. W3C, W3C Recommendation 10 February 2004, http://www.w3.org/TR/2004/REC-owl-ref-20040210/.

Chaudhri, V. K., Farquhar, A., Fikes, R., Karp, P. D., Rice, J. P., January 1998. Open Knowledge Base Connectivity 2.0.3. Technical Report KSL-98-06, Knowledge Systems Laboratory, Stanford, CA, http://www.ai.sri.com/ okbc/okbc-2-0-3.pdf.

Cohen, W. W., Ravikumar, P., Fienberg, S. E., 2003. A Comparison of String Distance Metrics for Name-Matching Tasks. In: Proc. IIWeb 2003 (IJCAI 2003 Workshop). pp. 73–78.

Corcho, O., Fernández-López, M., Gómez-Pérez, A., López-Cima, A., 2005. Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications. Springer-Verlag, Ch. Building legal ontologies with METHONTOLOGY and WebODE, pp. 142–157.

European Parliament, 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. The EU Water Framework Directive - integrated river basin management for Europe. L 327, 22/12/2000 pp. 0001-0073, European Parliament.

Farquhar, A., Fikes, R., Rice, J., 1997. The Ontolingua Server: A Tool for Collaborative Ontology Construction. International Journal of Human Computer Studies 46 (6), 707–727.

Fonseca, F. T., Egenhofer, M. J., Davis Jr., C. A., Borges, K. A. V., 2000. Ontologies and knowledge sharing in urban GIS. Computers, Environment and Urban Systems 24 (3), 251–271.

Ganter, B., 1987. Algorithmen zur formalen begriffsanalyse. In: Ganter, B., Wille, R., Wolff, K. E. (Eds.), Beiträge zur Begriffsanalyse. B.I.–Wissenschaftsverlag, Mannheim, pp. 241–254.

Ganter, B., Wille, R., 1999. Formal Concept Analysis: Mathematical Foundations. Springer, Berlin-Heidelberg.

Gómez-Pérez, A., Fernández-López, M., Corcho, O., 2003. Ontological Engineering. Springer-Verlag, London (United Kingdom).

IEEE, 1996. IEEE Standard for Developing Software Life Cycle Processes. IEEE Std 1074-1995. IEEE Computer Society, New York.

Noy, N., Musen, M., 2000. Prompt: Algorithm and tool for automated ontology merging and alignment. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000). Austin, Texas, pp. 450–455.
URL http://smi-web.stanford.edu/people/noy/publications.html

Stumme, G., Maedche, A., 2001. FCA-MERGE: Bottom-up merging of ontologies. In: Proc. 17th IJCAI. Seattle (WA US), pp. 225–230.

Vilches-Blázquez, L. M., Ángel Bernabé-Poveda, M., Suárez-Figueroa, M. C., Gómez-Pérez, A., Rodríguez-Pascual, A. F., 2007. Ontologies for Urban Development: Interfacing Urban Information Systems. Vol. 61 of Studies in Computational Intelligence. Springer, Ch. Towntology & hydrOntology: Relationship between Urban and Hydrographic Features in the Geographic Information Domain, pp. 73–84.