(Draft) Ontologies for Urban Development.Studies in Computational Intelligence.
Germany: Springer, 2007, vol. 61, p. 157-167. ISBN 978-3-540-71975-5.

Building an Address Gazetteer on top of an Urban Network Ontology

J. Nogueras-Iso, F. J. López, J. Lacasta, F. J. Zarazaga-Soria, and P. R. Muro-Medrano

University of Zaragoza, Zaragoza, Spain
{jnog,fjlopez,jlacasta,javy,prmuro}@unizar.es

Abstract. In order to create the contents of an address gazetteer service that forms part of a city council Spatial Data Infrastructure, all the existent repositories containing address information in the different council offices must be analyzed and harmonized. The problem is that usually these repositories are constrained by the use of different taxonomies for the identification of urban network feature types. The objective of this work will be to describe how to establish a formal ontology enabling the interoperability among the different taxonomies, and facilitating the construction of the gazetteer contents.

1 Introduction

The increasing relevance of geographic information for decision-making and resource management in diverse areas of government has promoted the creation of Spatial Data Infrastructures (SDI), which are usually defined as a coordinated approach to technology, policies, standards, and human resources necessary for the effective acquisition, management, distribution and utilization of geographic information at different organization levels and involving both public and private institutions. In the particular context of the development of an SDI for local administrations such as a city council, address gazetteer services represent one of the most important services that the councils must offer to their citizens [1]. The councils are responsible for the management of urban networks, and these networks are used as reference information for other services at national level such as cadaster or census services.

The creation of contents for an address gazetteer service requires SDI developers to perform a work of analyzing and harmonizing all the existent repositories containing address information in the different offices of the council. The main problem typically found is that different taxonomies are used for the identification of urban network feature types in different administrative processes. Frequently, when city councils need to exchange information with external organizations like National Cadaster Offices or National Statistics Institutes, the information needs to be reformatted in order to comply with the feature types accepted by these institutions. Moreover, it is usual that this reformatted information is stored at council level in parallel repositories (e.g., tax office databases, urban planning office databases) whose updates are not synchronized.

(Draft) Ontologies for Urban Development.Studies in Computational Intelligence.
Germany: Springer, 2007, vol. 61, p. 157-167. ISBN 978-3-540-71975-5.

In order to overcome the existent heterogeneity in the different repositories used for gazetteer contents, it seems sensible to establish a unified model of the feature types that can be found in this domain, and make the necessary mappings to the particular taxonomies that must be used in external organizations or in the different repositories maintained at council level. This feature type model could be formally represented by an ontology that defines explicitly the concepts and relationships between these concepts in a domain [2, 3]. On the one hand, this unified ontology would facilitate the interoperability with external administrative organizations. And on the other hand, it would enable the modelling of the contents served by the Gazetteer service.

Having observed this necessity of defining an ontology for feature types in the urban networks domain, the objective of this work will be to explore the mechanisms to build a unified urban network ontology on top of the existent taxonomies in the public administration for urban networks. The construction of an ontology upon existing vocabularies (textual dictionaries; glossaries; or even more structured vocabularies that can include taxonomies, thesauri or other existent ontologies) is a classical and widely used approach in ontological engineering whose main problem is how to reconcile the source taxonomies. For instance, [4] proposes the creation of a unified ontology by mapping different source ontologies against a common controlled vocabulary (the ADL Feature Type Thesaurus). The underlying problem, also known as *ontology alignment*, is how to find the relationships (e.g., equivalence or subsumption) that hold between the entities represented in different taxonomies. Moreover, ontology alignment methods may be useful for assisting conflict resolution among people having different conceptualisations of a given domain.

The remaining paper is organized as follows. Section 2 analyzes the use-case selected for this work explaining the different urban network databases (including their different feature type taxonomies) that must be used for the creation of a gazetteer. Then, the next two sections describe how to build the ontology that will guide the contents of the gazetteer. Whereas the first approach will describe an ad-hoc manual mapping among taxonomies used in the source repositories, the second one will describe how to apply *Formal Concept Analysis* techniques for the automatic creation of a formal urban network ontology that integrates the mappings among the different taxonomies. Finally, the paper ends with some conclusions and future lines.

2 Use case: Urban Network databases at the Zaragoza city council

The use-case selected for this work has been the SDI developed for the Zaragoza city council in Spain (IDEZAR, http://www.zaragoza.es/idezar/). Figure 1 sketches the flow of information concerned with urban transport networks between the different offices of the Zaragoza city council and from/to external administrative bodies. Three different categorizations are used for the urban network feature

(Draft) Ontologies for Urban Development.Studies in Computational Intelligence.
Germany: Springer, 2007, vol. 61, p. 157-167. ISBN 978-3-540-71975-5.

types in the different offices of the Zaragoza city council: *SIGLA*, *TVIAN* and *AYTO*.

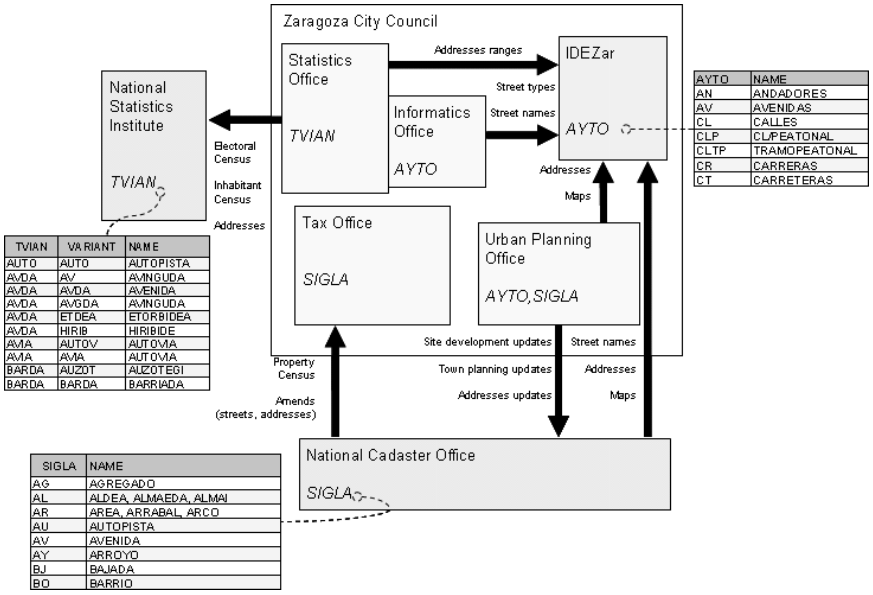

Fig. 1. Workflow

The *SIGLA* categorization (*Sigla de vía pública*) is a code list that describes the street types used in the transfer file format between the local government and the National Cadaster Office. This code list consists of acronyms and some of them are shared by two different street types (e.g., *CM* represents both *Camino/path* and C*ármen/southern kind of house*). *SIGLA* is used in the urban network databases managed by the Tax Office and the Urban Planning Office of the Council. Whereas the Tax Office is responsible of land taxes management, the Urban Planning Office is responsible of the land development. Its Geographical Information Service is responsible for the urban cartography and the parcel numbering. As regards the data flows where *SIGLA* is involved, three main data flows can be mentioned. Firstly, *SIGLA* is used for land address oriented data flows. The Tax Office informs the National Cadaster Office about tax management (owner addresses) and land management (property addresses). Secondly, *SIGLA* is used for tax management data flow. This flow comprises all the data sent from the National Cadaster Office to the city councils to help land tax management (property taxes, land valuation changes, amends). Also city councils may inform the National Cadaster office any change in the owner's data or mistakes. And thirdly, *SIGLA* is also used for town planning and development data flow. This flow informs the National Cadaster Office about any land related permissions, planning change or address change made by the city councils.

(Draft) Ontologies for Urban Development.Studies in Computational Intelligence.
Germany: Springer, 2007, vol. 61, p. 157-167. ISBN 978-3-540-71975-5.

The second categorization, *TVIAN* (*Tipo de Vía Normalizado*) is a partially normalized code list of street types used in the transfer file format between local governments and the Spanish National Statistics Institute. It establishes a mapping between a normalized key and a set of acronyms, which are variants in the different local languages. However, there is no hint of the language of each variant and, it is ill normalized as some concepts have more than one normalized key (e.g., the concept *callejón/alley* has the normalized key *CLLON* for the Spanish and the Basque language but *CXON* for the Galician language). *TVIAN* categorization is internally used in the council for the database managed by the Statistics Office, which is responsible of the inhabitant census and the poll census. As regards data flow, *TVIAN* is involved in the data flow concerned with citizen statistics. This flow, which goes from city councils to the National Statistics Office, comprises the inhabitant continuous census, the poll census and any change in streets, street number ranges and addresses.

And the third categorization is called *AYTO*. This code list is owned by the Zaragoza council which compiles the street types included in the local regulation (e.g., *caminos/paths*, *carreteras/roads*, *plazas/squares*, *calles/streets*, *paseos/boulevars*, *parques/parks*). Additionally, it integrates as well more specific street types "to avoid confusion" between streets with the same name. This categorization is used by the Culture Office and the Informatics Office. The council street names are proposed by the Culture Department, but their encoding and maintenance is the responsibility of the Informatics Office. Additionally, the Informatics office gives technical support to the applications based on the council gazetteer in hardcopy version.

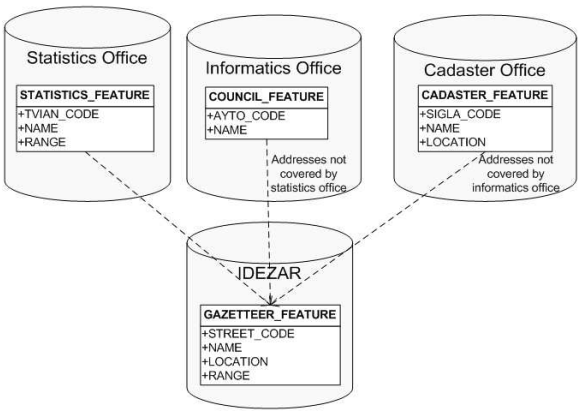

Fig. 2. Gazetteer contents

Finally, extracting the information from the three databases described above, the objective was to define an electronic gazetteer aggregating the information available in such databases. Figure 2 shows how the conceptual models from

(Draft) Ontologies for Urban Development.Studies in Computational Intelligence.
Germany: Springer, 2007, vol. 61, p. 157-167. ISBN 978-3-540-71975-5.

the above databases are matched and merged for the creation of the conceptual model that would define the contents of the desired electronic gazetteer. However, the main problem remains at the instance level in order to match the feature instances from the source databases and produce the feature instances the must be uploaded in the gazetteer database. That is to say, for each feature instance (e.g., *Plaza España*) found in both source databases, we need to integrate the normalized street name (from *COUNCIL_FEATURE* database at the Informatics Office) with the location (from *CADASTER_FEATURE* database at Tax Office) and the street range (*tramero* found in the *STATISTICS_FEATURE* database at the Statistics office). And for that purpose, it seems obvious that we need a mapping between the different feature types found in the three categorizations. On one hand, the matching of feature instances is not enough just using the feature names. For instance, *España* is a name that can be used for squares and streets. And on the other hand, it would be interesting that the gazetteer provides a consistent feature type categorization for the features served by the gazetteer, probably the common factor of the three source categorizations.

3 Ontology construction using a manual mapping approach

As explained in the introduction, we must face the problem of aligning the different taxonomies already available in order to identify equivalences between the entities represented in the different taxonomies and extract the most relevant concepts (including as well possible subsumption hierarchies).

One approach for this alignment is obviously the manual mapping between the different taxonomies. In particular, given that the taxonomies mentioned in previous section (*TVIAN*, *AYTO*, *SIGLA*) had no semantic description (in most cases just an acronym and the complete name), a manual mapping approach was tried in first place. That is to say, a human expert had the responsibility of comparing terms (acronyms+names) in the different taxonomies and establishing the mapping across the different taxonomies.

The objective was to use the *AYTO* as the bridge between the taxonomies and establish the manual mappings: *SIGLA* - *AYTO* (see figure 3), and *TVIAN*-*AYTO*. As *SIGLA* belongs to a property-oriented database, *TVIAN* belongs to a census-oriented database and *AYTO* belongs to an urban oriented database, it was expected that identical terms would overlap in the different databases. However, even in this specific domain, it was found that homonyms can arise (even with terms belonging to the same conceptual design).

The procedure for the manual mapping consisted of the following steps: collect acronyms from the different database; expand acronyms with their complete names; look up for definitions; and match equivalent terms based on their similar definitions. The matches that were obtained could be classified in the following categories:

- Exact match: the meanings of both concepts are identical.

(Draft) Ontologies for Urban Development.Studies in Computational Intelligence.
Germany: Springer, 2007, vol. 61, p. 157-167. ISBN 978-3-540-71975-5.

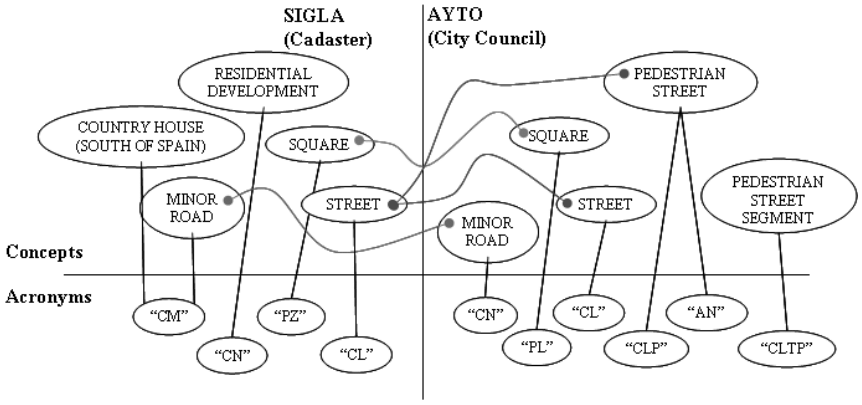

Fig. 3. Mapping *SIGLA-AYTO*

- Partial match: one concept is broader or narrower than the other. The concept represented by *CL* (street) in *SIGLA* has narrower concepts in *AYTO* such as the concepts represented by *CLP* and *AN* (different types of pedestrian streets).
- Provisional match: due to the design of *SIGLA* where different concepts share the same acronyms, the matching of concepts is provisional. The feature instances linked with this match should be verified.
- No match: the concept does not exist. We need add a residual category to cover these cases.

The experience from this first approach has shown that this non-systematic manual process results quite subjective, too time expensive and with little scalability. If a new taxonomy is added to the possible lists, a new mapping to the not very well structured *AYTO* taxonomy should be established. A more flexible approach could be the use of well-established shared common core and make mappings between the distinct sources and this common core.

Thus, our second experiment consisted in mapping the source taxonomies against the URBISOC thesaurus [5]. It is a thesaurus focused on Spanish terminology for Town Planning, which has been developed by the CINDOC/CSIC institute (Centre for Scientific Information and Documentation / Spanish National Research Council) to facilitate classification at the URBISOC bibliographic database, which is specialized in scientific and technical journals on Geography, Town Planning, Urbanism and Architecture. Additionally, we decided to use a proper ontology editor to facilitate communication and discussion between the experts in charge of the alignment. The tool selected was Towntology [6], which enables the storage of the ontology, the display of the ontology in visual graphic form, to navigate in the ontology and to query it. The main difference with respect to other ontology editors such as Protégé [7] is that it is

(Draft) Ontologies for Urban Development.Studies in Computational Intelligence.
Germany: Springer, 2007, vol. 61, p. 157-167. ISBN 978-3-540-71975-5.

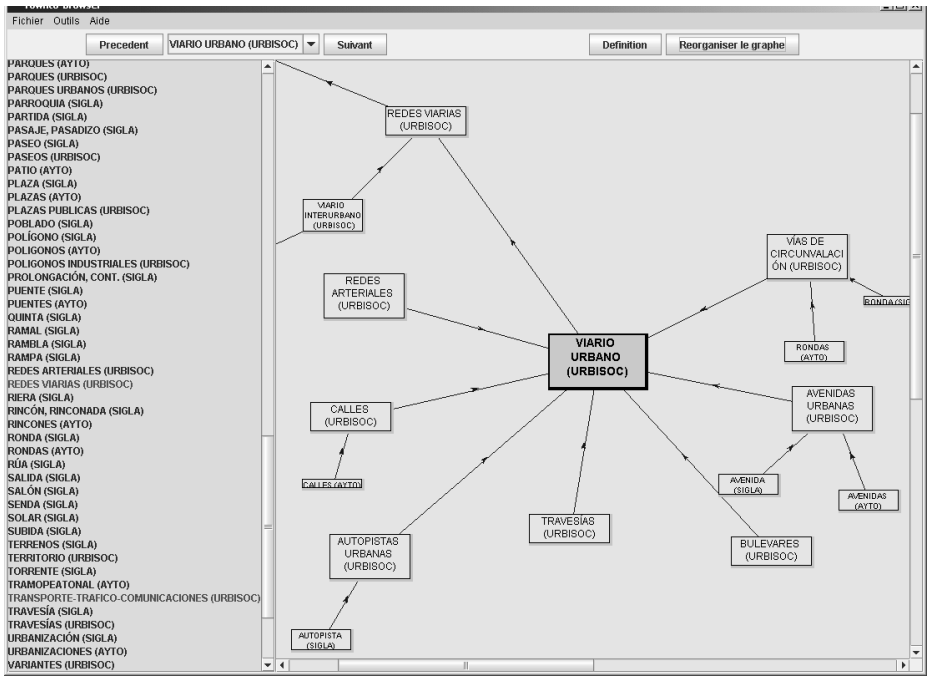

Fig. 4. Use of URBISOC as common core

not based on any formalism such as RDF(s) [8] or OWL [9]. But at this step we were more focused in the ontology construction process than in representing formally a built ontology. The Towntology tool is aimed for storing concepts with several definitions that are in a process of selection and characterization of these definitions.

Figure 4 shows a screenshot of the Towntology browser displaying some mappings between the URBISOC thesaurus and some of the terms available at *SIGLA* and *AYTO*. Although improving the scalability, this second attempt results still time expensive and error prone.

4 Ontology construction using an FCA approach

Having seen the difficulties in establishing a manual mapping of ontologies, it is highly beneficial to count on methods for the automatic alignment of these existent vocabularies, facilitating the rapid creation of a draft of the desired ontology.

In particular, this section describes the applicability of *Formal Concept Analysis* (FCA) techniques [10, 11] to output a hierarchy of concepts from the feature instances contained in the three databases shown in section 2. The basis of FCA is the definition of a *formal context*, which consists in a triple (G, M, I) where

(Draft) Ontologies for Urban Development.Studies in Computational Intelligence.
Germany: Springer, 2007, vol. 61, p. 157-167. ISBN 978-3-540-71975-5.

G and M are sets and $I \subseteq G \times M$ is a binary relation between G and M. The elements in G are called objects, those in M attributes and I the incidence of the context.

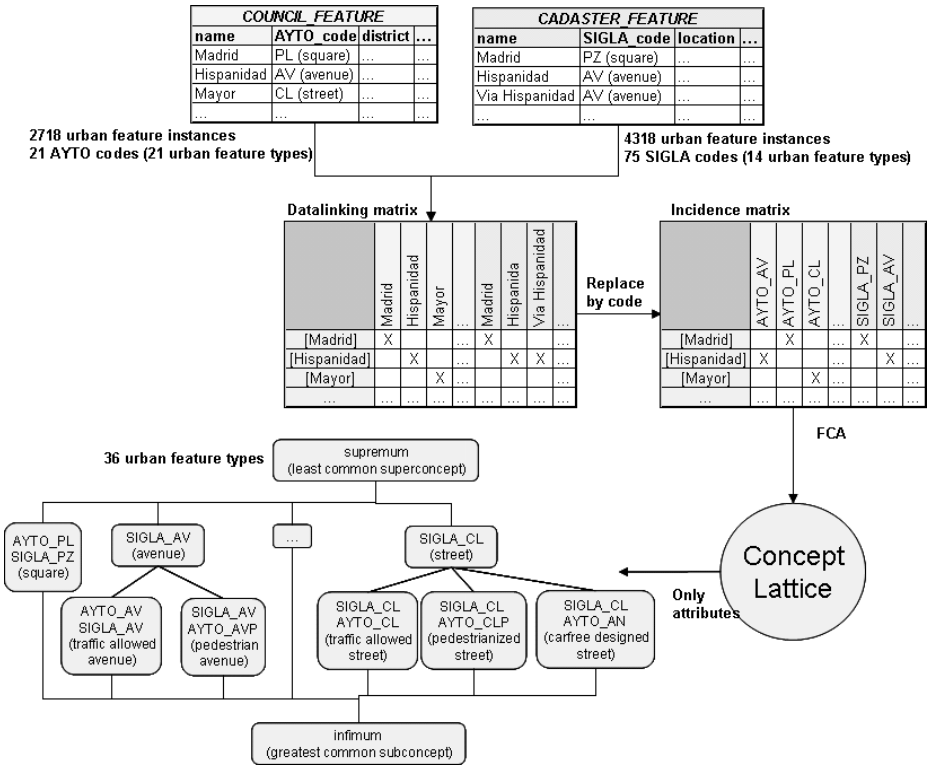

Fig. 5. Application of FCA

The objective of FCA is the extraction of a lattice of formal concepts, but previous to the definition of formal concepts we will define A' and B' for $A \subseteq G$ and $B \subseteq M$:

$$A' = \{m \in M | (g,m) \in I \; for \; all \; g \in A\}; \; B' = \{g \in G | (g,m) \in I \; for \; all \; m \in B\} \quad (1)$$

A' can be understood as the set of all the attributes common to the objects in A and B' is the set of all the objects which have in common with each other the attributes in B. And given these definitions, a pair (A, B) is a formal concept if and only if

$$A \subseteq G, B \subseteq M, A' = B \wedge A = B' \quad (2)$$

In other words, (A, B) is a formal concept if and only if the set of all attributes shared by the objects in A is identical with B and on the other hand A is also the set of all the objects which have in common with each other the attributes

(Draft) Ontologies for Urban Development.Studies in Computational Intelligence.
Germany: Springer, 2007, vol. 61, p. 157-167. ISBN 978-3-540-71975-5.

in B. Furthermore, the concepts of a given context are naturally ordered by the subconcept-superconcept relation defined by

$$(A_1, B_1) \leq (A_2, B_2) \Longleftrightarrow A_1 \subseteq A_2 \ (\Longleftrightarrow B_2 \subseteq B_1) \tag{3}$$

The ordered set of all formal concepts of (G, M, I) enables the definition of a concept lattice by linking it is possible to establish a concept lattice.

Figure 5 depicts the process of applying the FCA techniques to the instances contained in two source databases: the *COUNCIL_FEATURE* database using *AYTO* taxonomy, and the *CADASTER_FEATURE* database using *SIGLA* taxonomy. The main problem for the direct application of FCA techniques in our context was how to obtain a unique repository of instances, i.e. the formal context required by FCA. Therefore, in order to obtain this unique repository, traditional datalinking has been applied to the feature instances contained in the different databases. This datalinking has been based on the analysis of the lexical and spatial similarities of feature attributes, i.e. the lexical similarity of names (use of *SecondString* string similarity function library [12]) and the proximity of locations. Then, the datalinking matrix obtained as a result of this process together with the transformation of urban network feature type codes (e.g., *AYTO_CODE*, *SIGLA_CODE*) into proper attributes (with boolean values) enables the creation of the incidence matrix I of the formal context.

Once obtained the incidence matrix, a version of the algorithm *next closed set* [13] has been used to generate the concept lattice that establishes the alignment between the two source taxonomies. Thanks to the FCA technique and some minor adjustments, the source taxonomies can be transformed into a merged hierarchy of formal concepts. The technique not only identifies equivalent concepts in both taxonomies, but also subconcept-superconcept relations. An example of equivalent concept would be a *square* (*PL* in *AYTO* and *PZ* in *SIGLA*). And an example of subconcept-superconcept relation would be the identification of *street* as a broader concept in *SIGLA* (*CL*), which has narrower concepts in the *AYTO* taxonomy such as traffic-allowed streets (*CL*), pedestrianized streets (*CLP*) or carfree-designed streets (*AN*).

5 Conclusions

This paper has shown different mechanisms for the construction of an urban network ontology by means of the alignment of different source taxonomies. In particular, a manual mapping approach and an automated approach based on *Formal Concept Analysis* have been studied. Although minor problems must be supervised manually, it has been demonstrated that the second approach (based on FCA) provides more flexibility and scalability. Additionally, this technique enables the extraction of concepts independently from the encoding of the feature types. That is to say, it would be possible to analyze different data sources that have used a number encoding, without any apparent meaning, for the classification of features.

(Draft) Ontologies for Urban Development.Studies in Computational Intelligence.
Germany: Springer, 2007, vol. 61, p. 157-167. ISBN 978-3-540-71975-5.

Besides, the unified ontology obtained as result of this alignment process has been used to create the contents of an address gazetteer service integrated within the SDI of a local council. The unified ontology enables the union of the toponyms coming from the different databases used in the city council offices, detecting when necessary the intersections and avoiding duplications. Furthermore, this unified ontology would allow the construction of customized user query interfaces which can still use the original taxonomies according to the requirements of each city council office.

As future lines of this work, it is planned to make a refinement of the FCA-based approach in order to improve the efficiency and the formalization of the generated ontologies. On the one hand, it is believed that the detection and filtering of instances that may introduce noise will avoid generating spurious concepts. On the other hand, the formalization level could be enriched by means of extracting statistics over the attributes of original feature instances (e.g., conclusions about limits on the perimeter, area or geometry of the *square* concept). Finally, it is worth noting that this FCA-based approach could be also applied to other domains making use of toponyms and where ontologies help revealing the structure of separate repositories. For instance, it could be applied to the analysis of hydronyms, which are usually managed at national and regional levels by National Mapping Agencies and Water Agencies respectively.

Acknowledgements

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" and by the COST UCE C21 Action (Urban Ontologies for an improved communication in Urban Civil Engineering projects) of the European Science Foundation.

References

1. Portolés-Rodríguez, D., Álvarez, P., Muro-Medrano, P.: IDEZar: an example of user needs, technological aspects and the institutional framework of a local SDI. In: Proc. 11th EC GI & GIS Workshop, ESDI Setting the Framework. (2005)
2. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering. Springer-Verlag, London (United Kingdom) (2003)
3. Visser, P.R.S., Jones, D.M., Bench-Capon, T.J.M., Shave, M.J.R.: An Analysis of Ontological Mismatches: Heterogeneity versus Interoperability. In: AAAI 1997 Spring Symposium on Ontological Engineering, Stanford, USA (1997)
4. Berman, M.L.: Semantic Interoperability and Cultural Specificity: Examples from Chinese, Japanese, Mongolian and Uighur. In: Proc. of Social Science History Association meeting (SSHA'2003), Baltimore (2003)
5. Alvaro-Bermejo, C.: Elaboración del Tesauro de Urbanismo URBISOC. Una Cooperación Multilateral. In: Encuentro Hispano-Luso de Información Científica y Técnica. II, Salamanca (1988)

(Draft) Ontologies for Urban Development.Studies in Computational Intelligence.
Germany: Springer, 2007, vol. 61, p. 157-167. ISBN 978-3-540-71975-5.

6. Keita, A., Laurini, R., Roussey, C., Zimmerman, M.: Towards an Ontology for Urban Planning: The Towntology Project. In: CD-ROM Proc. 24th UDMS Symposium, Chioggia (2004)
7. Noy, N.F., Fergerson, R.W., Musen, M.A.: The knowledge model of Protege-2000: Combining interoperability and flexibility. In: Knowledge Engineering and Knowledge Management. Methods, Models, and Tools: 12th Int. Conf., EKAW 2000. Volume 1937 of LNCS., Juan-les-Pins, France (2000) 17–32
8. Manola, F., Miller, E., eds.: RDF Primer. W3C, W3C Recommendation 10 February 2004 (2004) http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.
9. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A.: OWL Web Ontology Language Reference. W3C, W3C Recommendation 10 February 2004 (2004) http://www.w3.org/TR/2004/REC-owl-ref-20040210/.
10. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin-Heidelberg (1999)
11. Stumme, G., Maedche, A.: FCA-MERGE: Bottom-up merging of ontologies. In: Proc. 17th IJCAI, Seattle (WA US) (2001) 225–230
12. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A Comparison of String Distance Metrics for Name-Matching Tasks. In: Proc. IIWeb 2003 (IJCAI 2003 Workshop). (2003) 73–78
13. Ganter, B.: Algorithmen zur formalen begriffsanalyse. Beiträge zur Begriffsanalyse. BIWissenschaftsverlag, Mannheim/Wien/Zürich (1987) 241–254