TOWARDS INTENSIONAL/ EXTENSIONAL INTEGRATION BETWEEN ONTOLOGIES

Eleni Tomai

Institute of Applied and Computational Mathematics, Foundation of Research and Technology – Hellas, P.O BOX 1529, 7110 Heraklion Crete
etomai@iacm.forth.gr

Commission II, WG II/6

KEY WORDS: Information Flow, Extension, Intension, Mappings, Ontologies, Tokens, Types

ABSTRACT:

This paper presents ongoing research in the field of extensional mappings between ontologies. Hitherto, the task of generating mapping between ontologies has been focused on the intensional level of ontologies. The term intensional level refers to the set of concepts that are included in an ontology. However, an ontology that has been created for a specific task or application needs to be populated with instances. These comprise the extensional level of an ontology. This particular level is being generally neglected during the ontologies' integration procedure. Thus, although methodologies of geographic ontologies integration, ranging from alignment to true integration, have, in the course of years, presented a solid ground for information exchange, little has been done in exploring the relationships between the data. In this context, this research strives to set a framework for extensional mappings between ontologies using Information Flow.

1. INTRODUCTION

A well-formed thematic or domain ontology should be able to provide answers to two types of questions:

1. What is a X? Or what it means to be X? Or can you define X? and
2. What is this? Or is this X?

The first type of questions refers to a process that humans often call description, explanation or definition, respectively. This process results into the demarcation of the different concepts in the ontology, as well as their definitions. No reference is made at this point to how these concepts are included in the ontology in the first place. Furthermore, the process helps identify semantic relations from one part and semantic properties from the other part (as defined and explained in Kokla and Kavouras 2002, Tomai and Kavouras 2004), which can produce the hierarchical structure of the ontology. The second type of questions refers to the process of categorization; namely the process of assigning members to category. The process itself accounts for allocating the instances of each concept in the ontology. Therefore, a well-formed ontology should include both concepts that stand in some kinds of relationships among them, as well as instances of these concepts.

The aforementioned procedures respectively refer to what linguistics define as:

- *Intension*, what you must know in order to determine the reference of an expression.
- *Extension,* the class of objects that an expression refers to (WordNet, 2003).

Consequently, we can distinguish between the intensional level (the set of concepts) and extensional level (the concepts' instances) of an ontology.

The growing interest in ontologies among geoscientists along with the plethora of data for the geographic domain have revealed the need for a form of unified information and has set the path for ontologies integration. The issue, therefore, is how to integrate two or more geographic ontologies in order to produce a new one, which contains all the pieces of information, contained by the original ones, or at least how to generate mappings between different ontologies, so that users can switch between them, reaching to semantic interoperability. Bearing in mind that an ontology includes two distinct levels of elements (concepts and instances) that both provide ontological information, we present, herein, a methodology for utilizing those two in order to reach to integration.

The methodology applies tenets from Information Flow Theory in order to perform integration of ontologies in two levels; intensional and extensional. The process of integration at this level aims at the analysis of definitions of the concepts, the extraction and statement of their semantic properties and relations and finally the revelation of heterogeneities that guide the establishment of the final/new schema. Thus far the majority of approaches to geographic ontologies integration have explored only the possibilities of integration at the intensional level. Herein, we explore the possibilities of adding the extensional level of the ontologies to the integration process.

2. INTEGRATING ONTOLOGIES; THE WAY SO FAR

Thus far, several methodologies of ontology integration have been presented by scholars. According to the framework presented by Kavouras (2005) we can identify four types of integration:

1. Alignment
2. Partial compatibility
3. Unification
4. True integration

In the case of alignment, mappings are generated between the concepts of the two ontologies; no distortion is made to either of them. This is the simplest integration case for it can be seen as a "translation" mechanism between the two ontologies' concepts. For a methodology on generating mappings between geographic ontology refer to Cruz et al. (2004).

On the other hand, partial compatibility refers to the unification of the common parts of the ontologies. The result is a single ontology but integration has just taken place for the common parts of the ontology with consequent distortion of the original ones.

Unification is an extension of partial compatibility, which results into a single ontology, by unifying every branch of the two ontologies into one. The two initial ontologies are fully distorted.

True integration refers to the procedure of producing a new ontology, which includes the initial ones without any alteration, however it includes some new concepts that are needed to associate the ontologies. The initial ontologies can be reused independently from the integrated one. A methodology of true integration between geographic ontologies using Formal Concept Analysis has been introduced by Kokla (2000).

The aforementioned methodologies have dealt with the intensional level of ontologies; they do not treat integration at the extensional level.

There has been the case where integration process has been generated between ontologies at the extensional level. Duckham and Worboys (2005) have proposed an algorithm of geographic ontologies integration depending on relationships between instances, which are able to infer taxonomic relations between the categories themselves. If prior knowledge of the taxonomies exists it can be taken into account but it is not a prerequisite.

The above-described methodology leaves a few open questions regarding the suitability of the assumption that extensional information can be used as an inference mechanism for the taxonomic structure of the intensional level. Two questions that should be addressed are:

- Identical instances of two categories refer to equivalent categories, or subsumed ones?
- How many instances to compare in order to achieve integration?

The first question reveals the problem of depending only on instances of ontologies to achieve integration. Two identical instances in two different ontologies may be members of two identical categories. However, this is not always the case, because these instances may belong to one category in the first ontology and to its subsumed category in the second ontology, given that the latter is more detailed than the former. Therefore, inferring taxonomic structure of the ontologies depending only on instances is not adequate.

Integration at the intensional level has never tackled the issue of how many categories are included in the ontologies to give a result. However, in the case of extensional information, there is an issue of sufficiency regarding the minimum number of instances that categories should have before the integration process can be pursued.

3. INTEGRATING ONTOLOGIES; THE WAY FORWARD

3.1 Information Flow

As mentioned, the method utilizes Information Flow Theory to provide the theoretical basis for the integration process. At this point it is essential to present and explain to the novice reader some key points of the theory.

The basic idea behind the Information Flow (Barwise and Seligman, 1997) is the notion of *containment*, which translates as the information an object contains about another. Information Flow is better understood within a *distributed system* Flow (Barwise and Seligman, 1997). Distributed systems are regarded as wholes with interrelated parts. Regularities within these systems ensure the flow of information between the parts. Consequently, the more random a system is the less information can flow (Bremer and Cohnitz, 2004). In literature (Lalmas, 1998) (Old and Priss, 2001), parts of a distributed system are considered as particulars; they are of some type.

3.1.1 Classification: The components of a distributed system are represented by a *classification A* (Devlin, 2001), which is a triple, $\left\langle \mathrm{A}, \Sigma_A, \models_A \right\rangle$, where A is the set of objects of *A* to be classified, called *tokens* of *A*, Σ_A the *types* of *A*, used to classify the tokens, while the tokens stand in relation $\models_A$ to the types (Fig.1).

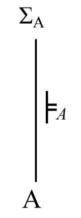

Figure 1. Classification *A*

Each classification has a Local Logic that governs its types (Bremer and Cohnitz, 2004). This logic allows inferences to be drawn at the type level of the classification. The sequent $\alpha \vdash \beta$, for types α, β indicates that the inference from α to β holds. For instance, the sequent house ⊢ building indicates that houses are buildings (Worboys, 2001).

3.1.2 Infomorphism: For relating two classifications, the notion of *infomorphism* (Devlin, 2001) is introduced. Let *A* $= \left\langle \mathrm{A}, \Sigma_A, \models_A \right\rangle$ and $B = \left\langle B, \Sigma_B, \models_B \right\rangle$ be two classifications. An infomorphism *f* between them consists of two functions; f^+ from types of *A* to types of *B*, and f^- from tokens of *B* to tokens of *A* (Fig.2).

$$\begin{tikzcd}
\Sigma_A \arrow[r, dash] \arrow[d, dash, "\models_A"] & \Sigma_B \arrow[d, dash, "\models_B"] \\
A \arrow[r, dash, "f^-"'] & B
\end{tikzcd}$$

Figure 2. Infomorphism f from *A* to *B*

3.1.3 Channel: The notion of *channel* is used to express relationships between situations (Devlin 2001). We write $s_1 \overset{c}{\longmapsto} s_2$ to denote that a situation s_1 delivers some of the information supported by a situation s_2 with respect to channel *c* (Lalmas, 1998). The channel allows formalizing the context in which the flow of information takes place. In other words, a channel *c* is the medium for Information Flow between two classifications *A* and *B* as those previously mentioned; it connects them through a core classification *C* via two infomorphisms *f* and *g* (Fig. 3).

$$\begin{tikzcd}
\Sigma_A \arrow[r, dash, "f^+"] \arrow[d, dash, "\models_A"] & \Sigma_C \arrow[r, dash, "g^+"] \arrow[d, dash, "\models_C"] & \Sigma_B \arrow[d, dash, "\models_B"] \\
A \arrow[r, dash, "f^-"'] & C \arrow[r, dash, "g^-"'] & B
\end{tikzcd}$$

Figure 3. A channel *c* between classifications *A* and *B*

3.2 Semantic Interoperability

The Information Flow framework can be used to enable semantic interoperability between different communities that use their own classifications, by providing the necessary mappings across them. Approaches of generating mappings based on Information Flow can be found in (Kalfoglou and Schorlemmer, 2002), (Kalfoglou and Schorlemmer, 2003) and (Schorlemmer and Kalfoglou 2003). In addition, Kent (2001 and 2004) has developed the Information Flow Framework for the standardization activity of Standard Upper Ontology, and proposed a methodology for ontology merging.

Interoperability in the context of Information Flow takes place both at the type and the token level as it can be understood from the notions of channel and infomorphism. Therefore, when generating mappings between classifications, the instances of the classifications are compared as well. Figure 4 illustrates how the concept of Information Flow works.

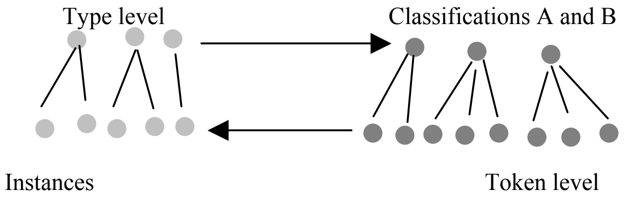

Figure 4. Applying the Information Flow framework to generate semantic interoperability between communities that use different classifications (Information Flow mappings – IF)

4. AN EXAMPLE OF TYPE/ TOKEN LEVEL MAPPINGS BETWEEN SOURCES OF GEOGRAPHIC INFORMATION; THE CASE OF THEMATIC MAPS

This section provides a framework for formalizing the information content of thematic maps. Maps are the most acknowledged graphic representations of spatial phenomena and are widely used due to their expressive power and convenience of conveying geographic information. This section compares the information content of two different thematic maps representing the same phenomenon, by generating mappings between the two maps. The objective of such a comparison, adopting the map-reader's point of view, is having a consistent overall view of the displayed phenomenon.

For demonstrating the aspects of this research, let us consider a map that represents population density of a certain part of Europe (l) at time t, and another one that also represents population density of another part of Europe (l') at the same time t. Due to the difference in population density classes, and to different symbols (areas of different colour intensities), it is complicated for the map user to have a complete view of the phenomenon for the two regions at time t.

In order to map between these two different graphic representations, we adopt the formal theory of Information Flow (evolution of Situation Theory) introduced by Barwise and Selingman (1997). We view maps as distributed systems (having separate parts that constitute a whole) with regularities. In this context, we define these distributed systems as classifications, and we attempt to provide mappings between these different systems (maps) to achieve semantic interoperability.

To make it more explicit to the novice reader, and paraphrasing Sowa's definition (Sowa 2005), a mapping of concepts and relations between two classifications *A* and *B* preserves the partial ordering by subtypes in both *A* and *B*. If a concept or relation *x* in classification *A* is mapped to a concept or relation *y* in classification *B*, then *x* and *y* are said to be *equivalent*. The mapping may be partial: there could be concepts in *A* or *B* that have no equivalents in the other classification. No other changes to either *A* or *B* are made during this process. The mapping process does not depend on the choice of names in either classification. Figure 5 portrays the procedure adopted in the paper to formalize the content of thematic maps and to generate mappings between them.

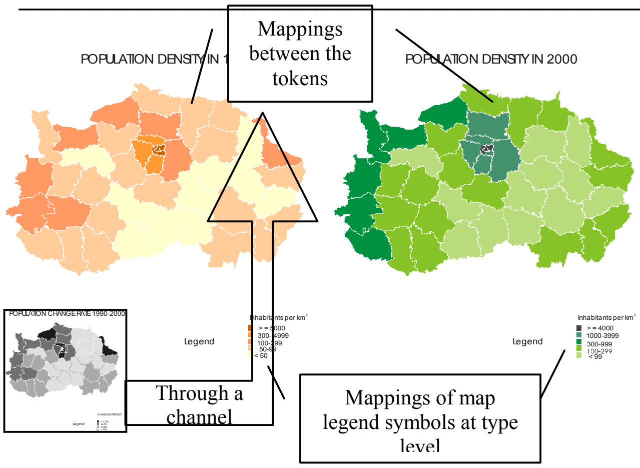

Figure 5. The interoperability aspect of information flow as applied to thematic maps

We give an example of IF-mappings between two symbol-sets of thematic maps, taken from Regions: Statistical Yearbook, (2002) and (2003), representing population density in a part of Europe at different times (years 1996, 1999) (Figures 6 & 7). In this task, we regard maps as classifications that are characterized by local logics as discussed in section 3.1.1. In the context of Information Flow, we consider classifications to be populated, which means instances of the categories must be included in the classification.

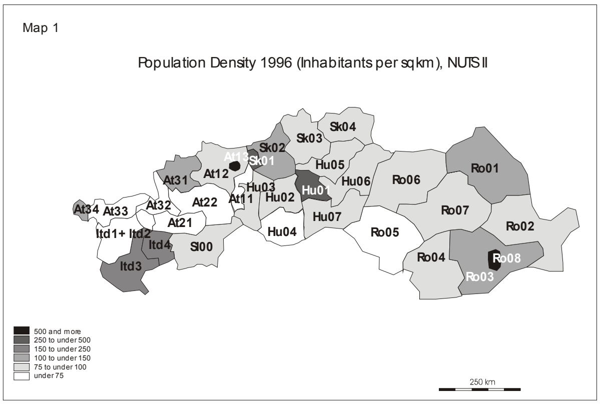

Figure 6. Population density in European regions (NUTSII) in 1996 (map_1)

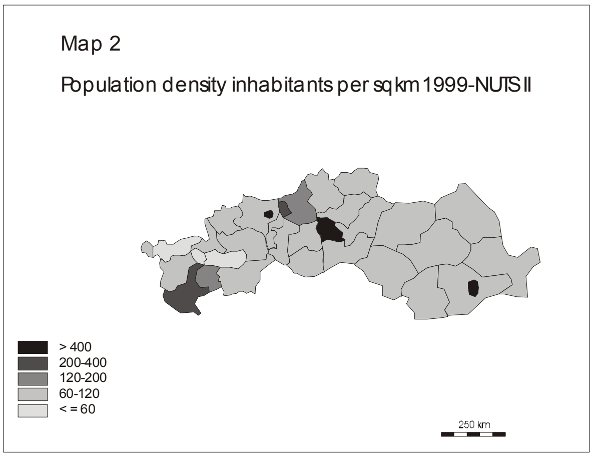

Figure 7. Population density in European regions (NUTSII) in 1999 (map_2)

The first map (Fig. 6) – map_1 represents population density (pd) in 1996 using six classes of pd. On the other hand, the second map (Fig. 7) – map_2 - represents the thematic concept of population density in 1999 (pd) using five classes of pd. As it stands, because of the different classifications, we are not able to reason whether population density of a region has increased or reduced over the three years time. Therefore, the goal is to produce mappings between these two different classifications to be able to draw secure inferences about the *phenomenon*. To generate these mappings in the framework of Information Flow, it is important to use a *channel* as described in the previous section.

Map_1 and map_2 are the two classifications; the legend symbols are the types of each classification, while the regions on the maps bearing the symbols are the tokens of the classification. The channel in this example is a third map, map_3, taken from Regions: Statistical Yearbook, (2003) (Fig. 8), which shows total population change rates between 1996 and 2000, in the same European regions.

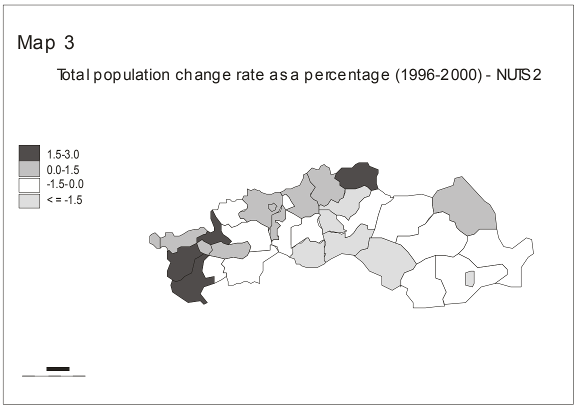
Map 3
Total population change rate as a percentage (1996-2000) - NUTS 2
1.5-3.0
0.0-1.5
-1.5-0.0
< = -1.5

Figure 8. The channel: map of total population change rate as a percentage for years 1996-2000 (map_3)

In our example, we have: different classifications of population density for different times for the same regions. In order to be able to compare these situations, we need a mapping from one classification to another in terms of types and tokens. The obvious relations for the classifications of map_1 and map_2 are shown in figure 9. Although these mappings at the type level are very easily generated, they do not hold at the token level because the values of population density are examined, herein, at different times.

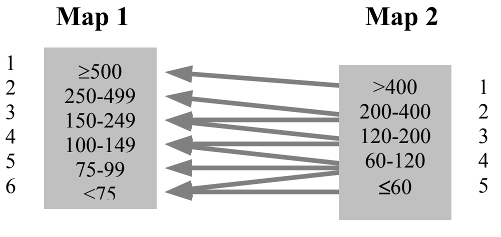

Figure 9. Mappings between classifications map_1 and map_2 at type level

For that reason, we need to find a way to compare these classifications at the token level as well. To do this, we need a source of information that is able to account for changes in population density across time that will serve as the channel within which the information between the two classifications can flow. As already mentioned, the channel in this case is map_3 showing population change rates between 1996 and 2000 for the study region (Fig. 9). Because of lack of other resources for this kind of information, we use map_3 as a channel assuming that *change rate of population is equally distributed within the four years period (A)*, assumption *A* falls in the case of background conditions discussed in section 3.2.

The steps that we follow are:

- With respect to map_3 and map_1 we calculate the population density in 1999 for the given regions
- Then we compare these to the population density deduced from map_2, and we end up with its true value for every region for the year.
- Finally, we establish mappings between the two classifications (column 3 of table 1).

Regions (NUTS II Nomenclature)	Relation	Class Map1 to Class Map 2
ITD3	Overlapping	4 TO 4
ITD4	Overlapping	4 TO 3
ITD1+ITD2	Overlapping	1 TO 2
AT33	Refinement	1 TO 1
AT34	Overlapping	3 TO 3
AT21	Refinement	1 TO 1
AT32	Overlapping	1 TO 2
AT22	Overlapping	1 TO 2
AT11	Overlapping	1 TO 2
AT31	Overlapping	3 TO 2
AT12	Inclusion	2 TO 2
AT13	Extension	6 TO 5
SK02	Overlapping	3 TO 3
SK03	Inclusion	2 TO 2
SK04	Inclusion	2 TO 2
SK01	Overlapping	5 TO 4
SL00	Inclusion	2 TO 2
HU03	Inclusion	2 TO 2
HU02	Inclusion	2 TO 2
HU04	Overlapping	1 TO 2
HU07	Inclusion	2 TO 2
HU01	Overlapping	5 TO 5
HU05	Inclusion	2 TO 2
HU06	Inclusion	2 TO 2
RO05	Overlapping	1 TO 2
RO06	Inclusion	2 TO 2
RO01	Overlapping	3 TO 2
RO02	Inclusion	2 TO 2
RO03	Overlapping	3 TO 2
RO04	Inclusion	2 TO 2
RO07	Inclusion	2 TO 2
RO08	Extension	6 TO 5

Table 1. Mappings between the instances (regions of the maps). The second column shows the relation that holds between map_1 and map_2 regarding the classifications' instances/ regions (token level). The third column shows the mappings at the instance (token) level

The result of the pursued procedure is that we transformed the classification of population density classes of map_1 into the classification of map_2. Consequently, we ended up with a classification of legend symbols of five (5) classes for map_1 identical to those of map_2 (Table 1). Furthermore, we established relations (Table 1) between the thematic contents of map_1 and map_2; namely, we provided mappings at the token level of the two classifications.

The resulting relations between the tokens of the two classifications can be described in terms of inclusion, overlapping, extension, and refinement. Inclusion is met in cases where a population density class of the first classification can be properly included in a population density class of the second (i.e., AT12 in Table 1). The case of overlapping holds when a part of a class of the first classification can be included in a class of the second (i.e., RO05). Extension regards expansion of the limits of the initial class (i.e., RO08), while refinement involves the opposite procedure when the limits of the initial class are confined (i.e., AT33). Recall at all times, however, that these relations hold only among the tokens of the two classifications.

5. FACTS, OPEN QUESTIONS, AND FUTURE RESEARCH

IF- Mappings can be easily generated in the case of thematic maps as the previous discussion has demonstrated. Future research comprises the application of information flow concepts in the field of ontologies. This has not been fully addressed, herein, because geographic ontologies are very hard to find, and even if this is the case, the majority of them is not populated with instances.

There is nevertheless a long way to go when trying to apply IF Theory to populated geographic ontologies. The problematic aspects of such an endeavour consist, among others, in several facts such as the following:

1. Ill-defined categories are more likely to include ill-defined instances.
2. Definition process produces less fuzzy results than categorization.
3. Relations among the tokens do not necessarily hold among the types.

The first aspect boils down to the fact, that whether integration at the intensional level has to tackle with not properly defined categories or not clear-cut taxonomies then these issues are likely to be inherited at the extensional level as well. This practically means that instances of ill-defined categories bear the ambiguity of the categories they belong to.

The second point distinguishes between the notions of categorization and definition. According to cognitive scientists and psychologists (Rosch, 1978) the process of assigning members to a category may result to overlapping categories, or categories with blurring edges. While, on the other hand, defining a category may has shortcomings like partial, or inadequate descriptions of a category, nonetheless, results in better demarcation among categories.

The third point can be easily clarified by the previous example. Figure 9 portrays the mappings at the intensional level, while the last column of table to shows the mappings at the extensional level. The second column of the table illustrates the relations that hold between the instances. These relations however do not hold at the intensional level, for extensional level relations are one-to-one, while intensional level ones can also be one-to-many, or many-to-one.

REFERENCES

Barwise J. and Seligman J., 1997. Information Flow. Cambridge University Press, Cambridge, England

Bremer M. and Cohnitz D. Textbook on Information and Information Flow - An Introduction, University of Dusseldorf, Faculty of philosophy, www.phil-fak.uni-duesseldorf.de/thphil/cohnitz/IF-BRM.pdf, (accessed 5 Mar 2005).

Cruz I. F., Sunna W., and Chaudhry A., 2004. Semi-automatic Ontology Alignment for Geospatial Data Integration in Max J. Egenhofer, Christian Freksa, Harvey J. Miller (Eds.): Geographic Information Science, Third International Conference, GIScience 2004, Adelphi, MD, USA, 2004, Proceedings. Lecture Notes in Computer Science 3234 Springer, ISBN 3-540-23558-2, pp 51-66

Devlin K., 2001. The Mathematics of Information, Lecture 4: Introduction to Channel Theory, 13th European Summer School in Logic, Language and Information, Helsinki, Finland, http://www.helsinki.fi/esslli/courses/readers/K1/K1-4.pdf (accessed 20 Jan. 2005).

Duckham M., Worboys, M.F., 2005. An algebraic approach to automated information fusion. *International Journal of Geographic Information Science,* 19(5), pp. 537-557.

Kalfoglou Y., and Schorlemmer M., 2002. Information Flow based ontology mapping, Proceedings of the 1st International Conference on Ontologies, Databases and Application of Semantics (ODBASE'02), California, USA,

Kalfoglou Y., and Schorlemmer M., 2003. IF-Map: an ontology mapping method based on Information Flow theory. *Journal on Data Semantics I*, LNCS 2800, pp. 98-127, Springer.

Kavouras M., A unified ontological framework for semantic integration, to appear in *Next Generation Geospatial Information* by A.A. Balkema Publishers - Taylor & Francis, The Netherlands, P. Agouris (ed). http://ontogeo.ntua.gr/publications/Kavouras_Marinos_01.pdf (accessed 25 Nov. 2005)

Kent R. E., 2001. The Information Flow Framework. Starter document for IEEE, the IEEE Standard Upper Ontology Working Group, http://suo.ieee.org/IFF/ (accessed 20 Jan. 2005)

Kent R. E., 2004. The IFF Foundation for Ontological Knowledge Organization, (N. J. Williamson and C. Beghtol, Eds.), Knowledge Organization and Classification in International Information Retrieval. Cataloging and Classification Quarterly. The Haworth Press Inc., Binghamton, New York. http://www.ontologos.org/Papers/CCQ/CCQ.pdf (accessed 20 Jan. 2005)

Kokla M. Kavouras M., 2000. "Concept Lattices as a Formal Method for the Integration of Geographic Ontologies" First International Conference on Geographic Information Science, Savannah, Georgia, USA.

Kokla M. and Kavouras M., 2002. "Extracting Latent Semantic Relations from Definitions to Disambiguate Geographic Ontologies", GIScience 2002, 2nd International Conference on Geographic Information Science, Boulder, Colorado, USA.

Lalmas M., 1998. The Flow of Information in Information Retrieval: Towards a general framework for the modeling of information retrieval, (F. Crestani, M. Lalmas and C.J. van Rijsbergen, Eds), Information Retrieval: Uncertainty and Logics - Advanced models for the representation and retrieval of information, Chapter 6, Kluwer Academic Publishers.

Old L. J., and Priss U., 2001. Metaphor and Information Flow, Proceedings of the 12th Midwest Artificial Intelligence and Cognitive Science Conference, pp. 99-104, Ohio, USA.

Regions: Statistical yearbook 2002, Collection: Panorama of the European Union, the Statistical Office of the European Communities (Eurostat), 2002.

Regions: Statistical yearbook 2003, Collection: Panorama of the European Union, the Statistical Office of the European Communities (Eurostat), 2003.

Rosch E., 1978. Principles of Categorization in E. Rosch, and B. Lloyd (Eds.): Cognition and Categorization.

Schorlemmer M., and Kalfoglou Y., 2003. On Semantic Interoperability and the Flow of Information, ISWC'03 workshop on Semantic Integration, pp. 80-86, Florida, USA.

Sowa J. F. Building, Sharing, and Merging Ontologies, http://www.jfsowa.com/ontology/ontoshar.htm (accessed 25 Nov. 2005)

E. Tomai & M. Kavouras, 2004. From "Onto-GeoNoesis" to "Onto-Genesis" The Design of Geographic Ontologies", *Geoinformatica*, 8(3): pp. 285-301.

M. F. Worboys, 2001. Communicating Geographic Information in Context, Meeting on Fundamental Questions in Geographic Information Science (draft) (M. Duckham, and M.F. Worboys, Eds.), pp.217-229, Manchester, UK. http://www.spatial.maine.edu/~worboys/mywebpapers/manchester2001.pdf.

WORDNET 2.0, 2003. A Lexical Database for the English Language, Cognitive Science Laboratory, Princeton University