Fuzzy Ontology-based Semantic Integration of Heterogeneous Data Sources in the Domain of Watershed Analysis

Vinícius R. T. Ferraz, Gustavo F. Afonso, Cristiane Yaguinuma,
Sérgio Borges, Marilde T. P. Santos

Department of Computer Science – Federal University of São Carlos (UFSCar)
13565-905 – São Carlos – SP – Brazil

{vinicius_ferraz,cristiane_yaguinuma,marilde}@dc.ufscar.br
gustavoferreiraafonso@yahoo.com.br, sergio@fatecriopreto.edu.br

Abstract. *In environmental research, data integration plays an important role given the increasing availability of heterogeneous data sources for specific features, such as soil type, climate, geographic location and so on. However, some of those features have inherently imprecise relationships and the lack of a suitable semantic model can be a major obstacle to its effective integration. In this context, this paper presents a semantic data integration system based on a fuzzy ontology, capable of perform query expansions based on imprecise aspects of real phenomena. We executed a real experience in the domain of watershed analysis. Consequently, watershed researchers obtained an homogeneous view of the data sources and more effective responses to their queries.*

1. Introduction

Like many environmental research fields, the watershed domain comprises an intricate set of parameters and relations, concerning not only water characteristics but also all its ecosystem. The research procedures of this field consider a wide range of activities such as collection of water samples, climatic evaluation, and other particular tasks that may focus on different influence factors and parameters. Hence, there is a great heterogeneity in watershed research and, consequently, in the generated data.

In this context, Data Integration Systems (DIS) [Halevy 2009] [Bleiholder and Naumann 2008] are an important instrument to support watershed researchers, by providing a homogeneous view of the data sources. Moreover, the adoption of a semantic model based on ontologies [Gruber 2009] can allow a DIS the solving of not only syntactic or structural issues, but also the reasoning over semantics of the watershed-related data.

Furthermore, as watershed research considers natural phenomena, some imprecise or vague information is usually required to describe them. For this purpose, representing fuzzy set concepts [Zadeh 1965] in ontologies is a feasible approach to express imprecise concepts and relationships, e.g. the similarity degree between two analysis parameters of water quality. By analyzing these fuzzy ontology constructs, it is possible to expand user queries and then retrieve relevant information [Yaguinuma et al. 2007a].

Given this scenario, we have developed DISFOQuE, a data integration system based on a global fuzzy ontology, which provides a homogeneous view of semantically heterogeneous data sources and also performs semantic query expansions, aiming to retrieve additional relevant results to user queries.

The remainder of this paper is organized as follows. Section 2 describes some characteristics of watershed research as well as its requirements to data integration. Section 3 presents more information on data integration systems and discusses the contributions of fuzzy ontologies to this field. Section 4 details DISFOQuE system, explaining its architecture and operation. Section 5 relates the application of DISFOQuE considering watershed analysis data. Finally, Section 6 describes conclusions and results from the experiments involving the domain of watershed analysis along with future work.

2. Watershed Analysis Research

By definition, a watershed, drainage basin or water basin is an extent of land drained by a river and its creeks, generally surrounded by water divides formed from geographic barriers like ridges, hills or mountains. Watershed analysis research is considered quite complex, since it regards not only the rivers constituting the drainage basin, but also a set of parameters and external relations that affect the main element of the basin - the water. Several influence factors are fundamental to this research field, such as use, occupation and composition of the soil; physical and chemical properties; topography; aquatic biodiversity; and so forth, encompassing a chain of relations among several different data [Dupas et al. 2006].

Geographic and temporal location are other relevant features in watershed research. Indeed, all water collection samples are registered with geographic coordinates (latitude and longitude) in order to identify regions, rivers and watershed they belong to. Regarding to temporal location, timestamps records support historical analysis as well as research on evolution and prediction of specific influence factors.

The development of this work was conducted in a partnership with International Institute of Ecology (IIE), so that it was possible to observe the typical work flow for watershed analysis research. First step is collection of water samples and their annotation. Following, researchers examine these samples, considering several factors (climate, soil characteristics, biodiversity). In the next step, generated data along with other information from external sources is combined to obtain spreadsheets associated with the sample set. Finally, researchers produce reports, maps, and databases based on analysis results.

Based on watershed analysis features and the methodology adopted by the researchers, it was possible to evaluate the main adversities for managing and integrating research data, such as: the lack of a suitable structure for data storage, hindering search and data correlation; insufficient data organization among several interrelated projects; data source heterogeneity, comprising structural and semantic diversity, the main problems to data integration; large and complex data sets due to the great amount of parameters and relations in the domain.

In this sense, data integration systems can be applied to handle the requirements pointed by watershed researchers in IIE. Next section describes more information on data integration systems and motivates the use of fuzzy ontology to enhance some of their features.

3. Using Fuzzy Ontology for Data Integration

Data integration systems (DIS) deal with two main problems: combining data located in different sources and providing the user with a unified view of gathered results

[Bleiholder and Naumann 2008]. Such features are required to search for desired information, since queries might be inappropriately answered or may have incomplete results if each data source is analyzed in isolation. By providing a unique, transparent and homogeneous view of heterogeneous data sources, it is possible to retrieve richer information, since different sources can have complementary data.

In order to build a integrated view of data sources, some conflicts must be addressed, such as syntactic (schematic and structural) [Rahm and Bernstein 2001] and semantic ones [Doan and Halevy 2005]. An approach to accomplish this task considers ontologies to provide a common and unambiguous understanding of terms and concepts of data sources, serving as a shared vocabulary for data integration [Noy 2004]. In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. These representational primitives are typically classes, attributes, and relationships [Gruber 2009], and can be used to model the semantics of the domain related to a set of data sources.

Several DIS-related works are based on ontologies [Fonseca et al. 2002] [Noy 2004] [Lenz et al. 2007] [Vidal et al. 2009], including applications in the environmental [Madin et al. 2008] and water resources domains [Latre-Abadía et al. 2009]. A common point in these systems is the use of traditional ontologies (*crisp* ontologies), which only consider predicates with "false" or "true" values. However, for representing real-world knowledge, including watershed characteristics, such ontologies are less suitable to express vague or imprecise information, so usual in human language. In this sense, fuzzy set concepts [Zadeh 1965] can be introduced in ontologies, so that one can build more suitable representations of imprecise aspects of real phenomena. For this purpose, several researches have been developed [Stoilos et al. 2006, Yaguinuma et al. 2007b, Lukasiewicz and Straccia 2008], aiming to define ontologies containing fuzzy concepts as well as fuzzy relationships, e.g. the *similarity* relationship. Besides providing a richer semantics, these fuzzy constructs can be analyzed to perform semantic query expansions, thus retrieving approximate results to user queries in DIS.

In order to illustrate how fuzzy ontologies can provide some relevant contributions to DIS, suppose that a user requests for information regarding a concept a and an ontology models the following fuzzy similarity relations: $similarTo(b, a) = 0.8$ and $similarTo(b, c) = 0.6$. As the similarity relation is a symmetric and transitive relation, it is possible to perform inferences based on symmetry ($similarTo(x, y) = similarTo(y, x)$) and max-min transitivity ($similarTo(x, z) \geq max_{y \in X} min[similarTo(x, y), similarTo(y, z)]$) axioms of the fuzzy ontology. So, if concepts a, b and c are distributed in distinct data sources and the user sets a similarity threshold of 0.5, all data sources will be searched to retrieve a and its similar results (b and c), since $similarTo(a, b) = 0.8$ and $similarTo(a, c) \geq 0.6$ are inferred from the fuzzy ontology. On the other hand, if the user sets the threshold to 0.7, only b should be considered as a similar result, since we cannot affirm for sure that $similarTo(a, c) \geq 0.7$ (it may assume a similarity degree between 0.6 and 0.7). Therefore, analyzing fuzzy ontologies in DIS not only increases recall, as more relevant results are covered, but also improves precision, by pruning answers that may not satisfy user requirements. Such fuzzy concepts and relationships are considered by the DISFOQuE system, a DIS that employs fuzzy ontologies to retrieve approximate results from heterogeneous data sources. Next section presents more details on DISFOQuE system and its main features.

4. DISFOQuE Architecture

DISFOQuE follows a mediated architecture, which implements a virtual approach to access the heterogeneous data sources. An overview of the mediated approach for information integration can be seen in [Halevy 2009]. The Figure 1 shows the architecture. The sets of elements separated by dashed lines comprise the layers of the architecture and the numbered arrows indicate the operational flow.

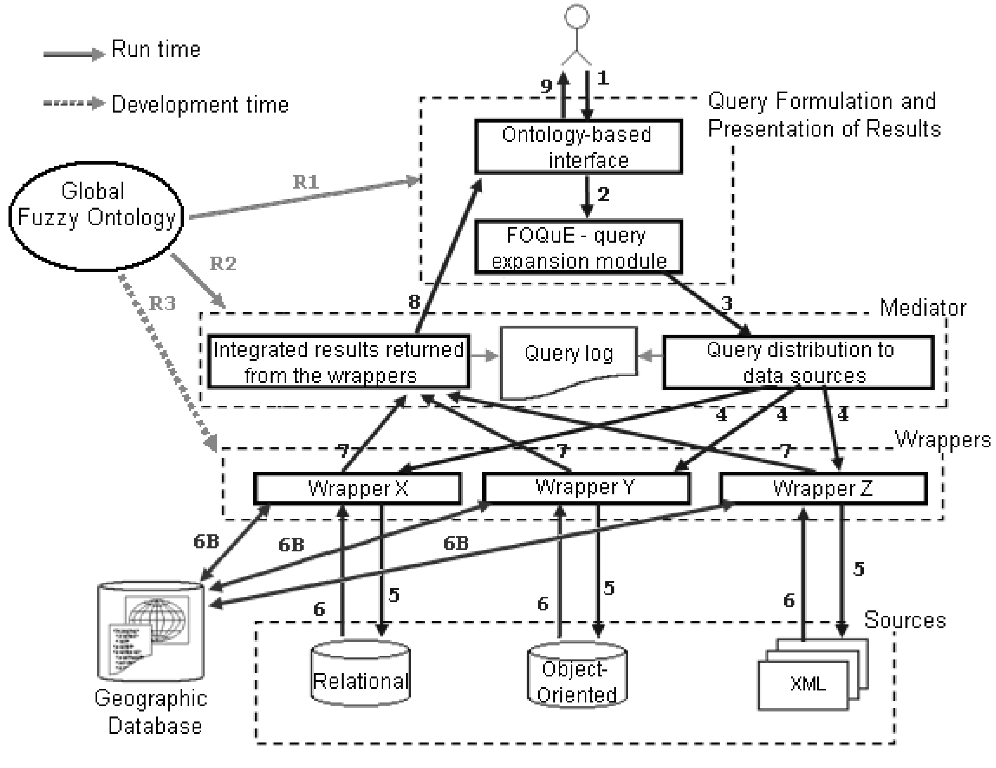


Figure 1. DISFOQuE Architecture.

The starting point is query formulation in user interface (1). It is typed in a text box, where the terms and value restrictions are denoted by relational operators (=,>,<) and logical operators (AND, OR). The adopted query structure allows users to input expressions more naturally than writing SQL or XQuery statements directly. Once the query is submitted to DISFOQuE, it is analyzed to check possible syntactic errors and homonyms in relation to the global ontology vocabulary. After the necessary corrections to the query, it continues to the FOQuE query expansion module (2). In this module, the query is modified if both fuzzy ontology constructs and user-defined expansion parameters indicate that semantic expansions are needed. For further information on semantic query expansion, see [Yaguinuma et al. 2007a].

In the next step, mediator receives the query (3), creates a thread for each wrapper and distributes the query to them (4). All wrappers keep two important types of XML documents containing: information on how to connect to its respective data source; and mappings for associating each ontology concept with a corresponding term in the data source. These mappings are an essential element that supports an accurate query translation, since mapping rules are written depending on the schema and the query language considered by a data source, where SQL is used for data sources in the relational model and XQuery for XML data. During the query translation, wrappers handle some heterogeneity problems, such as naming, lack of data, attribute, value, scale and identifier conflicts. After the query is translated, it is submitted to each respective data source (5).

As soon as results are returned from the data sources (6), wrappers check the mappings again to verify if there are terms related to geographic data. In this step, specific functions offered by the geographic database are called by the wrappers for filtering the results and discarding those whose coordinates that does not relate to the geographic terms in the query. Thus, standardized geometries are provided as base features, such as places and regions mentioned in the query. In the case of watershed analysis, such features would be boundaries of lakes, rivers, drainage basins, states and cities.

Finally, the mediator awaits the feedback from all queried wrappers. Once all results have been received (7), they are integrated and interrelated, checking for equalities and gathering them in a single tabular result. From the mediator layer, this tabular result is sent to the Query Formulation and Presentation of Results layer (8). Subsequently, it can be properly formatted and presented to the user (9), concluding the query flow.

5. Applying DISFOQuE in the Domain of Watershed Analysis

In an effort to handle requirements identified in IIE (Section 2), we have applied DISFOQuE system for integrating data sources related to watershed analysis research. The main goal of this experience was to provide an integrated environment so that researchers could conduct their work in a more efficient manner, by supporting an unified view of watershed factors and related geographic issues. Furthermore, researchers could be benefited from semantic query expansions, since more relevant results could be retrieved. Next subsections describe key points of this study case: the watershed analysis ontology and some queries highlighting DISFOQuE contributions to watershed studies.

5.1. Watershed Analysis Ontology

The first step of DISFOQuE deployment is the definition of a global fuzzy ontology, describing the semantics related to watershed analysis data sources. For this reason, we have defined a Watershed Analysis ontology with support of domain experts, the watershed researchers of IIE. The Watershed Analysis ontology was developed according to Methontology method [Lopez et al. 1999] and codified in OWL [Smith et al. 2004].

Figure 2 shows part of the modeled ontology. *Total nitrogen* and *Ammonia* are instances of different types of *Analysis Parameter*, which represents parameters collected in water, soil and air samples. *Miligrams per litre* is an instance of *Unit of Measurement*, which is associated to *Total nitrogen* through the property *hasUnitOfMeasurement*. Besides this property, *Analysis Parameters* has labels representing the alternative ways that a concept can be expressed. For example, *DTN* can be used instead of *Total nitrogen* on query formulation. An example of a fuzzy relationship in this ontology is the *similarTo* relation, which establishes a similarity degree of 0.65 between *Total nitrogen* and *Ammonia* parameters. This semantic construct can be analyzed by DISFOQuE in order to choose which databases should be queried (see Section 5.2 to see an example).

Furthermore, the Watershed Analysis ontology represents many other features, with special interest in water resources, like rivers, lakes and dams. Some aspects regard-

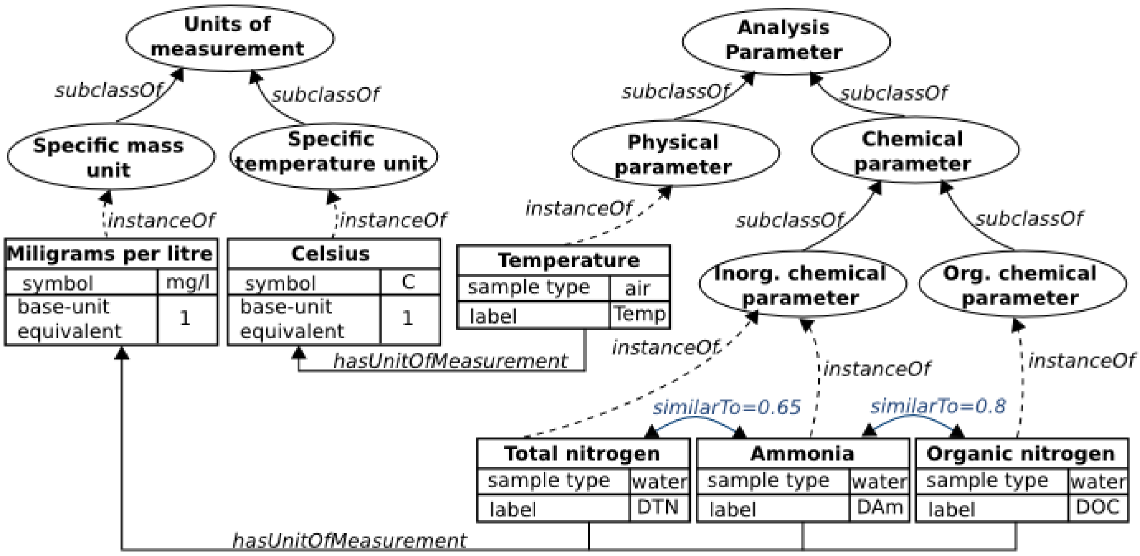


Figure 2. An extract of the watershed analysis ontology.

ing biodiversity (e.g. fishes, bentonitic organisms, planktons) and uses of soil are also considered.

The main challenge while building the ontology was to retrieve and explore a huge amount of non-digital documentation and technical protocols of watershed analysis research. Domain expert support was fundamental, making the definition and modeling of these concepts and relationships a less arduous task.

5.2. Query Expansion and Geospatial Processing in Watershed Data Sources

Once the global ontology has been defined, watershed data sources used by IIE were registered in DISFOQuE system. We used relational (MySQL and PostgreSQL) and semi-structured (in XML files) data sources, in a total of 5 sources with approximately 16000 non-integrated registers. All data sources have references to geographic information, while some specifically contain hydrological parameters, data on the use and occupation of the soil, among others.

Consider the query Q_1 in which researchers require information on the total of nitrogen and temperature in Tietê river:

Q_1: `Total nitrogen AND temperature AND river = Tietê`

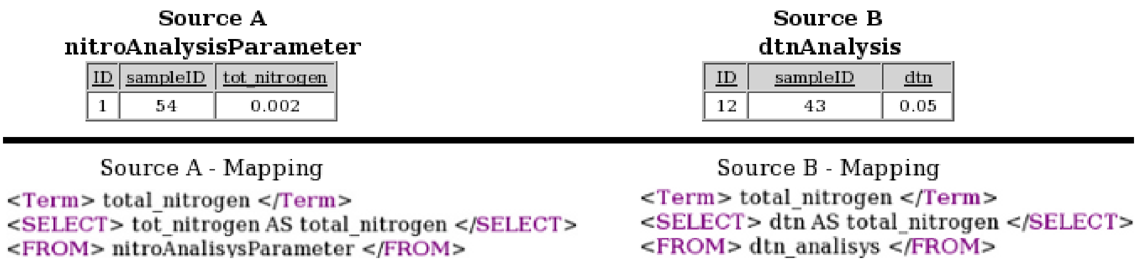
First step is analyzing Q_1 so that it can be rewritten in a common vocabulary language provided by the global ontology. In this step, the global ontology is analyzed to detect that `temperature` term can be associated with both `1 - temperature of the air` and `2 - temperature of the water` concepts. The system interacts with the user so that one option is selected; suppose that in this case the user has chosen the second one.

Supposing that researchers defined a $minSimilarity = 0.6$, the next step is when FOQuE query expansion module analyzes the global fuzzy ontology in order to check if it is possible to perform semantic expansions. According to Figure 2, the global fuzzy ontology has a similarity degree of 0.65 between *Total nitrogen* and *Ammonia*, which is higher than $minSimilarity$. Moreover, FOQuE also found similarity between *Total nitrogen* and *Organic nitrogen* through the transitivity property of the fuzzy relationships, since *Ammonia* and *Organic nitrogen* have a similarity degree of 0.8. Then, FOQuE module rewrites the original query, so that *Ammonia* and *Organic nitrogen* can be retrieved as a semantically similar result:

```
(Total nitrogen OR Ammonia OR Organic nitrogen)
 AND Temperature of the air AND river = Tietê
```

It is important to highlight that if data sources only stored information on *Organic nitrogen* term, the original query would not return answers, whereas the rewritten query is able to retrieve relevant results due to semantic query expansions.

When query reaches the mediator, it is redistributed to the registered wrappers, where XML documents containing data mappings are inspected. Figure 3 considers part of the mappings of two relational data sources. *Total nitrogen* concept is represented in *Source A* by the *tot_nitrogen* attribute, while *Source B* employs the *DTN* term. Such naming conflict is extremely common, since each data source may have a different schema attributes. The procedure is the same for the expanded terms and for *Temperature of the air*.



Source A
nitroAnalysisParameter

ID	sampleID	tot_nitrogen
1	54	0.002

Source B
dtnAnalysis

ID	sampleID	dtn
12	43	0.05

Source A - Mapping

```
<Term> total_nitrogen </Term>
<SELECT> tot_nitrogen AS total_nitrogen </SELECT>
<FROM> nitroAnalisysParameter </FROM>
```

Source B - Mapping

```
<Term> total_nitrogen </Term>
<SELECT> dtn AS total_nitrogen </SELECT>
<FROM> dtn_analisys </FROM>
```

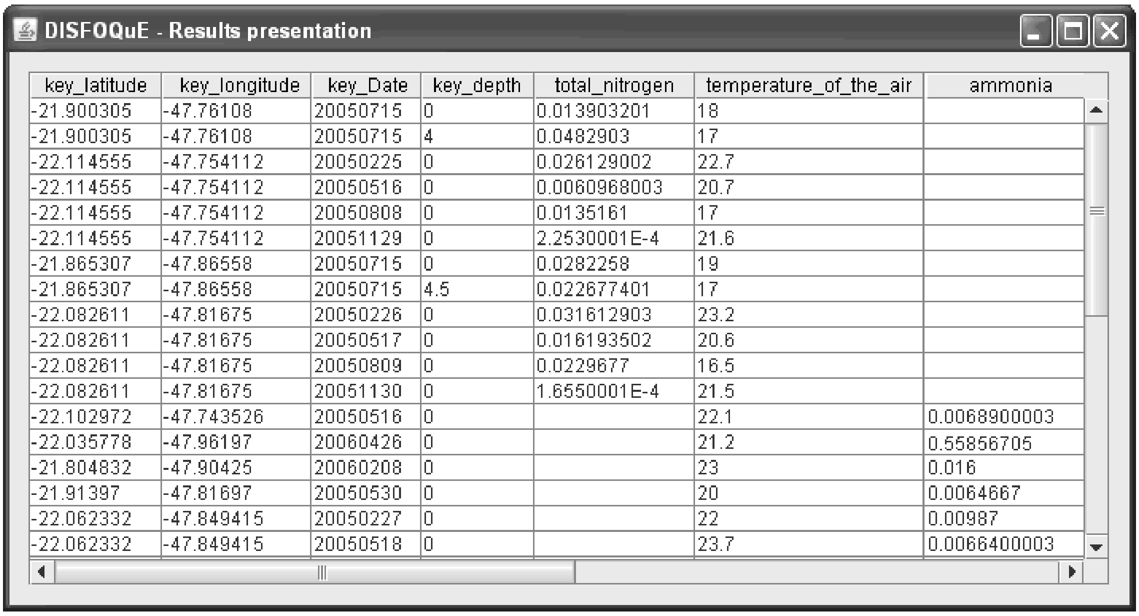
Figure 3. Mapping of the *Total nitrogen* concept for two relational data sources.

If a wrapper checks its mappings and they do not contain any of the terms of the query, the database-specific query will not contain this term as well. If no term is found, the wrapper will not activate its respective database and then closes the connection with mediator. After the mapping inspection, a rewritten query is sent to the sources A and B:

```
Source A: SELECT tot_nitrogen AS total_nitrogen AND ...
     Source B: SELECT dtn AS total_nitrogen AND ...
```

With regard to the term *river* in Q_1, wrappers are responsible for identifying it as a geographic concept. The mapping distinguishes that such term needs geospatial processing by the additional element `<ProcessingType> Geospatial </ProcessingType>`. Thus, after the data source answered the query, the geographic database is accessed by the wrappers in order to perform two operations: checking which of the returned coordinates are contained in the Minimum Bounding Rectangles (MBR) of the geographic features labeled as *Tietê river*; then checking a buffer stated to these features (6B arrows in Figure 1). Finally, results are sent back to mediator, where resultsets are merged.

The integrated results are then presented to the user in a tabular format, including values for the original query as well as values for expanded terms. Figure 4 shows the result presentation interface, both related to Q_1 example.



key_latitude	key_longitude	key_Date	key_depth	total_nitrogen	temperature_of_the_air	ammonia
-21.900305	-47.76108	20050715	0	0.013903201	18	
-21.900305	-47.76108	20050715	4	0.0482903	17	
-22.114555	-47.754112	20050225	0	0.026129002	22.7	
-22.114555	-47.754112	20050516	0	0.0060968003	20.7	
-22.114555	-47.754112	20050808	0	0.0135161	17	
-22.114555	-47.754112	20051129	0	2.2530001E-4	21.6	
-21.865307	-47.86558	20050715	0	0.0282258	19	
-21.865307	-47.86558	20050715	4.5	0.022677401	17	
-22.082611	-47.81675	20050226	0	0.031612903	23.2	
-22.082611	-47.81675	20050517	0	0.016193502	20.6	
-22.082611	-47.81675	20050809	0	0.0229677	16.5	
-22.082611	-47.81675	20051130	0	1.6550001E-4	21.5	
-22.102972	-47.743526	20050516	0		22.1	0.0068900003
-22.035778	-47.96197	20060426	0		21.2	0.55856705
-21.804832	-47.90425	20060208	0		23	0.016
-21.91397	-47.81697	20050530	0		20	0.0064667
-22.062332	-47.849415	20050227	0		22	0.00987
-22.062332	-47.849415	20050518	0		23.7	0.0066400003

Figure 4. Results retrieved by DISFOQuE system for Q_1.

With regard to implementation details, DISFOQuE was developed using Java, since it is multi-platform, including support to the concurrent programming. The Jena framework [Carroll et al. 2004] was also used to make inferences from the global ontology written in Web Ontology Language (OWL) [Smith et al. 2004]. The geographic database was built using PostgreSQL DBMS and its geospatial extension PostGIS, which follows the specifications recommended by Open Geospatial Consortium (OGC) [OGC 2008].

Tests were performed in a computer using an Athlon64 X2 4200 processor, with 2 GB of RAM memory and Linux operational system. We observed an average for query answering time of 15 seconds for most queries, reaching 347 seconds when the more complex query was executed, regarding over 26700 registers of integrated data.

6. Conclusion and Future Work

DISFOQuE overcomes the limitations of most DIS based on crisp ontologies, by handling imprecise information which are inherent in many domains. Thus, the main contribution of the DISFOQuE's approach is supporting the use of fuzzy ontologies to represent such semantics and then retrieve relevant integrated information that originally would be hidden in the data sources.

Our case study shows a real application of DISFOQuE in the watershed domain. Watershed researchers concluded that DISFOQuE provided a valuable support to achieve an integrated management of drainage basins and water resources, making it possible to access several data sources through a homogeneous view. Moreover, performed query expansions retrieve more relevant integrated data, consequently supporting researchers in building strategic reports and spreadsheets in less time. In this case study we have also modeled an ontology for the watershed analysis domain, built with supervision of watershed researchers. This ontology can support further research in computational solutions for this field.

As future directions, we intend to include a geovisualization module and reduce the response time of complex queries by materializing frequently searched data. Furthermore, the use of fuzzy ontology in DIS is the starting point to a wide range of possible improvements in this kind of system, such as sorting results by relevance concerning fuzzy concepts and relationships; query formulation with fuzzy linguistic terms such as "high", "low", "large", "small" and fuzzy modifiers like "very", "slightly", typical of natural language; among others.

7. Acknowledgments

We thank CAPES and INEP for providing funds to the WebPIDE project under the Education Observatory grant. We also thank CNPq and FAPESP for supporting correlated projects; and International Institute of Ecology (IIE) for providing a suitable environment to the case study.

References

Bleiholder, J. and Naumann, F. (2008). Data fusion. *ACM Computing Surveys*, 41(1):1–41.

Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., and Wilkinson, K. (2004). Jena: implementing the semantic web recommendations. In *Int. WWW Conf.*, pages 74–83, New York, NY, USA. ACM.

Doan, A. and Halevy, A. Y. (2005). Semantic-integration research in the database community. *AI Mag.*, 26(1):83–94.

Dupas, F. A., Souza, A. T. S., Tundisi, J. G., Tundisi, T. M., and Rohm, S. A. (2006). Indicadores ambientais para planejamento e gestão de bacias hidrográficas. In *Eutrofização na América do Sul: causas, conseqüências e tecnologias para gerenciamento e controle*, pages 491–506. IIE, São Carlos, SP [in portuguese].

Fonseca, F. T., Egenhofer, M. J., Agouris, P., and Câmara, G. (2002). Using ontologies for integrated geographic information systems. *Transactions in GIS*, 6(3):231–257.

Gruber, T. R. (2009). Ontology. In *Encyclopedia of Database Systems*, pages 1963–1965. Springer.

Halevy, A. (2009). Information integration. In *Encyclopedia of Database Systems*, pages 1490–1496. Springer.

Latre-Abadía, M. Á., Lacasta, J., Mojica, E., Nogueras-Iso, J., and Zarazaga-Soria, F. J. (2009). An approach to facilitate the integration of hydrological data by means of ontologies and multilingual thesauri. In *AGILE Conf.*, pages 155–171.

Lenz, R., Beyer, M., and Kuhn, K. A. (2007). Semantic integration in healthcare networks. *Int. Journal of Medical Informatics*, 76(2-3):201 – 207. Connecting Medical Informatics and Bio-Informatics - MIE 2005.

Lopez, M., Gomez-Perez, A., Sierra, J., and Sierra, A. (1999). Building a chemical ontology using methontology and the ontology design environment. *IEEE Intelligent Systems and their Applications*, 14(1):37–46.

Lukasiewicz, T. and Straccia, U. (2008). Managing uncertainty and vagueness in description logics for the semantic web. *Journal of Web Semantics*, 6(4):291–308.

Madin, J. S., Bowers, S., Schildhauer, M. P., and Jones, M. B. (2008). Advancing ecological research with ontologies. *Trends in Ecology & Evolution*, 23(3):159–168.

Noy, N. F. (2004). Semantic integration: a survey of ontology-based approaches. *SIGMOD Record*, 33(4):65–70.

OGC (2008). Ogc reference model - ogc 08-062r4. <http://www.opengeospatial.org/standards/orm>. Access date: Dec. 2008.

Rahm, E. and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350.

Smith, M. K., Welty, C., and Mcguiness, D. L. (2004). Owl web ontology language guide. <http://www.w3.org/TR/2004/REC-owl-guide-20040210>. Access date: Jan. 2006.

Stoilos, G., Simou, N., Stamou, G., and Kollias, S. (2006). Uncertainty and the semantic web. *IEEE Intelligent Systems*, 21(5):84–87.

Vidal, V. M., Sacramento, E. R., Macêdo, J. A., and Casanova, M. A. (2009). An ontology-based framework for geographic data integration. In *Proc. of the ER 2009 Workshops*, pages 337–346, Berlin, Heidelberg. Springer-Verlag.

Yaguinuma, C. A., Biajiz, M., and Santos, M. T. P. (2007a). Sistema foque para expansão semântica de consultas baseada em ontologias difusas. In *XXII Brazilian Symposium on Databases*, pages 208–222, João Pessoa, PB. [in portuguese].

Yaguinuma, C. A., Santos, M. T. P., and Biajiz, M. (2007b). Meta-ontologia difusa para representação de informações imprecisas em ontologias. In *Workshop on Ontologies and Metamodeling in Software and Data Engineering*, pages 57–67, João Pessoa, PB. SBC. [in portuguese].

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.