

## **Water Protection Information Management by Syntactic and Semantic Interoperability of Heterogeneous Repositories**

Domenico Gendarmi<sup>1</sup>, Filippo Lanubile<sup>1</sup>, Oriana Licchelli<sup>3</sup>, Giovanni Semeraro<sup>1</sup>, Attilio Colagrossi<sup>2</sup>

<sup>1</sup> Dip. di Informatica, University of Bari

<sup>2</sup> Dip. Tutela delle Acque Interne e Marine, APAT

<sup>3</sup> ESEO 4

**Abstract.** One of the most prominent themes concerning environment is the inner and marine water protection. The information required for developing such a theme is very large, having to cover numerous and various fields of knowledge: geography, chemistry, hydrology, geology, biology, physics, economics and social sciences. The national Institutions devoted to cope with this theme, such as the Italian Agency for Environmental Protection and Technical Services (APAT), have to solve two fundamental problems: how to collect such a large mass of information, and how to access the information and retrieve coherent data to process in order to perform easy, fast and reliable decision making. The first problem has been solved by APAT in successive steps, as the technology progressively changed, producing large collections of data stored in several technologically distinct repositories. However, accessing data sources individually and then combining the results manually every time an information is needed can be awfully time-consuming.

In this paper we present a project which addresses interoperability of environmental information systems both at the syntactic and semantic level. The former is addressed through a federated database system which provides a global view of independent data sources. The latter is addressed by building an ontology upon the federated schema to enable autonomous information systems to share domain knowledge. The goal of the project was to develop an information broker which provides a full and user-transparent integration of the heterogeneous data sources maintained by APAT, ensuring, at the same time, the existing legacy applications that operates on them to continue operating autonomously, without undergoing any sort of modification. The information broker is also

augmented with a domain ontology, which facilitates knowledge sharing among different departments, and a web ontology browser for the online navigation of the APAT domain.

**Keywords:** environment protection, water information, heterogeneous repositories, interoperability, ontologies, federated databases

## 1. Introduction

One of the most prominent themes concerning environment is the inner and marine water protection. The information required for developing such a theme is very large, having to cover numerous and various fields of knowledge: geography, chemistry, hydrology, geology, biology, physics, economics and social sciences. As an example, it is almost obvious that the protection of inner and marine water requires to consider chemical and biological pollutant agents coming from anthropic activities, such as sea navigation, industrial factories and agricultural farms, but also phenomena such as river floods and coastal erosion that depend much more by the geomorphological, hydrographical and physical characteristics of the territory. The Institutions devoted to cope with this theme, have to solve two fundamental problems: how to collect such a large mass of information, and how to access the information and retrieve coherent data to process in order to perform easy, fast and reliable decision making.

The centrality of such a theme is also asserted by the European Community, which has launched since 2003 the WISE project (Water Information System for Europe)<sup>1</sup> as a joint initiative of DG Environment, The European Environment Agency (EEA), Eurostat (ESTAT) and the Joint Research Centre (JRC) in order to implement the data upload, sharing and analysis requirements of the Water Framework Directive 2000/60/CE. As an information system WISE includes all possible WISE nodes, data and viewer providers as well as the common WISE public web site and their interactions. It is not a centralized database but rather a decentralised system at EU level which will have capabilities to interoperate with existing national systems. It is planned that WISE will be fully operational by 2010.

In this paper we report our experience in developing an information broker for the Italian Agency for Environmental Protection and Technical Services (APAT)<sup>2</sup>. Our approach addresses interoperability both at the syntactic and semantic level [Kajan and Stoimenov, 2005; Park and Ram, 2004]. The former allows underlying systems to exchange information, resolving technical issues such as data integration. The latter ensures that exchanged information is meaningful for any cooperating applications, including those not initially conceived to work together. We address syntactic interoperability through a federated database system, which provides a global view of independent data sources [Sheth and Larson, 1990]. All the information needed can be

---

<sup>1</sup> WISE: Water Information System for Europe,  
<<http://wise.jrc.cec.eu.int/wfdview/php/index.php>>.

<sup>2</sup> APAT: Agenzia per la Protezione dell'Ambiente e per i Servizi Tecnici,  
<<http://www.apat.gov.it/site/it-IT/>>.

accessed uniformly, transparently and independently of the physical storage location, as if it was stored into a single data source. Semantic interoperability, on the other hand, is addressed by building an ontology upon the federated schema to enable autonomous information systems to share domain knowledge.

The remainder of this paper is structured as follows. Section 2 presents the project domain. Our solution for retrieving coherent data is described in Section 3. Finally, conclusions and future works are drawn in the last section.

## **2. Application Domain**

APAT carries out scientific and technical activities in the national interest to protect the environment, water resources and the soil. It is subject to the supervision of the Italian Ministry of the Environment and Territorial Protection and is integrated into the Environmental Agency System, which provides consulting services and support to other governmental agencies. Moreover, APAT is structured in several departments, which are themselves divided into smaller units, each offering intradepartmental services to the Agency.

Currently, the whole Agency manages six main repositories about water information:

- The hydrological time series, which refer to measures of rain, temperature and river levels on the Italian territory for about eighty years;
- The hydrological real time monitoring, which refers to real time monitoring of rain, temperature and river level;
- The hydrological reports, which describe the most relevant flood events occurred in Italy;
- The state of Italian rivers, which reports the main characteristics of the Italian rivers;
- The quality of the inner and marine waters, which refers to the monitoring of rivers, lakes, coast line and seas, in order to detect the degree of pollution deriving from anthropic activities;
- The forecast of meteorological parameters useful to predict important phenomena, such as, for example, floods, heavy sea, the level of the Venice Lagoon.

These repositories have been realized in many years using heterogeneous technologies, depending on the kind and the size of information to process as well as on the information technology available [Colagrossi, 2003]. The repositories used to store hydrological measures are relational data bases implemented in MySQL; the Hydrological Reports and the state of the Italian rivers are stored in XML repositories implemented using Tamino XML Server; the information on water quality is stored in a relational data base implemented in PostgreSQL and, finally the meteomarine forecast parameters are stored in compressed files and managed through the Unix file system.

Although all the information is owned by the same organization, the huge amount of information gathered is managed by different departments and units. Besides, given the large diversity in syntax and semantic of the data collected, measures are stored into several independent systems. While the construction of such information repositories has been solved in successive steps as the technology progressively changed, at the present the major need of the Agency is to extract

coherent data from these several and technologically heterogeneous repositories [Colagrossi and Gasparri, 2003].

The key word here is ‘coherent data’, that could be expressed as ‘every information concerning the same concept, meaning’. As an example, if we are investigating for a possible imminent flood in a given river basin X, then all the following data related to the concept ‘flood’ have to be retrieved: hydrological real time and time series measures for ‘river basin X’, hydrological reports for ‘river basin X’, water quality for ‘river basin X’, forecast of meteorological parameters for ‘river basin X’, and so on. The problem of retrieving coherent data has been approached by developing an Information Broker as shown in the next section.

### **3. Enabling Interoperability of Environmental Information Systems**

The goal of the presented project was to develop an information broker to provide an integrated and user-transparent access to all the heterogeneous data sources available within APAT. The developed Information Broker is a Federated Database System, enhanced with a set of web services which ensure the provision of data on demand, whilst keeping existing legacy applications to continue operating autonomously, without undergoing any sort of modification [Calefato et al., 2006]. The Information Broker is also augmented with a domain ontology, which facilitates knowledge sharing among different departments.

#### *3.1 Information Broker Architecture*

Figure 1 shows the overall architecture of the Information Broker: in the Wrapper Layer, a Data Access Service (DAS) has been developed to wrap each data source available and to extract the information required on demand. Each DAS is a wrapper implemented as a web service, that provides a WSDL interface to allow the remote invocation of the service via SOAP. The information about how actually it is possible to gain access to the data sources is hidden inside the DAS, and the components at an upper layer have a unique way to access the data. This solution ensures a transparent access to all distributed, heterogeneous and autonomous repositories available, solving the technical differences among them.

The Federation Layer offers a uniform and transparent access to the data stored in data sources through the Query Processor and the Federated Schema Browser components. The Query Processor performs the task of decomposing a global query in a set of local queries and integrating all the obtained results in a single response. This component does not know how data sources have to be actually queried, because this information is hidden in the Wrapper Layer. The Federated Schema Browser provides a high-level access to the federated schema, and is used by the Query Processor to discover the appropriate DAS which, in turn, provides access to a specific concept.

The Presentation Layer represents the communication medium between the broker and the end-users. It consists of two main components: the User Interface and the Ontology. Two different user-interfaces have been developed: an hydrological query wizard, used to perform global queries and view consequent results in a common web browser, and a web ontology browser, enabling users to navigate through the hydrological concepts within the APAT domain.

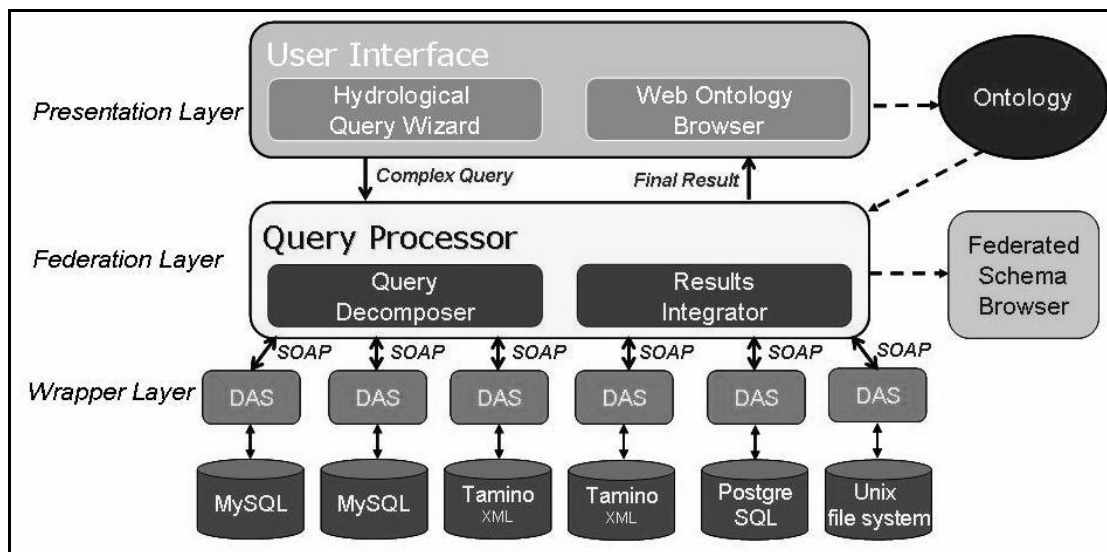


Figure 1. The Information Broker

### 3.2 Interoperability at the syntactic level

The global view of the distributed data is achieved through the creation of a federated schema, according to the four-step bottom-up process depicted in Figure 2.

The local schemas of the different component databases within the federation represent the starting point for building a federated schema. Due to the heterogeneity of the underlying data models, the first step in the integration process is to transform the local schemas into so-called export schemas, expressed in a common data model (CDM). Several options were weighted up to choose the most appropriate CDM, including relational data model, XML and OWL. Our choice was to use the XML data model because today it is the de-facto standard language to exchange information between applications and it is supported by most of the existing DBMS.

Once the data model heterogeneity is overcome, the next step is the creation of the export-schema mappings, which are XML files manually generated at design time from each export schema. Such files contain the mappings between the local and export schemas, that is, the correspondences between low-level data stored and high-level domain concepts.

The third step in the integration process is the construction of the federated schema, which is supposed to represent the logical model of the virtual database containing all the data available within the federation. The federated schema is the result of the merging of all the export schemas.

During the merge, all the possible conflicts have to be identified and solved. This is accomplished through two distinct activities. The Correspondence Investigation activity searches

for correspondences among the export schemas. The output of this activity is a set of conflicts, grouped in:

- naming conflicts, i.e. either different names are used to identify the same concept, or the same name is used to identify different concepts;
- structural conflicts, i.e. the same concept is represented with different structures in different schemas.

This set of conflicts is the input of the next activity, the Conflict Resolution. For naming conflicts where the same concept is identified with distinct names, a common name to use in the federated schema is chosen. For the other kind of naming conflicts the name of one of the two distinct concepts with the same name is changed. For structural conflicts, instead, each conflict is solved defining a new ad-hoc structure.

Once the federated schema has been obtained, the last step in the process is to manually generate the federated-schema mapping file, an XML file that stores the correspondences between: complex concepts and simple concepts distributed in the different export schemas; simple concepts and constraints that characterize them; simple concepts and services able to retrieve them.

A complex concept is a concept that can be decomposed in a set of simple concepts. A simple concept, instead, is an atomic concept which cannot be further decomposed and has to be instantiated by a set of constraints.

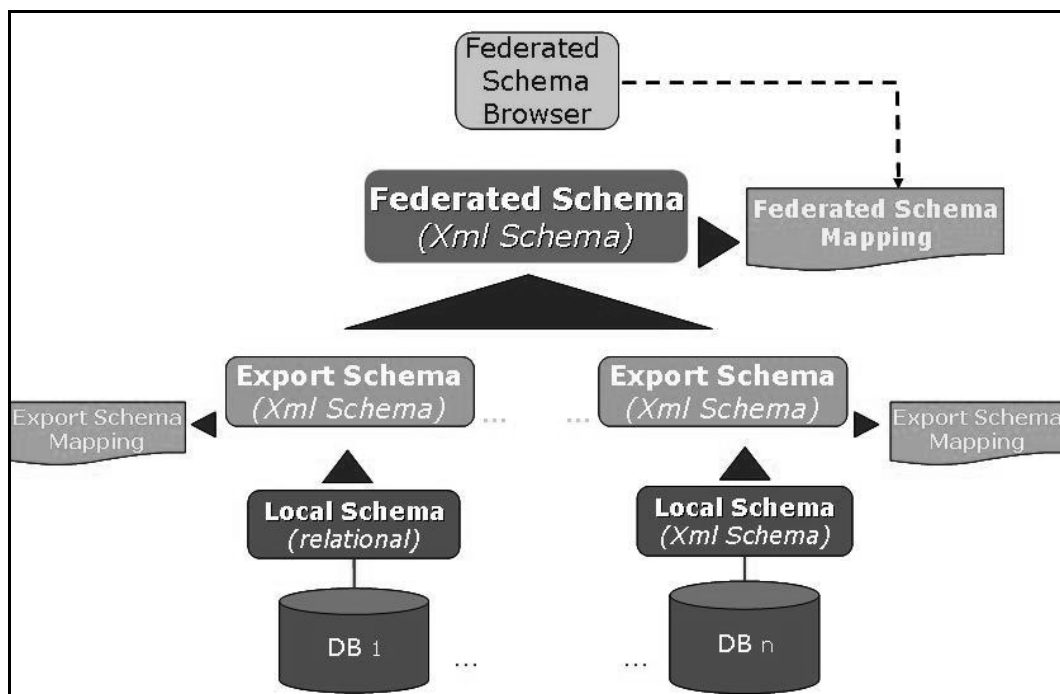


Figure 2. Schema Integration Process

### 3.3 Interoperability at the semantic level

Ontologies allow to obtain a shared and common understanding of a domain, that can be exploited by people as well as applications [Uschold and Grüninger., 2004]. The Information Broker exploits an Ontology API for sharing the APAT knowledge.

Figure 3 shows the relation between the Ontology API and two components named Domain Ontology and Ontology Schema Mapping: The Domain Ontology formalizes the main features of both water basins and APAT points of survey (data in the Federated System). The Ontology Schema Mapping, defines which individuals (ontology instances) can be derived from the results of queries to the Federated System and how to create them. These queries are defined according to the instructions of the Federated Schema Mapping and are executed by the Query Processor. Other individuals, not mentioned in the Ontology Schema Mapping, are embedded in the code of the Domain Ontology.

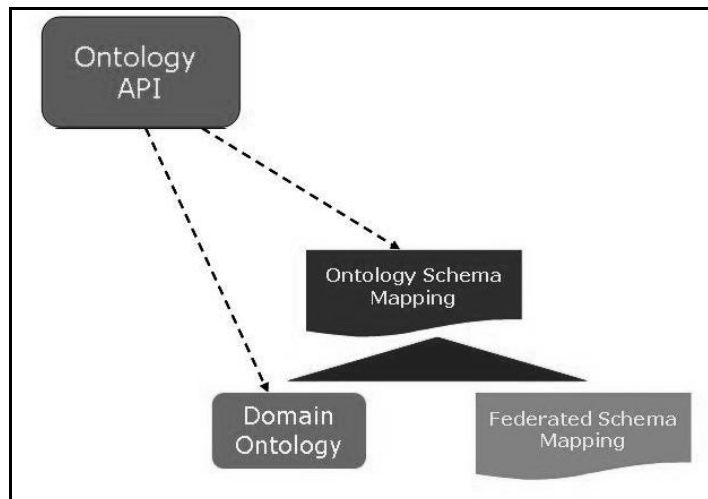


Figure 3. Ontology components

The Domain Ontology, which has been developed following the principles of the 101 Methodology [Noy and McGuinness, 2001], contains knowledge about topics concerning waters and covers aspects such as generic characteristics of Italian water basins and data directly originated from APAT infrastructure of data collection. Its goal is to create a shared dictionary to answer to the need of having a unique point of access to data for the whole system. The Domain Ontology wraps a complex net of concepts which are defined through classes. Each class has one or more individuals, which represent concrete objects in the domain. For example, *Basin* is the class that defines the concept of hydrographical basin, while *Po* and *Adige* are two individuals that represent real Italian basins.

Relations between different classes are defined through properties; each of them has a domain and a range. For example, *Basin* has the property *contains\_water\_bodies* whose range is the class

*Water\_body*; the same property in the individual *Po* has a list of values containing individuals like *Borbore* and *Cervo*.

Figure 4 reports a representation of the most considerable relationships between ontology classes. In particular, the classes *Basin* and *Point\_of\_survey* can provide information concerning:

- Recent measurements, data collected in recent sessions of measurement. Every measurement is related to a specific tool of survey from which it is possible to know the meaning of the measurement.
- Historical measurements, data collected in the past, which concern information about rainfalls.

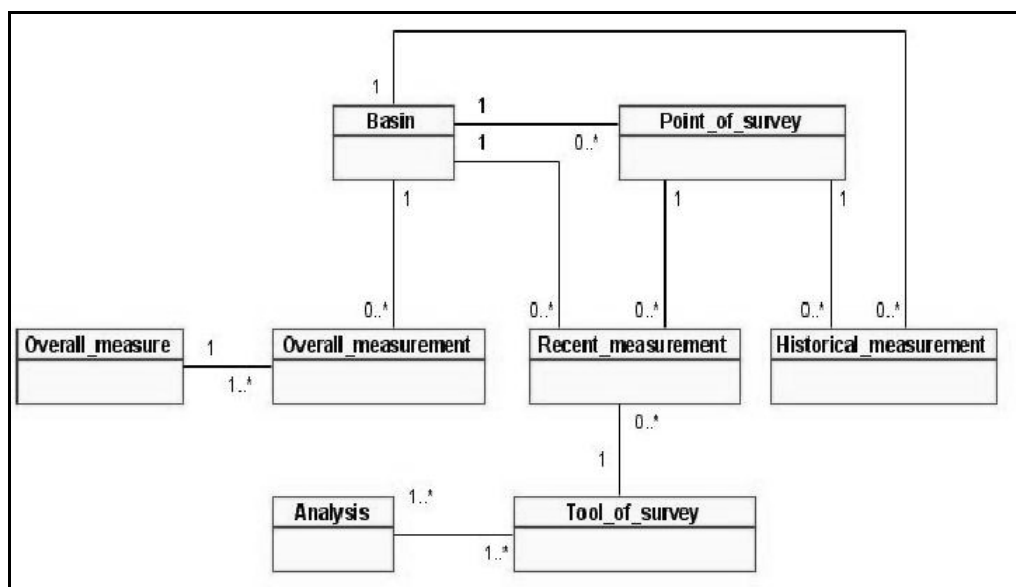


Figure 4. Representation of the most relevant ontology classes and relationships.

These kinds of data are tied to temporal information, since they are continuously updated day after day due to data flowing between hundreds of points of survey and a central repository; thus, they require a complex and flexible management of individuals. Once the Domain Ontology was built, it was important to understand how to populate it. The number of individuals depends on time frequency of observations, thus it is not advisable to instantiate these individuals in the same model containing the Domain Ontology, since this would require too much memory and too much time to load. For example, the class *Basin* has the property *overall\_measure*, and the Domain Ontology contains about 750 basins. Each individual in the class *Basin* has at least five dimensional measurements. Then it is not very maintainable to embed these data in the code of the Domain Ontology. Another relevant example concerns the thousands of measurements that every day the APAT points of survey bear; also in that case, data have not been embedded in the definition source of the ontology.

According to these motivations, the instantiation of individuals can happen in two ways:

Directly in the source code of the ontology; in this way individuals, classes and properties of the ontology are loaded at the same time;

Through volatile objects, which are instantiated at run time and only on a direct request of the end user. These individuals represent data extracted from the Federated System.

### *3.4 Web Ontology Browser*

As shown in the whole architecture (Figure 1), the user interface consists of two components: the Hydrological Query Wizard and the Web Ontology Browser (WOB). The Web Ontology Browser is a tool to navigate an ontology; it bears to allow the reading of an ontology through a web interface. Its main characteristics are:

- Tree visualization of classes
- Class details with visualization of properties and individuals
- Properties details with visualization of domain and range
- Individual details with visualization of values of its properties

An important function is the assignment of some parameters that are necessary to get values of ontology individuals for specific classes or properties (this is the case where individuals are retrieved, at run time, from the Federated System).

WOB is based on MVC pattern<sup>3</sup>, the framework selected for the implementation is Jakarta Struts<sup>4</sup>. Differently from traditional web applications, WOB does not implement a fully thin-client approach; Ajax is used for some tasks [Paulson, 2005]. The user interface (Figure 5) is composed by five main parts: the view of ontology classes which is navigable in tree-like way, and it is possible to zoom on the class details, the property details, the individual details.

---

<sup>3</sup> Sun Microsystems, Core J2EE Patterns: Patterns index page ,  
<<http://java.sun.com/blueprints/corej2eepatterns/Patterns/index.html>>.

<sup>4</sup> Struts: Jakarta Struts web framework, <<http://struts.apache.org/>>.

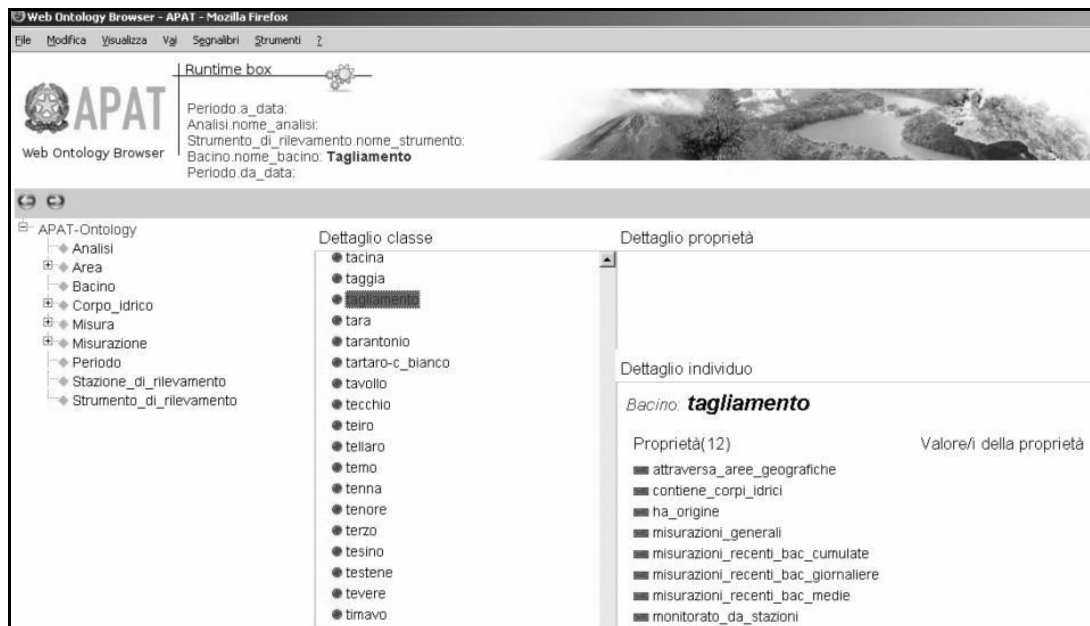


Figure 5. WOB user interface (Italian version)

#### **4. Conclusion**

Environmental Information Systems use a variety of tools and technologies to facilitate the interpretation of environment-related information. This paper reports about an experience in developing an information broker which provides an integrated and user-transparent access to all the heterogeneous data sources. Our information broker integrates a federated database system in order to achieve the syntactic interoperability, and it is augmented with domain ontology to share domain knowledge and to reach the semantic interoperability.

In the future, our plan is to endow the information broker with machine learning techniques to induce profiles of the users. User profiles will be exploited to enhance the current hydrological wizard with a personalized interface that helps the user during the process of query formulation.

#### **References**

- Calefato, F., Colagrossi, A., Gendarmi, D., Lanubile, F., Semeraro, G. 2006. An Information Broker For Integrating Heterogeneous Hydrologic Data Sources: A Web Services Approach. In: Tjoa, A.M., Xu, L., Chaudhry, S. (Eds.), *Research and Practical Issues of Enterprise Information System*, IFIP Series (Springer), Vol. 205, ISBN 9780387343457, pp 41-50.
- Colagrossi, A. 2003. Technologies For Storing, Processing and Fruition of Hydrological Data. Proc. of the Second International Conference on River Basin Management.
- Colagrossi, A., Gasparri, P.M. 2003. New Trends in Hydrology. The Role Played by Innovative Technologies. Proc. of the Fourth International Conference on Ecosystems and Sustainable Development.
- Kajan, E., Stoimenov, L. 2005. Toward an ontology-driven architectural framework for B2B. *Communications of the ACM* 48, 12, pp. 60-66.
- Noy N. F., McGuinness D. L. 2001. *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05.
- Park, J., Ram, S. 2004. Information systems interoperability: What lies beneath? *ACM Trans. on Information Systems* 22, 4, pp. 595–632.
- Paulson, L.D. 2005. Building Rich Web Applications with Ajax. *IEEE Computer* 38, 10, pp. 14-17.
- Sheth, A.P., Larson, J.A. 1990. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys* 22, 3, pp. 183-236.
- Uschold, M., Grüninger, M. 2004. Ontologies and semantics for seamless connectivity. *SIGMOD Record*, 33, 4, pp. 58-64.