

ONTOFUSION: Ontology-based integration of genomic and clinical databases

D. Pérez-Rey^{a,*}, V. Maojo^a, M. García-Remesal^a, R. Alonso-Calvo^a, H. Billhardt^b,
F. Martín-Sánchez^c, A. Sousa^d

^a*Biomedical Informatics Group, Artificial Intelligence Laboratory, School of Computer Science, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Spain*

^b*Artificial Intelligence Group, Universidad Rey Juan Carlos, Madrid, Spain*

^c*Bioinformatics Medical Group, Institute of Health Carlos III, Majadahonda, Madrid, Spain*

^d*IEETA, University of Aveiro, Aveiro, Portugal*

Received 3 June 2004; accepted 7 February 2005

Abstract

ONTOFUSION is an ontology-based system designed for biomedical database integration. It is based on two processes: mapping and unification. Mapping is a semi-automated process that uses ontologies to link a database schema with a conceptual framework—named virtual schema. There are three methodologies for creating virtual schemas, according to the origin of the domain ontology used: (1) top-down—e.g. using an existing ontology, such as the UMLS or Gene Ontology—, (2) bottom-up—building a new domain ontology— and (3) a hybrid combination. Unification is an automated process for integrating ontologies and hence the database to which they are linked. Using these methods, we employed ONTOFUSION to integrate a large number of public genomic and clinical databases, as well as biomedical ontologies.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: Ontologies; Biomedical informatics; Biomedical ontologies; Database integration; Ontology unification

* Corresponding author. Tel.: +34 91 336 74 67; fax: +34 91 352 48 19.

E-mail address: dperez@infomed.dia.fi.upm.es (D. Pérez-Rey).

1. Introduction

New technologies are being created to facilitate information search, access, retrieval and gathering from remote sources over the World Wide Web. In this scenario, developers are looking forward to the Semantic Web and related technologies that should facilitate information-related tasks in many areas. One such area is biomedicine, where collaborative efforts over the Web have led to significant scientific advances and accelerated efforts such as the Human Genome Project, among others. In this regard, research carried out during the last few decades has led to controlled vocabularies and taxonomies such as the UMLS [1], Gene Ontology [2], and others.

Ontologies provide the basis for the Semantic Web. Historically, the concept of ontology has a philosophical meaning, related to metaphysics. In informatics, ontologies provide a conceptual framework for modeling a knowledge domain. Considering medicine and biology, ontologies can contribute to bridging the gap between both fields by providing new conceptual frameworks. For instance, in the area of heterogeneous database integration, ontologies will provide the platform for sharing common vocabularies by modeling scientific domains. This exchange should prove fundamental in issues such as genomic medicine, where genomic and medical information will be jointly collected and analyzed to create new models of health care.

Biological and medical databases have traditionally been separate. Recent developments, such as the Iceland database [3], the biobanks and a clinical/genomic database under construction at the Mayo Clinic in Rochester, USA [4], are being carried out to gather biological and medical information. In this sense, ontologies can be particularly helpful for providing integrated approaches to data collection and analysis.

In this paper, we describe a project carried out over the last few years with support from the European Commission. This project, called INFOGENMED, aimed to develop various methods and tools for database integration from remote sources, based on intelligent agents and ontologies. The focus of this paper is related to the components of the system that are directly linked to ontologies. The system has been implemented and evaluated with biological and medical information. However, given its domain independent features, the ONTOFUSION system can be also used in other application domains.

The paper is organized as follows. Section 2 gives background on existing database integration methods and ontologies, especially from the biomedical point of view. In Section 3, we present the ONTOFUSION approach to database integration. Section 4 describes the evaluation of the system and Section 5 provides some discussion. Finally, Section 6 gives some conclusions and directions for further research.

2. Background

Biomedical institutions are producing an increasing amount of data. Given this scenario, professionals are demanding new models and tools to search, store and analyze information. Since the development of the World Wide Web, collaborative efforts among remote institutions and researchers have increased the need for information exchange and distributed data processing. We provide below a description of recent research on database integration. Since the latest efforts on database integration include the use of ontologies, we also provide a brief description of research related to ontologies, particularly in the area of biomedicine.

2.1. Database integration

Database integration is an area of informatics related to information exchange and gathering information from disparate and heterogeneous sources. The location of the information to be gathered is one of the first issues to be considered. From the viewpoint of data location, we can consider two different approaches to database integration: distributed vs. centralized approaches. In distributed approaches, databases are located at different and remote places and are connected over computer networks (e.g., the Internet). In centralized approaches, data are produced and accessed from various locations and gathered at centralized sites, usually called “data warehouses”. These systems have special features as compared with traditional relational databases [5].

While distributed architectures have several disadvantages, such as cost, performance, security, availability or reliability, they provide local autonomy to the independent organizations that share their information. By contrast, data warehousing is usually preferred at large enterprises and unified institutions, where homogeneity, rapid access to data and on-line analytical processing is a key issue for success. Regarding the biomedical area, the distributed approach has been preferred in areas such as genomic research, where independent institutions have collaborated and exchanged information, while keeping control of the data they generate and store [6]. This approach faces several challenges, since the data sets used by computational biologists are highly heterogeneous in structure, content, and semantics. They include data stored in flat files, relational databases, object-oriented databases, etc. This heterogeneity has led to the development of different informatics approaches to achieve database integration. The most important techniques are summarized in Table 1. Database integration requires bridging the syntactic and semantic gaps existing across data sources, a problem for which ontologies are particularly suitable [21]. According to this framework, database integration is currently evolving towards ontology-based approaches. Examples in the biomedical area include projects such as SEMEDA [22], KAON Reverse [23], or D2R MAP [24]. In these systems, developers use ontology-based views to facilitate the mapping from objects belonging to a specific database to concepts of a shared vocabulary. If two different databases contain the same concept, but this concept is represented by different names, ontologies are used to map these names to the same concept descriptor. While these novel, ontology-based approaches are highly promising in the field of database integration, there are still significant issues that need further investigation. Given their potential to facilitate the integration of heterogeneous data sources, the next subsection presents an overview of ontologies from an informatics perspective.

2.2. Ontologies

In general, ontologies are a conceptual framework of the structured representation of a given domain or meaning [25]. Gruber proposed the most widely accepted definition, considering ontologies as “an explicit specification of a conceptualization” [26]. Ontologies can be used either by humans or computerized systems. Usually, they include elements such as concepts, relationships between concepts, definitions for these concepts and properties. They may also provide the capability to generate logical inferences by defining rules and axioms. These mechanisms also control the objects that can be included within an ontology. Thus, ontologies are much more than simple controlled vocabularies or taxonomies, they have been conceived to represent the underlying meaning of a scientific domain.

Most of the recent ontology languages are based on XML, the resource description framework (RDF), and RDF schema (RDFS) [25]. Although other languages have been proposed, there are no universal

Table 1
Database integration approaches

Method	Main features	Examples	Drawbacks
Information linkage	Related pieces of information (web pages, records of DBs, etc.) are linked together using static links that can be followed by users	Hypertext based systems: e.g. World Wide Web Web-based biomedical DBs: e.g. MEDLINE, PDB, Prosite, etc.	Many of the links are uni-directional Queries are limited or pre-defined by the static links
Data translation	Data from different databases are converted into a common format Transformed data are stored in a centralized repository Users can retrieve data from different sources by accessing the centralized repository	Data warehouses	When data from the source databases are modified, the centralized repository has to be updated The stored data has to be checked for redundancy
Query translation Pure mediation	Mediators and wrappers are used to execute user queries. A mediator is a system capable of filtering the information from one or more sources [7]. It contains all the necessary knowledge to retrieve the required information and to present it to the users	TSIMMIS [8], DISCO, DIOM, HERMES, BioKleisli, BioDataServer	This approach is less intuitive for users than other approaches that use ontology schemas
Single conceptual schema	A single global conceptualization is provided containing all the information from the system data sources User queries are built based on this global conceptual schema Results are returned at the conceptual level—i.e., results are instances of entities and relationships belonging to the global conceptual schema	SIMS [9], ARIADNE [10], Pegasus [11], Garlic [12], Discovery-Link [13], TAMBIS [14], BACIIS [15]	Any change in the system (e.g., addition, removal, or modification of databases) may require a modification of the global domain conceptualization This approach is only applicable if all data sources store data from a common application domain
Multiple conceptual schemas	Different virtual schemas are used to describe the semantics of each database integrated into the system It solves some of the problems of the single schema approach, since it supports addition, removal or modification of data sources. These operations require the modification of the affected schemas only User queries may be expressed by using terms from specific domain ontologies	OBSERVER [16]	It cannot be assumed that individual schemas share the same vocabulary It is hard to find semantically similar or equivalent concepts in two or more conceptual schemas Mappings between semantically similar entities in different virtual schemas have to be established to solve this problem

Table 1 (continued)

Method	Main features	Examples	Drawbacks
Hybrid approach	Separate conceptual schemas (ontologies) are used to describe the semantics of each data source These schemas have to be developed using a common domain ontology	PICSEL [17], COIN [18], MECOTA [19], BUSTER [20], SEMEDA [21]	An existing domain ontology is required

standards yet. In the short history of ontology languages, there are numerous examples, such as Ontolingua, LOOM, OCML, FLogic, CARIN, SHOE, XOL or OML [27] among others. In general, these languages are very similar. Usually, existing services and applications can be updated when new ontology languages appear. Such updates have occurred when languages such as DAML, OIL, DAML+OIL and the last W3C recommendation, the Ontology Web Language (OWL), were released [28].

Ontologies have been recently introduced in the domain of genetics. In 1998, a consortium began to develop an ontology in the area of genetics, named “gene ontology” (GO) [2]. The GO project is a collaborative effort to create a controlled vocabulary of gene and protein roles in cells, addressing the need for consistent descriptions of gene products in different databases. GO developers have published three different ontologies, describing gene products in terms of their associated biological processes, cellular components, and molecular functions.

In the field of genomics there is a lack of naming conventions—e.g. like suffixes in chemistry—leading to naming duplications and confusions. To overcome inconsistencies, information from different research groups should be standardized—e.g. by using common nomenclatures. For instance, for each known human gene, Human Genome Nomenclature Committee (HGNC) approves a gene name and symbol (short-form abbreviation). Although HGNC is a controlled vocabulary rather than an ontology and does not provide relationships or a hierarchy, it is frequently used to avoid naming inconsistencies. Each symbol is unique and it is ensured that each gene is only assigned to one approved gene symbol. HGNC has already approved over 13,000 symbols for genes, approximately a third of the estimated total amount of 30,000 human genes.

In medicine, substantial contributions have been carried out for the development of standards, medical terminologies and coding systems. HL7, SNOMED, MeSH, and the Unified Medical Language System (UMLS, which integrates most of them) [1] have been widely adopted. The UMLS, launched by an initiative from the US National Library of Medicine (NLM), has been designed to facilitate information retrieval and integration from multiple, machine-readable biomedical information sources. The UMLS gathers more than 100 of the most relevant vocabulary sources in medicine. The April 2004 release of the Metathesaurus, for example, includes more than 1 million concepts and nearly 2.4 million unique concept names.

Until recently, few efforts have been made to develop an ontology integrating both genomic and clinical terminologies. One such effort was made by the US NLM. This initiative aimed to include GO as an additional source in the UMLS’ Metathesaurus since the 2003AB release. Another effort is being carried out by genomics organizations, such as the Microarray Gene Expression Data Society (MGED). The MGED has launched an ambitious project to build an ontology for genomic data obtained from microarrays and clinical data.

Some reports and a recent Workshop on Ontologies in Medicine held in Rome have stated the current problems of ontologies [29,30]. According to these sources, ontologies should be developed using more strict and formal approaches and methodologies. Whereas these calls for more research on developing more formal ontologies might provide more consistent approaches to current examples of ontologies, the ONTOFUSION approach has been designed to easily update these desirable improvements within biomedical ontologies. Therefore, once a new ontology is available, the process for updating ONTOFUSION is straightforward.

3. The ONTOFUSION approach to database integration

Database integration at a semantic level is a key issue for providing homogeneous access to clinical and genetic databases. The integration approach used in ONTOFUSION is based on two processes: mapping and unification. In the mapping process, the physical schema of each database is mapped to what we call a “virtual schema”. Virtual schemas are ontologies representing the structure of the information contained in a given database at a conceptual level. In the unification process, several virtual schemas corresponding to distinct databases are merged into a unified virtual schema. Unified virtual schemas are ontologies that reflect the conceptual structure of the information stored in various databases. Thus, they act as descriptions of virtual databases, matching “real” data from physical databases.

Fig. 1 illustrates the ONTOFUSION processes. The first row from the bottom shows the three different types of physical databases that are used in ONTOFUSION (private and public databases and biomedical ontology databases). The next row shows the mapping process linking these physical databases to virtual schemas, considered in our approach as ontologies. At the top, the unification process integrates several virtual schemas into new unified virtual schemas, which can be accessed by users in order to retrieve

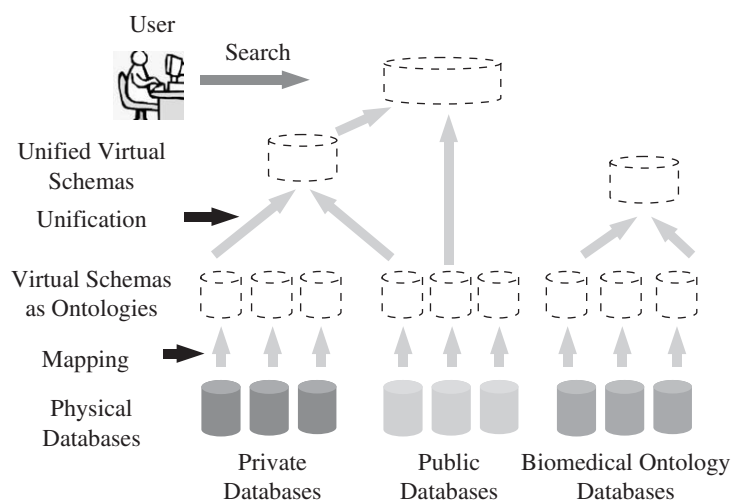


Fig. 1. ONTOFUSION mapping and unification process.

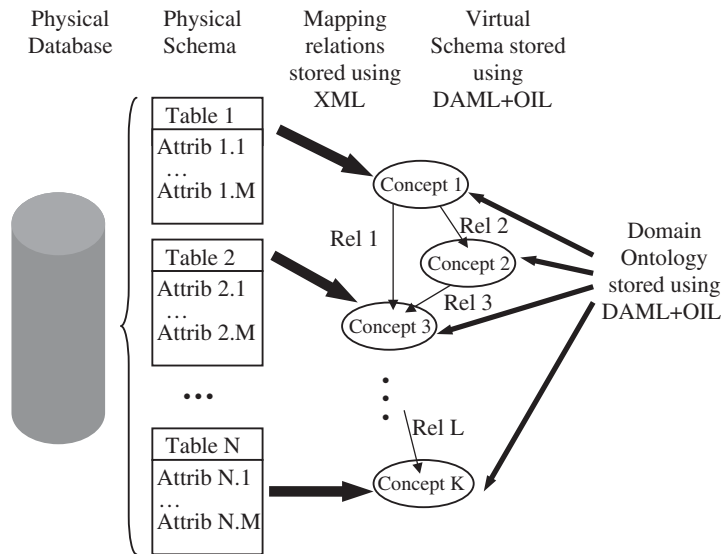


Fig. 2. ONTOFUSION virtual schema construction.

information from various sources at the same time. Unification can be applied to different sets of virtual schemas. In this way, it is possible to create several unified virtual schemas, each of which covers special data types.

3.1. Building virtual schemas

The first process in our database integration approach is to create virtual schemas. Virtual schemas provide new views, defined by users, of physical databases. If necessary, they can also be used to modify the actual schema of a physical database. In addition, virtual schemas provide methods to transform user requests into physical database queries (using specific database query languages).

Since database schemas do not have a conceptual environment, users cannot interact with the meaning underlying the stored data. Mapping a new database into the system is a semi-automated process, needed to link a physical database schema with an existing conceptual framework. In ONTOFUSION, specific domain ontologies are used to provide such a framework. They guide the administrator in the process of creating virtual schemas. Domain ontologies act as standardized vocabularies and contain conceptual hierarchies, concepts (classes), attributes and relationships, and the dependencies among them. The purpose of a domain ontology is to guarantee that only generally accepted concept names are used to describe the different elements in virtual schemas. Ontologies ensure that all semantically equivalent elements in different schemas use the same concept name.

When a new virtual schema is being developed, concept names cannot be chosen arbitrarily. They should be selected from the respective domain ontology. In this sense, domain ontologies provide the semantic links between different schemas, which are needed to carry out the schema unification process later.

As represented in Fig. 2, virtual schema elements are obtained from the corresponding domain ontology. Tables and attributes of physical database schemas are mapped to virtual schema elements, storing these relations using XML.

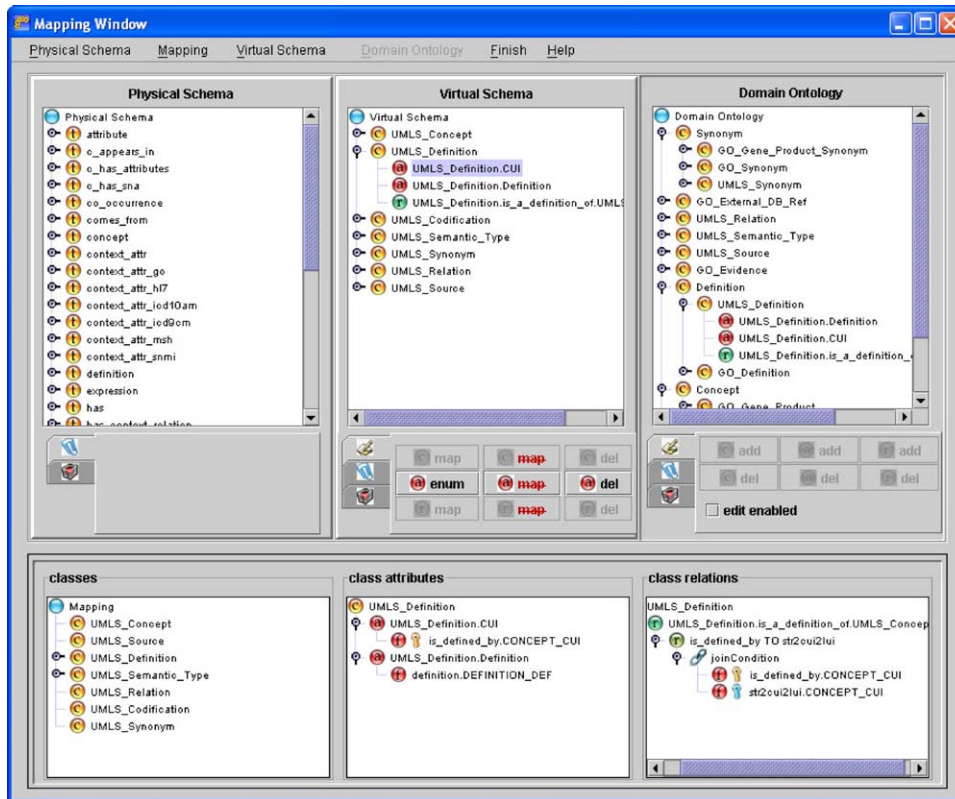


Fig. 3. ONTOFUSION mapping tool.

We have developed a mapping tool to facilitate the creation of virtual schemas. Fig. 3 shows a screenshot of this computer program:

As shown in Fig. 3, the mapping tool is divided into six sections. The top three frames, representing, from left to right, the physical schema, the virtual schema, and the domain ontology, can be modified by users. The three frames at the bottom show the mapping information relating the elements in the virtual schema to elements in the physical database schema.

The mapping process is carried out as follows. First, an existing domain ontology is loaded or a new one is created. Then the database's physical schema is obtained. Afterwards, the virtual schema is built by selecting classes, relationships and properties from the domain ontology (using a drag and drop mechanism). At the end, every element in the virtual schema is mapped to a respective element in the physical schema (again using a drag and drop mechanism).

The mapping tool and the domain ontologies provide guidance on the creation of virtual schemas. The process itself is carried out by an administrator. This person is responsible for identifying the important conceptual elements within a database, assigning them to appropriate concept names, and identifying the existing relationships among them.

There are different methods for building virtual schemas depending on the type of source database. In our system, we consider three different types of databases: private, public, and databases containing

biomedical ontologies. Private databases are privately maintained and have known and accessible physical schemas. In ONTOFUSION, their physical schema is automatically obtained by accessing the database. Public databases are usually accessed through the Internet using web-based forms. For instance, genomic data is often stored in public databases, like SwissProt, OMIM, Prosite, and others, and are publicly available over the WWW. ONTOFUSION can map these databases to ontologies as well, allowing users to query them from the same interface. Public database physical schemas have to be obtained by analyzing the web forms that provide access to these databases. The physical schemas have to be stored as XML files, which can then be loaded into the mapping tool. The last type of databases that we consider in our system are private databases storing biomedical ontologies or vocabularies. As discussed earlier in Section 2.2, biomedical ontologies are valuable vocabulary sources usually with a huge number of terms. Storing such biomedical sources—e.g. the UMLS—using Ontology Description Languages such as RDF, DAML+OIL or OWL could be a difficult task. The high number of elements would involve unmanageably large files. Consequently, storing biomedical ontologies as databases is currently a more efficient approach.

Regarding the domain ontology we use, three different mapping methodologies are possible with ONTOFUSION: top-down, bottom-up and hybrid. The top-down methodology uses an already existing domain ontology, such as UMLS, GO or HGNC. The newly created virtual schemas will be subsets of these vocabularies. The bottom-up methodology creates a new domain ontology from scratch. In this case, the mapping process begins by building a new domain ontology to model the concepts underlying a database schema. This domain ontology is extended as new concepts appear in subsequent mapping processes for other databases. Finally, the hybrid methodology combines both approaches. Initially an existing domain ontology is created with terms from standard vocabularies and is extended over time as required.

3.2. Virtual schema unification

To integrate databases, ONTOFUSION can unify virtual schemas that have been created using a common domain ontology. This unification process is completely automatic. To identify which concepts must be unified, the unification algorithm checks the domain ontology. When two or more terms match the same concept in the domain ontology, they are unified, including the associated attributes. The unification algorithm is described elsewhere [31]. Concepts stored in public databases can also be unified with concepts stored in private databases, providing standardized access to these different types of sources. Within the unification tool, the administrator selects a set of virtual schemas from what are available in the virtual schema pool, and the common domain ontology used to develop them. Based on this information, the unification tool generates the new unified virtual schema.

3.3. User interface

In our system, physical and virtual databases are represented by ontologies. This approach facilitates the creation of a user interface, which is dynamically modified to adjust to the databases that are accessible at a given time. The ONTOFUSION interface is an ontology navigator. Using this program, users can browse the ontologies that represent the physical or virtual databases that are connected to the system.

First, users navigate through the set of virtual schemas by means of the ONTOFUSION virtual database navigator. Based on a textual description of the different databases, they select the database that they want to browse. Once a user has selected a virtual schema, he/she can navigate through the concepts of the respective ontology using the virtual schema navigator. The user can select a concept and specify a query to retrieve instances of that concept. For this, he/she fills out a form to specify the filter criteria for one or more of the concept's attributes. For integers, users can choose from 'greater than' (>), 'less than' (<), 'equal to' (=), 'greater than or equal to' (>=), 'less than or equal to' (<=) and 'different' (!=) comparison operations. For strings, they can choose between 'logical and' (AND) and 'logical or' (OR). Furthermore, users can specify whether or not string comparison should be based on exact match. Finally, for every attribute and relationship, users can specify whether or not they are interested in retrieving that information. Once the user has specified the requested information, the query is automatically composed, divided (if required) and translated to the corresponding database query language.

Since ontologies are used to represent virtual schemas, the information—stored in public, private or vocabulary databases—is retrieved by the system as ontology instances. Results must first be unified (if required) according to the respective unified virtual schema. Then, results are presented to the users as instances of the virtual schema ontology. This approach permits navigation through the different instances and their relationships with other instances. In addition, results can also be deployed as an ontology description language file, e.g. DAML+OIL file, if required.

4. System evaluation

The system has been successfully tested with twenty databases:

- Eight private databases containing biomedical information of various types and stored in database management systems, such as MySQL, PointBase, Access, and others.
- Nine public databases: Ensembl, SwissProt, OMIM, Prosite, SNP, PDB, ENZYME, LocusLink, and InterPRO.
- Three databases containing biomedical ontologies: UMLS, GO and HGNC.

Although ONTOFUSION is a research tool and needs additional refinement, results are promising. A large number of different DBMS can be added to the system, including all major DBMS from different software companies. We have created a driver server using which the system can be easily configured to integrate new DBMS.

We have executed several comprehensive sets of queries against the different mapped databases. From these performance tests, data results from virtual schemas were compared to the corresponding results from physical databases. Furthermore, we also tested unified versus separate virtual schemas. We present below a case study using UMLS, GO and HGNC. All three sources have been mapped to the respective virtual schemas, and these have then been unified into a common virtual schema.

Let us consider that a researcher wants to retrieve the information included in UMLS, GO and HGNC about the muscle protein “desmin”. If he/she accesses each database separately, it takes four queries. Actually, a query asking for “desmin” in the “Name” attribute is sent to the class “UMLS_Concept” in UMLS, to the classes “GO_Term” and “GO_Gene_Product” in GO, and the class “HGNC_Gene” in HGNC. The results are shown in [Fig. 4](#).

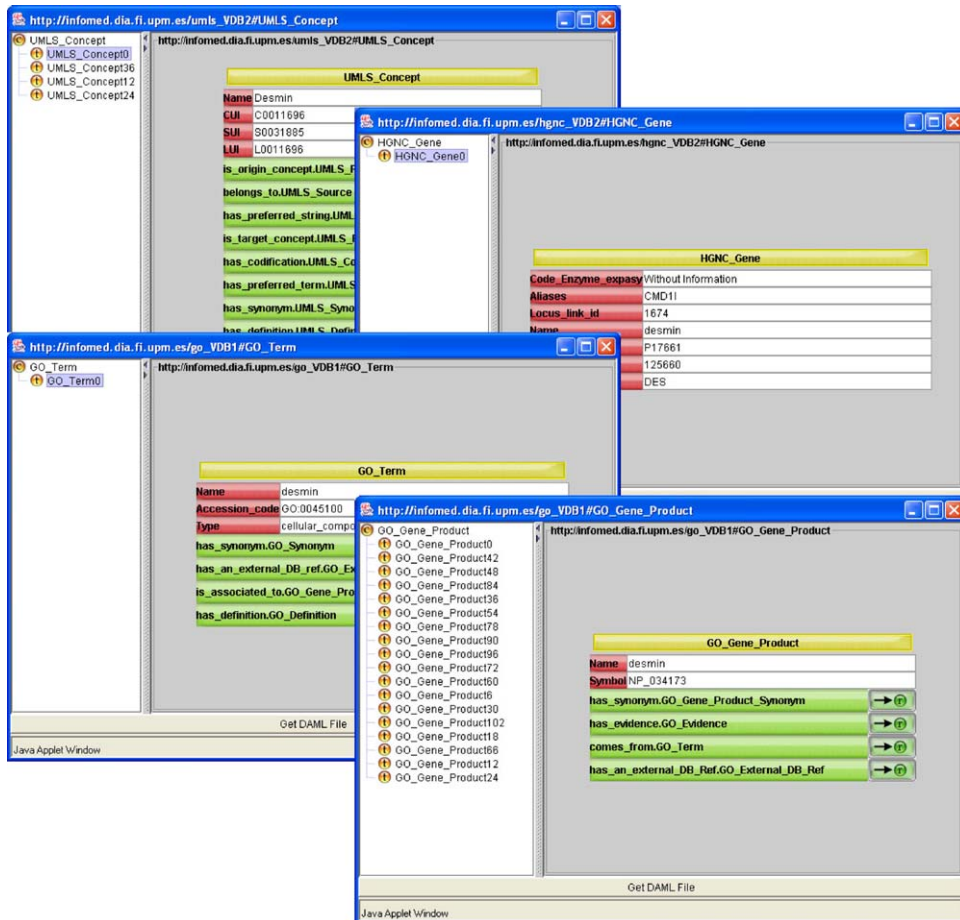


Fig. 4. Separate results for “desmin” query in UMLS, GO and HGNC.

The top left-hand window shows the four instances containing “desmin” in UMLS. The top right-hand window contains the instance retrieved from HGNC, and the bottom windows show the results retrieved from GO: 1 term and 18 gene products.

Fig. 5 shows the results obtained by submitting an equivalent query to the unified virtual schema that integrates UMLS, GO and HGNC.

The 24 instances shown in Fig. 5 are the same as in the case of the separate queries. They were retrieved with just one ONTOFUSION query. Besides the fact that the retrieved data instances were correct, they also corresponded to the results that a user would expect to retrieve when formulating the queries through the interface. That is, from the user point of view, there is an intuitive correspondence between the self-descriptive search forms and the retrieved data instances. This process proves the reliability of the proposed mapping and unification approach. However, databases from very different domains should not be unified. Such unifications would produce unified virtual schemas that are overly generic, and the

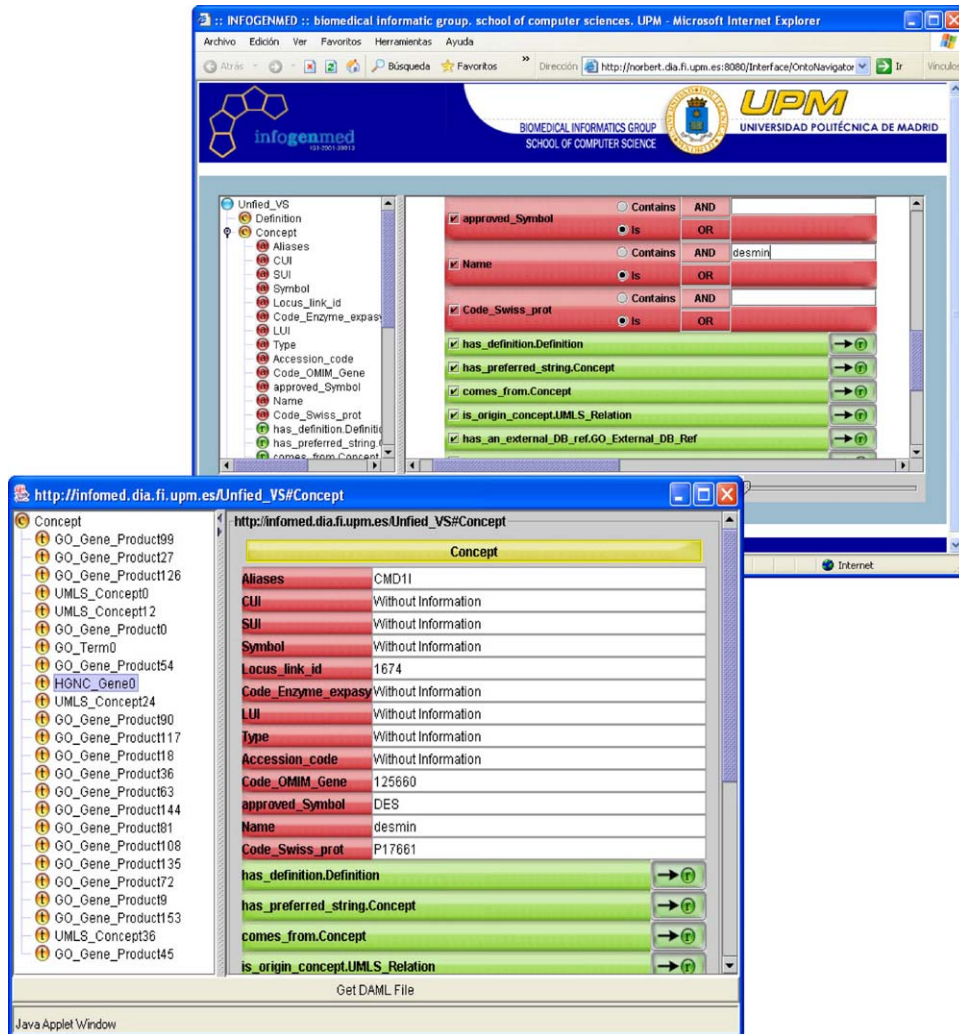


Fig. 5. Unified query and results for “desmin” in UMLS, GO and HGNC.

intuitive correspondence between the virtual schema that describes the underlying data and the actual data may be lost.

5. Discussion

ONTOFUSION has been implemented using a multiagent architecture. User agents play the role of users in the system and virtual schema agents act as wrappers of physical or virtual (unified) databases. Fig. 6 shows an example of the agent messages involved when a query is received within the system. Virtual agents are connected following the unification hierarchy of the databases. When a user submits a query to the system, it is translated, divided and transferred by the virtual schema agents until it reaches

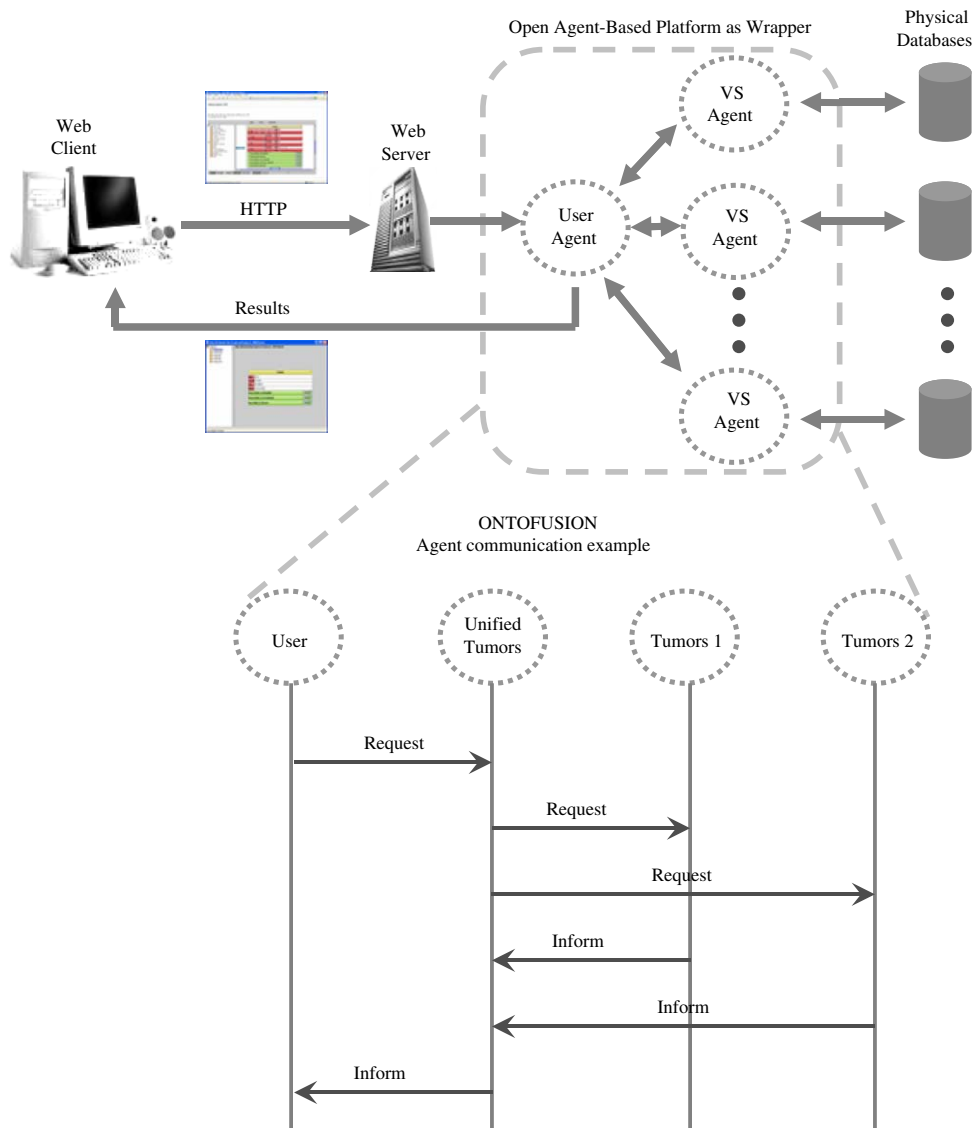


Fig. 6. Agent communication in ONTOFUSION.

the underlying physical databases. The retrieved results are sent back the same way and are represented through the user interface. This agent-based approach has the advantage that the whole system or part of it can be easily used for other tasks (e.g., epidemiological studies, data mining, etc.).

We believe that ontologies are the most suitable representation formalism for virtual schemas in our database integration system. Since virtual schema construction is based on the development of the model underlying a given domain—a database in this case—, we decided to use ontologies as the basis for our conceptual representation approach. In addition, the virtual schema construction is the process in which expert knowledge is more important and consequently where ontologies play their most important

Table 2

Comparison of ontology-based database integration approaches

	D2RMAP*	SEMEDA*	KAON reverse*	ONTOFUSION
Ontology description language	RDF	RDF	RDFS	DAML+OIL
OWL compatible	Yes	No	No	Yes
Ontology editor	No	Yes	Yes	Yes
Graphical ontology navigator	No	No	Yes	Yes
Public databases	No	Yes	No	Yes
Physical schema redesign	No	No	No	Yes
Virtual schema unification	No	No	No	Yes

*According to the latest information available at the time of writing this paper.

role within ONTOFUSION. Within our system, ontologies are stored using DAML+OIL, a high level ontology description language. Furthermore, we use XML to store the relationships between the elements of physical and virtual schemas—i.e. concepts or classes, attributes and relationships. ONTOFUSION is also compatible with RDF, RDF Schema and the new OWL. Our system can be easily adapted to a new ontology description language if needed.

The ONTOFUSION approach allows biomedical vocabulary sources to be managed and unified, enabling users to correlate concepts from different sources. Frequently, biomedical ontologies provided in the shape of databases do not follow a clear and standard structure. For instance, the database schema of the UMLS database release has several problems—such as redundancies—which must be taken into account. In this regard, the ONTOFUSION mapping tool can be used to normalize such databases to a common view, which could be more understandable for users. As previously stated, ONTOFUSION can be easily updated to include new and more consistent versions of UMLS, GO and other biomedical ontologies.

The ontology-based database integration approach of ONTOFUSION is more than just mapping elements from databases to ontologies. We have compared ONTOFUSION with other database integration systems that use an ontology-based approach. This comparison is shown in Table 2.

Table 2 shows the different features that characterize several ontology-based database integration systems. All of them have been developed for general-purpose database integration, not only for biomedical databases, although SEMEDA has been tested with biological data. In our case, ONTOFUSION has been tested primarily with biological and medical data. At the time of writing this paper, a demo is being built, using military databases, by a private Spanish company for a NATO exhibition. To our knowledge,

ONTOFUSION is the first database integration system that uses a high-level ontology description language to represent the virtual schemas. Like D2RMAP, ONTOFUSION is also compatible with the new OWL. Our system incorporates an ontology editor to build the virtual schemas (KAON and SEMEDA have a similar feature as well) and a graphical ontology navigator (as KAON does) to retrieve the results. Regarding public databases, ONTOFUSION shares with SEMEDA the capability to map and integrate these databases, but only ONTOFUSION can unify public and private databases.

Another main contribution of ONTOFUSION is its capability to redesign database schemas. Using the ONTOFUSION mapping tool, physical database schemas can be improved by removing structural differences. For instance, suppose we have two databases with clinical information from patients, and their genetic profiles, which have different schemas and must be unified. One of them has only one table with all the information (“Patient+Genetic Profile”), whereas the other has two tables (“Patient” and “Genetic Profile”). Both databases should have two tables. Using ONTOFUSION, these two different physical schemas can be mapped to a common virtual schema with two concepts. Unification is then automatic.

Regarding scalability, ONTOFUSION has been tested with a total of twenty databases, eight private databases, nine public databases and three databases containing biomedical ontologies. In theory, the only constraint on the number of databases that can be integrated into the system is related to computer processing capability. More tests—including some tests carried out with a larger number of databases—will be needed to ensure efficiency.

6. Conclusions

Modern database integration tools are moving towards ontology-based approaches. Our approach, ONTOFUSION, has followed this trend. Ontology-based systems offer the possibility of navigating through the ontology concepts and exploring their relationships. These approaches ease the understanding of these concepts and their underlying knowledge. This is especially important in fields such as biology and medicine, where the number of concepts is very large and new concepts are appearing all the time.

ONTOFUSION provides a flexible architecture. It is based on a two-level integration process. First, databases are mapped to virtual schemas that describe their information content at a conceptual level. Second, virtual schemas can be unified, creating new, virtual databases that cover multiple physical databases and provide integrated access to their data. Using this approach it is easy to integrate additional databases at any time.

A key concept in the integration process are domain-specific ontologies. They provide the conceptual framework to establish semantic links between the data stored in different databases. Existing domain ontologies, such as the UMLS, GO or other future improvements of existing or new biomedical ontologies, can be used in the mapping/unification processes. Domain ontologies can also be the result of an integration process. That is, the system can also be used to create domain ontologies.

Whereas the unification process is fully automated, human intervention is required in the mapping process. An administrator has to define the conceptual structure of the information stored in a given database. In the future we plan to study semi-automated methods that provide additional help with this task. In any case, we believe that the mapping process should be supervised by human experts, because it is an essential process for successful database integration.

Regarding biomedical ontologies, we have currently considered only the UMLS, GO and HGNC. Nevertheless, other vocabularies or ontologies can be easily added. Since virtual schemas developed in

the mapping process for biomedical ontologies are in fact metaontologies, ONTOFUSION can provide developers with methods to redesign and manage existing biomedical ontologies.

7. Summary

New technologies are being created to facilitate information search, access, retrieval and gathering from remote sources over the World Wide Web. In this paper, we describe ONTOFUSION, an approach to information integration that has been developed as part of a project carried out over the last few years with support from the European Commission. This project aimed to develop various methods and tools for integrating databases from heterogeneous sources, using intelligent agents and ontologies. In ONTOFUSION, specific domain ontologies are used to link a database's physical schema with an existing conceptual framework. At the moment, ONTOFUSION considers three types of data sources: private and public databases or databases containing biomedical ontologies.

The integration approach used in ONTOFUSION is based on two processes: mapping and unification. In the mapping process, physical database schemas are transformed into intuitive and conceptual ontologies, called virtual schemas. Names for virtual schema concepts are picked from a domain ontology. Domain ontologies contain conceptual hierarchies used to develop several virtual schemas, with general concepts, attributes and relations. The mapping method is semi-automated and is an aid for the mapping process, but the administrator is responsible for establishing relationships and correspondences between concepts. There are three different mapping methodologies within ONTOFUSION: top-down, bottom-up and hybrid. The top-down methodology uses an already existing domain ontology such as UMLS, GO, HGNC or other. The new virtual schemas representing databases will be subsets of these vocabularies. The bottom-up methodology creates a domain ontology from scratch. In this case, the ontology, which is initially empty, is enhanced during subsequent mappings of new databases whenever a concept appears that has not yet been incorporated. Finally, a hybrid methodology is used when a new domain ontology is created with terms from standard vocabularies and new concepts are incorporated as they appear.

The unification process takes a set of virtual schemas and creates a new unified virtual schema. This schema is a conceptual description of the information provided by a virtual database, a database that integrates the data from several physical databases. The process is fully automatic. It checks the domain ontology used to develop the virtual schemas to identify which concepts must be unified. When two or more classes from different original schemas match the same concept in the domain ontology, they are unified, including their respective attributes.

The ONTOFUSION interface is an ontology navigator for virtual schemas. For every concept containing attributes, users can fill in a form to filter the required information and submit the query to the system. Results retrieved from a database (virtual or physical) are presented as instances of the virtual schema. This approach permits navigation through the different instances and their relationships.

The system has been successfully tested on eight different private databases with biomedical data stored in different database management packages such as MySQL, PointBase, Access, and others. Furthermore, we included nine public databases—Ensembl, SwissProt, OMIM, Prosite, SNP, PDB, ENZYME, LocusLink and InterPRO, and three biomedical ontologies: UMLS, GO and HGNC.

Compared to modern ontology-based systems like SEMEDA, D2RMAP or KAON Reverse, ONTOFUSION includes original work on graphical ontology navigation, physical schema redesign and virtual schema unification.

Acknowledgements

This research has been supported by funding from the EC INFOGENMED project and the INFO-BIOMED Network of Excellence, the INBIOMED project, Ministry of Health, Spain, and the Ministry of Science and Technology, Spain.

References

- [1] C. Lindberg, The Unified Medical Language System (UMLS) of the National Library of Medicine, *J. Am. Med. Record Assoc.* 61 (5) (1990) 40–42.
- [2] The Gene Ontology Consortium, Gene ontology: tool for the unification of biology, *Nat. Genet.* 25 (2000) 25–29.
- [3] G.J. Annas, Rules for research on human genetic variation-lessons from Iceland, *New England J. Med.* 342 (2000) 1830–1833.
- [4] P.C. de Groen, A healthy database, IBM creating a system for millions of Mayo clinic patient files, in: *Renee Berg Post-Bulletin*, Rochester, MN, 2002.
- [5] R. Kimball, M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, Wiley, New York, 2002.
- [6] A.S. Lopatenko, Information retrieval in current research information systems, *Workshop on Knowledge Markup and Semantic Annotation at K-CAP'2001*, 2001.
- [7] G. Wiederhold, Mediators in the architecture of future information systems, *IEEE Comput.* 25 (3) (1992) 38–49.
- [8] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, J. Widom, The TSIMMIS project: integration of heterogeneous information sources, In: *Proceedings of IPSJ Conference*, Tokyo, Japan, 1994, pp. 7–18.
- [9] Y. Arens, C.N. Hsu, C.A. Knoblock, Query processing in the SIMS information mediator, in: M.N. Huns, M.P. Singh (Eds.), *Readings in Agents*, Morgan Kauffmann, San Francisco, CA, 1998.
- [10] C.A. Knoblock, S. Minton, J.L. Ambite, N. Ashish, I. Muslea, A.G. Philpot, S. Tejada, The Ariadne approach to Web-based information integration, *Int. J. Cooperative Inform. Syst.* 10 (1–2) (2001) 145–169.
- [11] M.C. Shan, R. Ahmed, J. Davis, W. Du, W. Kent, Pegasus: a heterogeneous information management system, in: W. Kim (Ed.), *Modern Database Systems*, ACM Press, New York, Addison-Wesley, Reading, MA, 1994 (Chapter 32).
- [12] M.J. Carey, L.M. Haas, P.M. Schwarz, M. Arya, W.F. Cody, R. Fagin, M. Flickner, A.W. Luniewski, W. Niblack, D. Petkovic, J. Thomas, J.H. Williams, E.L. Wimmers, Towards heterogeneous multimedia information systems, *Proceedings of the IEEE Fifth International Workshop on Research Issues in Data Engineering*, Taipei, Taiwan, 1995, p. 124.
- [13] L.M. Haas, P.M. Schwarz, P. Kodali, E. Kotlar, J.E. Rice, W.C. Swope, Discoverylink: a system for integrated access to life sciences data sources, *IBM Syst. J.* 40 (2) (2001) 489–511.
- [14] P.G. Baker, A. Brass, S. Bechhofer, C. Goble, N. Paton, R. Stevens, TAMBIS: transparent access to multiple bioinformatics information sources, *Bioinformatics* 16 (2) (2000) 184–186.
- [15] Z.B. Miled, N. Li, G. Kellett, B. Sipes, O. Bukhres, Complex life science multidatabase queries, in: *Proceedings of the IEEE*, 90, vol. 11, 2002, pp. 1754–1763.
- [16] E. Mena, A. Illarramendi, V. Kashyap, A.P. Sheth, OBSERVER: an approach for query processing in global information systems based on interoperability between pre-existing ontologies, *Distrib. Parallel Databases* 8 (2) (2000) 223–271.
- [17] F. Goasdoué, V. Lattes, M.C. Rousset, The use of CARIN language and algorithms for information integration: the PICSEL project, *Int. J. Cooperative Inform. Syst.* 9 (4) (2000) 383–401.
- [18] C.H. Goh, Representing and reasoning about semantic conflicts in heterogeneous information sources, Ph.D. Dissertation, Massachusetts Institute of Technology, 1997.
- [19] H. Wache, T. Scholz, H. Stieghahn, B. König-Ries, An integration method for the specification of rule-oriented mediators, in: Y. Kambayashi, H. Takakura (Eds.), *Proceedings of the International Symposium on Database Applications in Non-Traditional Environments (EFIS 99)*, Kühlungsborn, Germany, 1999.
- [20] H. Stuckenschmidt, F. van Harmelen, D. Fensel, M. Klein, I. Horrocks, Catalogue integration: a case study in ontology-based semantic translation, Technical Report IR-474, Computer Science Department, Vrije Universiteit Amsterdam, 2000.
- [21] A. Silvescu, J. Reinoso-Castillo, V. Honavar, Ontology-driven information extraction and knowledge acquisition from heterogeneous, distributed, autonomous biological data sources, *Proceedings of the IJCAI2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources*, 2001.

- [22] J. Köhler, S. Philippi, M. Lange, SEMEDA: ontology based semantic integration of biological databases, *Bioinformatics* 19 (18) (2003) 2420–2427.
- [23] <http://kaon.semanticweb.org/alphaworld/reverse/>
- [24] C. Bizer, D2R MAP—A database to RDF mapping language, in: *Proceedings of the International World Wide Web Conference (WWW2003)*, Budapest, Hungary, 2003.
- [25] T.R. Gruber, A translation approach to portable ontology specifications, *Knowledge Acquisition* 5 (2) (1993) 199–220.
- [26] J. Davies, D. Fensel, F. van Harmelen, *Towards the Semantic Web: Ontology-Driven Knowledge Management*, Wiley, London, 2003.
- [27] A. Gómez-Pérez, M. Fernández-López, O. Corcho, *Ontological Engineering*, Springer, London, 2003.
- [28] The Ontology Web Language. <http://www.w3.org/2001/sw/WebOnt/>
- [29] D.M. Pisanelli, A. Gangemi, If ontology is the solution, what is the problem?, in: D.M. Pisanelli (Ed.), *Ontologies in Medicine*, IOS Press, Amsterdam, 2004, pp. 1–19.
- [30] W. Ceusters, B. Smith, A. Kumar, C. Dhaen, Mistakes in medical ontologies: where do they come from and how can they be detected?, in: D.M. Pisanelli (Ed.), *Ontologies in Medicine*, IOS Press, Amsterdam, 2004, pp. 145–163.
- [31] H. Billhardt, J. Crespo, V. Maojo, F. Martín, J. L. Maté, A new method for unifying heterogeneous databases, in: *Proceedings of the ISMDA*, 2001, pp. 54–61.

David Pérez del Rey is a research assistant at the Biomedical Informatics Group at the Polytechnical University of Madrid (Spain). He received a B.S. in Computer Science from the Complutense University of Madrid, including a year in the University of Southampton as a visitor student. He is currently finishing his Ph.D. thesis on Ontology-based KDD process for biomedical information. His research interests include data integration, data mining, KDD and the Semantic Web. Contact him at the School of Computer Science, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid (Spain); dperez@infomed.dia.fi.upm.es

Victor Maojo got his MD degree at the University of Oviedo (Spain) in 1985 and his Ph.D. in Computer Science at the Universidad Politécnica de Madrid (UPM) in 1990. At the UPM, he is currently an associate professor and associate director of the Artificial Intelligence Lab. Before entering the faculty of the UPM, he was a postdoctoral researcher and consultant in Georgia Tech (Atlanta, USA, 1990–1991), and a research fellow at the Decision Systems Group (Harvard University-MIT, Boston, USA, 1991–1993). He has been the principal investigator in more than 20 national and international projects and has authored more than 100 scientific papers and books. He has been a member of numerous committees at international conferences and journals and served as an expert for the IV and V Framework Programmes of the European Commission.

Miguel García Remesal is a research assistant at the Biomedical Informatics Group at the Universidad Politécnica de Madrid (Spain). He received a B.S. in Computer Science from the Universidad Politécnica de Madrid. He is currently finishing his Ph.D. thesis on Ontology-based Information Retrieval for biomedical information resources. His research interests include information retrieval, text mining, and the Semantic Web. Contact him at the School of Computer Science, Polytechnical University of Madrid, 28660 Boadilla del Monte, Madrid (Spain); mgarcia@infomed.dia.fi.upm.es

Raúl Alonso Calvo is a research assistant at the Biomedical Informatics Group at the Universidad Politécnica de Madrid (Spain). He received a B.S. in Computer Science from the Universidad Politécnica de Madrid. He is currently finishing his Ph.D. thesis on Content-Based Image Retrieval and Ontology-based Information Retrieval for biomedical information resources. His research interests include image analysis, information retrieval, and mathematical morphology. Contact him at the School of Computer Science, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid (Spain); ralonso@infomed.dia.fi.upm.es

Holger Billhardt received his M.Sc. in Computer Science from the Technical University of Leipzig, Germany, in 1994. He has been working from 1997 to 2001 as a research fellow at the Medical Informatics Group at the Universidad Politécnica de Madrid, Spain, where he received his Ph.D. in Computer Science in 2003. Dr. Billhardt is currently an Associate Professor at the Department of Informatics, Statistics and Telematics at the University Rey Juan Carlos of Madrid. His research interests include information retrieval, the use of multiagent systems for information access and retrieval, and its applications in the field of biomedicine.

Fernando Martin Sanchez earned his bachelor's degree in Biochemistry and Molecular Biology in 1986 from the Autonomous University of Madrid and received a MSc in Knowledge Engineering in 1987 and a Ph.D. in Computer Science in 1990 from the Polytechnic University of Madrid. He was a postdoctoral fellow at the Emory University Hospital-Georgia Institute of Technology Joint Research Program in Biomedical Informatics. Dr. Martin-Sanchez serves currently as Head of the Medical Bioinformatics Department of the National Institute of Health "Carlos III" of Spain, where he currently leads a multidisciplinary research team focused on Biomedical Informatics and microarray applications in genomic medicine. He regularly teaches on these subjects in public health schools, universities and hospitals.

Antonio Sousa Pereira received the degree in Electrical Engineering from the University of Porto, and Ph.D. degree, in Electrical Engineering from the University of Aveiro, where he is currently full professor. He is Director of IEETA, a R&D Institute, and Coordinator of the Information Systems and Telematics Lab. His main research interests are in telematics in healthcare and biomedical informatics.