INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team gemo

Management of Data and Knowledge Distributed Over the Web

Futurs

THEME 3A

Activity Report

2003

Table of contents

1. Team

Gemo is a joint project with Laboratoire de Recherche en Informatique (UMR 8623 CNRS-University Paris-Sud), located in Orsay. The project started in January 2003 and is a descendant of the Verso project.

Managers

Serge Abiteboul [DR-INRIA]
Marie-Christine Rousset [Professor, Univ. Paris 11]

Administrative Assistant

Geneviève Grisvard [until September]
Stéphanie Meunier [since September]

INRIA personnel

Ioana Manolescu [CR-INRIA]
Luc Segoufin [CR-INRIA]

University personnel

Hélène Gagliardi [Assistant Professor, Univ. Paris 11]
François Goasdoué [Assistant Professor, Univ. Paris 11]
Ollivier Haemmerle [Assistant Professor, INA P-G]
Nathalie Pernelle [Assistant Professor, Univ. Paris 11]
Chantal Reynaud [Professor, Univ. Paris 10]
Brigitte Safar [Assistant Professor, Univ. Paris 11]
Véronique Ventos [Assistant Professor,Univ. Paris 11]

Scientific Advisors

Bernd Amann [Assistant Professor, CNAM]
Christine Froidevaux [Professor, Univ. Paris 11]

Invited researchers

Tova Milo [Professor, U. Tel Aviv]
Victor Vianu [Professor, U.C. San Diego, 4 months]

Engineers

Jérôme Baumgarten
Nicolaas Ruberg [Bank NDES of Brazil]

Ph.D students

Philippe Adjiman [Allocataire MENRT, Paris 11]
Omar Benjelloun [Allocataire MENRT, Paris 11]
Grégory Cobéna [X-Télécom]
Gloria-Lucia Giraldo [Paris 11]
Hassen Kefi [Allocataire MENRT, Paris 11]
Amar-Djalil Mezaour [Allocataire MENRT, Paris 11]
Benjamin Nguyen [Allocataire MENRT, Paris 11]
Gabriella Ruberg [Federal University of Rio de Janeiro]
Mathias Samuelides [ENS Cachan]
Alexandre Termier [Allocataire MENRT, Paris 11]

3. Scientific Foundations

Key words: *Databases*, *knowledge representation*, *data integration*, *semantic integration*, *query language*, *query optimization*, *distributed query*, *peer-to-peer (p2p)*, *semi-structured data*, *XML*, *World Wide Web*, *Web services*, *change control*, *logic*, *complexity*.

Information available online is more and more complex, distributed, heterogeneous, replicated, and changing. Web services, such as SOAP services, should also be viewed as information to be exploited.

The goal of this project is twofold: first of all to study the fundamental problems that are raised by modern information and knowledge management systems, and secondly to determine novel solutions to solve these problems. Such systems will contain rich information and must be connected to networks. Gemo's main theme is the integration of information, seen as a general concept; more precisely, to discover meaningful information or services, understand their content or goal, integrate them, and finally monitor their evolution over time.

We would like to offer environments that are both powerful and flexible to simplify the development and deployment of applications that give fast access to meaningful data. The creation of data warehouses and mediators offering a wide access to multiple heterogeneous sources provides a good means of achieving these goals. These new problems combine Artificial Intelligence techniques (such as classification) and Database techniques (such as indexing).

Gemo is a project born from the merging of INRIA-Rocquencourt project Verso, with members of the IASI group of LRI. It is located in Orsay-Saclay.

4. Application Domains

4.1. Introduction

Key words: *Web*, *telecommunications*, *electronic commerce*, *enterprise portal*, *search engine*, *data warehousing*, *multimedia*.

Databases do not have specific application fields. As a matter of fact, all applications that involves processing large amounts of data need to use databases. Technologies recently developed within our project focus on novel applications in the context of the Web, both telecom, multimedia, enterprise portals, or information systems.

Gemo has chosen to focus mainly on applications that involve the Web. As an example, we demonstrate the intelligent use of data from the Web in a warehouse approach.

4.2. A warehouse on food risk

Key words: *Internet*, *web*, *enterprise portal*, *search engine*, *data warehouse*, *food risk*.

The warehouse allows to acquire information from the Web about food risk, to enrich this information, e.g., by classifying and integrating it, and finally provide a unique entry point for it.

Our goal is to develop tools allowing us to build domain specific data warehouses, which automatically integrate information found on the Web with private information, and information provided by content providers. This work takes place within the RNTL project e.dot that started during 2003. The project is based on XML, and on new services, such as those provided by Xyleme, high level queries and Web monitoring. Experimentation will lead to the construction of a data warehouse on food risk.

This project is a cooperation between the INRA BIA group, the Xyleme start up, and Gemo. BIA was chosen by the Ministry of Agriculture and the Ministry of Research to be the center of computer related skills with regards to the national research program on food risk.

5. Software

- XQueC: a prototype for storing and querying compressed XML data [30].
- WebQueL : a multi-criteria filtering tool for Web documents, developed in the setting of the e.dot project.
- XyDiff: a *diff* tool for XML (freeware and open source)
- Thesus: a prototype of a Web warehouse of thematic documents, constructed using links semantics, in collaboration with the Athens University of Economics and Business.

- SPIN: a prototype of Web warehouse definition and construction, based on a declarative language, and implemented using Active XML
- Active XML: a language and system based on XML documents containing Web service calls.
- STYX: definition and construction of a generic platform to integrate and query relevant XML resources concerning a Web community.
- OntoClass and OntoQuery: two tools (patented by France Telecom R&D) for the automatic classification of concepts, and the rewriting of conjunctive queries into query plans, which was developed within the PICSEL project framework.
- TreeFinder: a prototype system that discovers frequent tree patterns within a collection of XML data.
- Zoom: a prototype to construct and refine a lattice of classes of semi structured documents, developed within the GAEL project framework.
- OntoMedia: a prototype for the automatic construction of ontology components, using DTDs, developed within the PICSEL2 project.

6. New Results

6.1. Theoretical foundations

Participants: Serge Abiteboul, Tova Milo, Luc Segoufin, Victor Vianu.

Key words: *Semi-structured data*, *query languages*, *automata*.

One of the reasons for the success of the relational data model was probably its clean theoretical foundations. On the mathematical side it simply consists of relations equipped with the first-order logic as a query mechanism. This is accompanied by the equivalent relational algebra which allows evaluation of queries and facilitates optimization issues. Last but not least, there is SQL as an easy-to-use query language which, thanks to its simplicity, evolved de facto as a standard.

Obtaining such a clean foundation for the semistructured data model and XML is still an open issue. We studied XML as well as Active XML, a novel extended variant of XML where part of the documents data is generated dynamically via embedded calls to Web services. We detail below our results, first for XML and then for Active XML.

XML Most of the current proposals are based on the tree structure of XML data and make use of the fundamental connection between Monadic-Second-order (MSO) logic and automata on trees. Most of our theoretical work follows this approach.

In [41] we study the precise complexity of testing whether an unranked tree is accepted or not by a tree automata. We deduce from it the precise complexity of checking whether an XML document conforms to a DTD or an XML-schema. We also study, in the same paper, the precise complexity of evaluating queries of various fragments of XPath (a W3C standard which is also a subset of MSO).

Active XML (AXML) extends XML by allowing documents where some of the data is given explicitly while other parts are defined only intensionally by means of embedded calls to Web services (see Section 6.4). Each such call is typed using XML schemas. When a user request some of the data, the system has to decide whether to materialize some of the intensional data or not and the order of the materialization may affect the final result. [32] addresses the problem of what to materialize and when and presents a tractable case. In [44] the general problem is shown to be non computable and an overview of many decidable cases is presented.

6.2. XML and Service Mediation

Participants: Hélène Gagliardi, Gloria Giraldo, Nathalie Pernelle, Chantal Reynaud, Marie-Christine Rousset, Michele Sebag, Alexandre Termier, Veronique Ventos.

Key words: *Semantic integration*, *ontologies*, *clustering*, *structure extraction*.

Mediation consists of providing an integrated and uniform view over multiple data and services that are possibly heterogeneous. We focus on data and services that are described in XML documents.

In the PICSEL2 project, we have studied how to semi-automatically construct a class-based ontology from a set of DTDs related to a given application domain. A first prototype, called OntoMedia, has been developed for extracting ontology components from DTDs. In the continuation of this work, we have developed an approach for generating automatically service descriptions relatively to the resulting ontology [37]. The approach is based on XML standardized and industry-wide specifications of e-business messages (e.g., those defined by Open Travel Alliance for travel industry : www.opentravel.org).

We have designed and implemented a suite of algorithms ($TreeFinder\ TreeFinderCC$) for extracting frequent tree patterns in large collections of heterogeneous XML documents. This provides tree mediated schema over the semistructured data represented in the XML documents. Our method is robust for handling variability in the input data and easily parameterizable for controlling efficiency and scalability.

ZooM [35] is a two-step clustering tool which is formally based on Galois lattice techniques. The first step provides a coarse clustering. The second step refines a small subpart of the coarse lattice delimited by two nodes chosen by the user. We have improved ZooM by implementing an extensional refinement enabling it to deal either with noise in the instances or with heterogeneous user-defined basic types.

In the setting of the e.dot project, mediation is guided by a pre-defined ontology for discovering relevant HTML documents on the Web, extracting from them tables of useful data and transforming them into XML documents in which terms of the ontology are used as semantic tags.

A preliminary work has been done on automatic structure extraction from raw texts in order to transform unstructured fragments of text into XML documents. This work has been done in collaboration with LIMSI, in the setting of a BQR project supported by Paris-Sud University.

6.3. Mediation for the Semantic Web and Peer to Peer Systems

Participants: Philippe Adjiman, Bernd Amann, Francois Goasdoué, Hassen Kefi, Chantal Reynaud, Marie-Christine Rousset, Brigitte Safar.

Key words: *Semantic web*, *ontologies*, *semantic mapping*, *peer to peer*, *cooperative query answering*.

The Semantic Web envisions a world-wide distributed architecture where data or computational resources will easily interoperate thanks to a semantic marking up of Web resources using ontologies (e.g., taxonomies of terms shared by some communities of users or practitioners). In this vision of the Semantic Web based on simple but distributed ontologies, the key point is the mediation between data, services and users, using mappings between ontologies. We have obtained results in several complementary directions, detailed below.

First, we have investigated the problems raised by mediation between a network of semantically related peers. In [29] we have defined a simple framework appropriate for practical applications, which is based on a class-based language for defining peer schemas as hierarchies of (possibly disjoint) atomic classes, and semantic mappings between them as inclusions of logical combinations of atomic classes. On this subject, we are starting a collaboration with the SoMeOne project [36] of France Telecom R&D.

As a complementary direction, in the context of the Active XML project, we have designed and implemented a Web service directory based on a semantic description model using taxonomies for describing and querying services and data. The prototype has been implemented by Radu Pop during his internship. This work is a starting point of a new research activity around Web services and the Semantic Web.

In another direction, in the e.dot project, we are investigating how to map two distinct ontologies related to the same domain, each one having been designed independently of the other. The goal, in this work, is to study the automation of the mapping process.

Finally, we also work on the use of ontologies for cooperative query answering in a mediator approach. In [40], we have designed and implemented a tool (OntoRefiner) for refining queries with too many answers. By analyzing the class hierarchy of the ontology and the descriptions of the answers, OntoRefiner clusters the answers and provides a description of each cluster as a set of terms of the ontology. The user can then specialize his query by selecting the cluster whose description corresponds the best to his demand.

6.4. Active XML and Web Applications

Participants: Serge Abiteboul, Jerome Baumgarten, Omar Benjelloun, Gregory Cobena, Ioana Manolescu, Tova Milo, Benjamin Nguyen.

Key words: *Data integration*, *web services*, *peer-to-peer*.

Web services can be seen as the building blocks for complex software applications, see, e.g. Microsoft .Net, or BEA Web Logic. We have continued the development of the Active XML (AXML, for short) system, a declarative framework that harnesses Web services for data integration, and is put to work in a peer-to-peer architecture. The AXML system is centered around XML documents where some of the data is given explicitly while other parts are defined only intensionally by means of embedded calls to Web services. We considered the exchange of such documents between applications, and their distribution and replication among peers.

Exchange of AXML documents When such documents are exchanged between applications, one has the choice to materialize the intensional data (i.e. to invoke the embedded calls) or not, before the document is sent. This choice may be influenced by various parameters, such as performance and security considerations. Our research addressed the problem of guiding this materialization process [32]. We use types (ala DTD and XML Schema) to control the exchange of intensional data. We studied the problem and developed an implementation that complies with real life standards for XML data, schemas, and Web services, and is used in the Active XML system.

Distribution and replication Since dynamic documents may contain calls to services on other sites, some minimal form of distributed computation is inherently part of the model. A higher level of distribution, that also allows (fragments of) dynamic documents to be distributed and/or replicated over several sites is highly desirable in today's Web architecture. Starting from the data model and query language, we described a complete framework for distributed/replicated dynamic XML documents, in a peer-to-peer context including a cost model for query evaluation and an algorithm for recommending replication [21].

To validate our ideas in the above two directions, two prototypes were built and demonstrated in VLDB'03 [18][23]. The first demo illustrates the use of the exchange of AXML data in a peer-to-peer news publication and syndication system. The second demo illustrates the use of AXML as a platform for the management of distributed workspaces.

In the field of Web application design, conceptual modeling has already demonstrated its advantages, allowing for declarative specification, easier correctness checks, and automatic deployment, from a high-level model to implemented code. However, the conceptual modeling of applications using Web services has not yet been addressed. This problem has two orthogonal facets: the integration of basic Web service primitives within Web applications; and the design of Web applications driven by a given process specification, since Web services are often used to implement complex, "workflow-style" interactions among several peers. In collaboration with the WebML team from Politecnico di Milano, Italy, we have extended the WebML (http://www.Webml.org) modeling language with support for Web service interactions [9][27]. In parallel, we have demonstrated how the workflow dimension can be seamlessly integrated into the WebML model, allowing thus for the design of process-driven hypertext applications [24]. These results provided the basis for a tutorial on Web application modeling [26].

6.5. Thematic Web Warehousing

Participants: Serge Abiteboul, Omar Benjelloun, Jerome Baumgarten, Gregory Cobena, Amar-Djalil Mezaour, Tova Milo, Benjamin Nguyen, Marie-Christine Rousset.

Key words: *Warehouse*, *thematic information*, *declarative specification*.

Our research involves the development of a flexible and generic approach, which would let us specify in a declarative way the information necessary to create and enrich a thematic warehouse. We also want to simplify the acquisition of the documents that should be stored in the warehouse from the Web, monitor this warehouse, and organize the information it contains, for future querying.

We have begun a first experimental prototype, based on the Active XML language. To this end, we have programmed a library of Web services useful in order to construct a Web warehouse.

Another project on this theme has been running for nearly two years with the Athens University of Economics and Business, called Thesus. In Thesus, we follow a relational approach to specify the choice and enrichment of the resources we are interested in for a given theme. The major contribution of the Thesus project is to enrich the information retrieved, using links between documents. Semantic information is extracted from the neighborhood of links in the body of documents. The definition of measures of semantic similarity between such documents is also one of our interests.

The theme of focused crawling is also present in the e.dot project, and the Bibliothèque Nationale de France, French Web archiving project. In both cases, we want to integrate information found on the Web with high quality information already stored in a database. This approach combines calls to existing Web services, and machine learning techniques. These features concur to define a "best first" strategy for crawling Web pages, depending on what themes we are interested in.

6.6. Compressed XML data management

Participant: Ioana Manolescu.

Key words: *Compression techniques*, *semi-structured Data*, *query Optimization*.

XML suffers from the major limitation of high redundancy. Various compression techniques for XML data have been proposed, however, none of them allows efficient browsing and querying of the data. To address this problem, we propose a new approach for storing, compressing and querying XML data. We developed the *XQueC* system, an *XQue*ry processor and *C*ompressor, aiming at covering a large set of XQuery queries in the compressed domain. XQueC persistently stores XML documents using an embedded storage system, and is able to apply a large class of XML queries directly on compressed data. Due to its storage and query processing model, XQueC only uses decompression to construct the final results in a user-readable format. Furthermore, XQueC may take advantage of a query workload (if one is known) to exploit data commonalities and choose the compression algorithms and the compression granules. The XQueC system has been demonstrated at the VLDB conference [23]. Research is still ongoing within the XQueC system, in particular, on the aspect of transformation-based XQuery optimization.

7. Contracts and Grants with Industry

7.1. Introduction

Gemo has a continuous collaborations with France Telecom R&D, Xyleme and INRA. With the Bibliothèque Nationale de France, we are working on the archiving of the French Web. A RIAM proposal around this theme has been approved by the Ministry of Culture.

7.2. PICSEL2 Project

PICSEL2 is the continuation of PICSEL, which was the starting point of a collaboration with France Telecom R&D which has started in December 1997.

PICSEL2 aims at scaling up to the Web the mediator approach which has been implemented in PICSEL. The goal is to facilitate the automatic construction of a mediated schema over several XML sources described by DTDs and related to a same domain. A prototype (OntoMedia) has been developed, which extracts ontology components automatically from a set of DTDs. In PICSEL2, we also develop methods initiated in PICSEL for cooperative query answering.

7.3. RNTL Project e.dot

The goal of e.dot is to develop an XML warehouse for information concerning food risk. It is composed of Gemo, together with the BIA Group of the Institut National de Recherche en Agronomie that is specialized in this application, and the Xyleme company.

We specified the global architecture to meet the application needs and to take advantage of the existing data stored at INRA. We investigated several technical problems and the way to solve them, either by adapting existing Gemo technology or by developing new tools. The key point is that for constituting the XML warehouse, we are guided by a pre-existing ontology that was designed by INRA people as the uniform schema of their data, in order to acquire relevant documents from the Web, extract useful data from them, and transform them into XML documents using terms of the ontology as tags. We are investigating how to develop or reuse software components by packaging them as Web services, and how to use the SPIN architecture and Active XML for integrating them in a declarative way.

8. Other Grants and Activities

8.1. National Actions

Very close links exist with the database research group in the LRI (N. Bidoit, P. Rigaux, E Waller), the bioinformatics group in the LRI (C. Froidevaux, C. Rouveirol), the machine learning group in the LRI (M. Sebag), the Cedric Group in CNAM-Paris, the Atlas project in INRIA-Bretagne (F. Valduriez), The BIA group at INRA (O. Haemmerle, P. Buche, C. Dervin), the LISI of the University of Lyon 1 (M. Hacid), and the LIRMM of the University of Montpellier(M. Chein, M-L. Mugnier).

8.1.1. ACI Project ACI-MDD

This project is funded by the *ACI (Action Concertée Incitative) Masses de Données* and has just started. It is a joint project with Patrick Gallinari’s group of LIP6 (University of Paris 6) and Remi Gilleron’s Mostrare group of INRIA Futurs (Lille).

Faced to the rapid growth of structured documents (eg., XML documents) available online, the goal of this project is to to build tools for retrieving and extracting information, which fully and jointly exploit the structure and content of the XML documents.

Our goal is to bridge the gap between the search engines such as Google (exploiting the textual content only) and the data-centric approaches for querying XML documents. In those approaches, XML documents are considered as semi-structured data, which are queried almost exclusively through their structure by sophisticated query languages such as Xquery whose functionalities for dealing with content are yet very limited.

The distinguishing feature of our approach is to use machine learning techniques for building flexible and robust tools applicable to large corpora of structured documents, which are possibly heterogeneous, varied and dynamic.

8.1.2. ACI Project ACI-MDP2P

This project on Massive Data Management in Peer-to-Peer Systems is also funded by the ACI Masses de Données and has just started. It is a joint project with the Atlas, Paris and TexMex teams from INRIA-Bretagne. The goal of this project is to provide efficient data management tools in a peer-to-peer architecture. In this context, the Gemo team will study the management of distributed and replicated XML data in a peer-to-peer network [17], based on the existing Active XML platform [21]. In particular, the platform will be developed with functionalities for indexing and locating relevant data and services within a network of peers.

8.2. European Commission Financed Actions

Very close links exist with the University of Mannheim (G. Moerkotte), University of Marburg (T. Schwentick), University of Athens (M. Vazirgiannis), ETH Zurich (H. schek, R. Weber), University of Madrid (A. Gomez-Perez), University of Manchester (I. Horrocks), University of Rome (M. Lenzerini) and Politecnico di Milano (S. Ceri).

Particular projects that we conduct are detailed next.

8.2.1. Procope

This year we started a new project funded by Procope with the database group of Bernhard Seeger and Thomas Schwentick at Marburg University, Germany. The project is expected to last three years until the end of 2005. Its goal is to generate interactions between theory and practice in the context of systems for semi-structured data. More specifically we would like to find out whether we could develop automata- and logic-based methods for XML query evaluation and optimization.

8.2.2. European Project DBGlobe

DBGlobe, a database approach to solve the problem of distributed calculus at world wide scale is an IST project composed of the University of Ioannina, Hellas, Computer Technology Institute, Hellas, Research Center of the Athens University of Economics and Business, Hellas, University of Cyprus, University of California, Riverside, Technical University of Crete, Hellas, Aalborg University, Denmark, and INRIA, France.

The DBGlobe project aims at developing novel data management techniques to deal with the challenge of global computing. On the premise, global computing is a database problem: how to design, build and analyze systems that manage large amount of data.

However, the traditional database approach of storing data of interest in monolithic database management systems becomes obsolete in such environments. In current database research, data are relatively homogeneous, exhibit a small degree of distribution (just a few network sites) are passive in that they remain unchanged unless explicitly updated.

All these assumptions do not hold in the global computing world. This creates the need for new theoretical foundations in all aspects of data management: modeling, storage and querying.

8.3. Bilateral International Relations

8.3.1. Cooperation with the Middle-East

Very close links exist with the Hebrew University (C. Beeri) and the University of Tel-Aviv (T. Milo who is currently a visiting researcher at the group).

8.3.2. Cooperation with North America

Close links also exist with the Stanford University (J. Widom), AT&T (S.Amer-Yahia), Lucent-Bell Labs (J. Simeon), University of Washington (A. Halevy), University of Rutgers (A. Borgida), University of Toronto (A. Mendelzon and L. Libkin) and the University of California in San Diego (V. Vianu).

8.3.3. GemSaD

Since the beginning of the year, Gemo and the data management group at the University of California at San Diego (V. Vianu, A. Deutch, Y. Papakonstantinou) form an associated team funded by INRIA International. This association is expected to last at least three years. The two groups met at San Diego in June for a three day workshop. Two seniors and two Ph.d students from UCSD came to visit Gemo in Paris and stayed from one week to four months. Luc Segoufin spent two months at UCSD. The home page of GemSaD can be found at http://www-rocq.inria.fr/~segoufin/GEMSAD/

GemSad will now also be supported by the National Science Foundation for 3 years.

8.4. Visiting Professors

This year the following professors visited Verso :

- Tova Milo, professor at the University of Tel-Aviv (all year round)
- Victor Vianu, professor, UC San Diego (4 month)
- Michael Benedikt, researcher at Lucent Bell-labs (1 month)

9. Dissemination

9.1. Participation in Conferences

Our team has published a lot of papers in various international conferences and workshops (see bibliography). Some members of the project have participated in program committees. The list is given next.

S. Abiteboul

- PC Chair of the international VLDB conference (Very Large DataBases), 2003.
- International Conference on Database Theory (ICDT), Siena, Italy (2003)

I. Manolescu

- Workshop on Web-based Collaboration 2003; Web technologies and applications track in the ACM Symposion on Applied Computing
- Tutorials: “Constructing and integrating data-centric Web Applications: Methods, Tools, and Techniques”, tutorial at the VLDB 2003 conference [31]; “The nuts and bolts of DBMS construction: building your own prototype”, tutorial at the SBBD (Brazilian database) conference 2003 [20].
- Web Technologies and Applications, special track at the 18th ACM Symposium on Applied Computing (SAC 2003), Melbourne, Florida, USA.

T. Milo

- PC Chair of the ACM International Symposium on Principles of Database Systems, San Diego, CA, June 2003
- XML Database Symposium (Xsym) 2003, Berlin, Sept, 2003
- Extending DataBase Technology (EDBT) 2004

C. Reynaud

- Co-chair of the Semantic Web action spécifique du département STIC du CNRS (November 2001-September 2003)
- Co-organizer of Journée de l’action spécifique Web Sémantique du département STIC du CNRS, “ Web Sémantique et SHS”, Paris, France, 2003.
- DEXA Workshop on Web Semantics, Prague, Czech Republic, 2003.
- K-CAP Workshop on Knowledge Management and the semantic Web, Florida, USA, 2003.
- Première Journée Web Sémantique médical, Rennes, France, 2003
- Journée d’étude “Le Web Sémantique : de nouveaux enjeux documentaires ?”, Paris-la Défense, 2003.
- Journées Francophones de la Toile (JFT’2003), Tours, France, 2003.
- Journées Francophones d’Ingénierie des Connaissances (IC’2003), Laval, France, 2003.
- Coordination of the Web Semantic special report in Bulletin de l’AFIA, No 54, June 2003.

M-C. Rousset

- ACM International Symposium on Principles of Database Systems, 2003
- International Joint Conference on Artificial Intelligence (IJCAI), 2003
- IJCAI Workshop on Information Integration on the Web, 2003
- Journées Francophones d’Extraction et de Gestion des Connaissances (EGC), 2003
- Conférences sur les Bases de Données Avancéés, 2003
- Congrés Francophone de Reconnaissances des Formes et Intelligence Artificielle (RFIA), 2004

V. Vianu

- ACM International Symposium on Principles of Database Systems, San Diego, CA, June 2003

9.2. Invited Tutorials

V. Vianu

- Symposium on Theoretical Aspects of Computer Science, Berlin, January 2003
- 18th Brazilian Symposium on Databases, Manaus, Brazil, Oct. 2003

9.3. Scientific Animations

9.3.1. Editors

S. Abiteboul

- Information and Computation
- Journal of Digital Libraries

B. Amann

- Revue Information - Interaction - Intelligence (I3)

C. Reynaud

- JEDAI (Journal Electronique d'IA de l'AFIA)
- Revue Information - Interaction - Intelligence (I3)

M-C. Rousset

- ACM Transactions on Internet Technology (TOIT)
- AI Communications (AICOM)
- Electronic Transactions on Artificial Intelligence (ETAI) (for the areas : Concept-based Knowledge Representation and Semantic Web).
- Revue Information - Interaction - Intelligence (I3)

10. Bibliography

Books and Monographs

[1] J. CHARLET, C. REYNAUD, R. TEULIER.. *Ingénierie des connaissances pour les systèmes d'information.* 2003.

[2] C. REYNAUD, B. SAFAR, H. GAGLIARDI. *Une expérience de représentation d'une ontologie dans le médiateur PICSEL.* Eyrolles, 2003.

Doctoral dissertations and "Habilitation" theses

[3] G. COBÉNA. *Change management of semi-structured data on the Web.* Ph. D. Thesis, Ecole Polytechnique, 2003.

[4] B. NGUYEN. *Construction and maintenance of a Web warehouse.* Ph. D. Thesis, U. Paris Sud, 2003.

Articles in referred journals and book chapters

[5] N. Alon, T. Milo, F. Neven, D. Suciu, V. Vianu. *Typechecking XML views of relational databases.* in « ACM Trans. Comput. Logic », number 3, volume 4, 2003, pages 315-354.

[6] N. Alon, T. Milo, F. Neven, D. Suciu, V. Vianu. *XML with data values: typechecking revisited.* in « J. Comput. Syst. Sci. (JCSS) », number 4, volume 66, 2003, pages 688-727.

[7] M. Benedikt, M. Grohe, L. Libkin, L. Segoufin. *Reachability and Connectivity Queries in Constraint Databases.* in « JCSS », number 1, volume 66, 2003, pages 169-206.

[8] M. Benedikt, L. Libkin, T. Schwentick, L. Segoufin. *Definable Relations and First-Order Query Languages over Strings.* in « J. ACM », volume 50, 2003, pages 694-751.

[9] M. Brambilla, S. Ceri, S. Comai, P. Fraternali, I. Manolescu. *Specification and design of Workflow-Driven Hypertexts.* in « Journal of Web Engineering », 2003.

[10] S. de Amo, N. Bidoit, L. Segoufin. *Order independent temporal properties.* in « Journal of Logic and Computation, to appear », 2003.

[11] C. Delobel, C. Reynaud, M.-C. Rousset, J.-P. Sirot, D. Vodislav. *Semantic Integration in Xyleme: a Uniform Tree-Based Approach.* in « Data and Knowledge Engineering Review », number 3, volume 44, March, 2003, pages 267,298.

[12] M. Halkidi, B. Nguyen, I. Varlamis, M. Vazirgiannis. *THESUS: Organizing Web Document Collections Based On Semantics And Clustering.* in « The International Journal on Very Large Databases », number 4, volume 12, November, 2003, pages 320–332.

[13] T. Milo, D. Suciu, V. Vianu. *Typechecking for XML transformers.* in « J. Comput. Syst. Sci. (JCSS) », number 1, volume 66, 2003, pages 66-97.

[14] E. Pitoura, S. Abiteboul, D. Pfoser, G. Samaras, M. Vazirgiannis. *DBGlobe: a service-oriented P2P system for global computing.* in « SIGMOD Record », number 3, volume 32, 2003, pages 77-82.

[15] M.-C. Rousset, C. Reynaud. *Knowledge Representation for Information Integration.* in « Information Systems », 2003.

[16] R. Thomopoulos, P. Buche, O. Haemmerle. *Representation of weakly structured imprecise data for fuzzy querying.* in « Fuzzy Sets and Systems », number 1, volume 140, October, 2003, pages 111–128.

Publications in Conferences and Workshops

[17] S. Abiteboul. *Managing an XML Warehouse in a P2P Context.* in « Advanced Information Systems Engineering, 15th International Conference, CAiSE », volume Springer-Verlag, 2681, pages 4–13, 2003.

[18] S. ABITEBOUL, B. AMAN, J. BAUMGARTEN, O. BENJELLOUN, F. D. NGOC, T. MILO. *Schema-driven Customization of Web Services.* in « Conference on Very Large Data Bases, Demo », 2003.

[19] S. ABITEBOUL, V. BANSAL, G. COBÉNA, A. POGGI, B. NGUYEN. *Model, Design and Construction of a Service-Oriented Web-Warehouse.* in « European Conference on Digital Libraries, Demo », 2003.

[20] S. ABITEBOUL, J. BAUMGARTEN, A. BONIFATI, G. COBÉNA, C. CREMARENCO, F. DRAGAN, I. MANOLESCU, T. MILO, N. PREDA. *Managing Distributed Workspaces with Active XML.* in « Conference on Very Large Data Bases, Demo », 2003.

[21] S. ABITEBOUL, A. BONIFATI, G. COBÉNA, I. MANOLESCU, T. MILO. *Dynamic XML Documents with Distribution and Replication.* in « ACM Sigmod Conference on the management of data », 2003.

[22] S. ABITEBOUL, M. PREDA, G. COBÉNA. *Adaptive On-Line Page Importance Computation.* in « World Wide Web Conference », May, 2003.

[23] A. ARION, A. BONIFATI, G. COSTA, S. DÆAGUANNO, I. MANOLESCU, A. PUGLIESE. *XQueC: Pushing Queries to Compressed XML Data.* in « Conference on Very Large Data Bases, Demo », 2003.

[24] M. BRAMBILLA, S. CERI, S. COMAI, P. FRATERNALI, I. MANOLESCU. *Specification and design of workflow-driven hypertexts.* in « World Wide Web Conference, Poster », 2003.

[25] P. BUCHE, O. HAEMMERLE, R. THOMOPOULOS. *Integration of heterogeneous, imprecise and incomplete data: an application to the microbiological...* in « 14th International Symposium on Methodologies for Intelligent Systems, ISMISÆ2003 », series LNAI, volume 2871, Springer, pages 98–107, October, 2003.

[26] S. CERI, I. MANOLESCU. *Constructing and integrating data-centric Web Applications: Methods, Tools, and Techniques.* in « VLDB (Very Large Databases) Conference (tutorial) », Morgan Kauffman, J. C. FREYTAG, P. LOCKEMANN, S. ABITEBOUL, M. CAREY, P. SELINGER, A. HEUER., editors, pages 1151-1151, September, 2003.

[27] S. COMAI, I. MANOLESCU. *Conceptual Modeling Issues in Web applications enhanced with Web services.* in « Workshop on E-Services and the Semantic Web », May, 2003.

[28] F. GOASDOUE, M.-C. ROUSSET. *Answering Queries using Views: a KRDB Perspective for the Semantic Web.* in « ACM Journal - Transactions on Internet Technology (TOIT) », 2003.

[29] F. GOASDOUE, M.-C. ROUSSET. *Querying Distributed Data through Distributed Ontologies : a simple but scalable approach.* in « Information Integration on the Web », 2003.

[30] A. LERNER, I. MANOLESCU. *The nuts and bolts of DBMS construction: building your own prototype.* in « Simposio Brasileiro de Banco de Dados », 2003.

[31] I. MANOLESCU, S. CERI, M. BRAMBILLA, P. FRATEMALI, S. COMAI. *Exploring the combined potential of Web sites and Web services.* in « World Wide Web Conference, Poster », 2003.

[32] T. Milo, S. Abiteboul, B. Amann, O. Benjelloun, F. D. Ngoc. *Exchanging Intensional XML Data.* in « ACM Sigmod Conference on the management of data », 2003.

[33] B. Nguyen, G. Cobéna, M. Vazirgiannis. *Organization of Web Document Collections Based on Link Semantics.* in « European Conference on Digital Libraries, Demo », 2003.

[34] B. Nguyen, I. Varlamis, M. Halkidi, M. Vazirgiannis. *Construction de Classes de Documents Web.* in « Journees Francophones de la Toile », 2003.

[35] N. Pernelle, V. Ventos, H. Soldano. *ZooM : Alpha Galois Lattices for Conceptual Clustering.* in « Managing Specialization/Generalization Hierarchies (MASPEGHI) », October, 2003.

[36] M. Plu, P. Bellec, L. Agosto, W. V. de Velde. *The Web of People: A dual view on the WWW.* in « International World Wide Web Conference WWW2003 », 2003.

[37] C. Reynaud, G. Giraldo. *An application of the mediator approach to services over the Web.* in « Concurrent Engineering », 2003.

[38] C. Reynaud, G. Giraldo. *Mediation de services sur le Web.* in « Journee Francophones de la Toile », pages 59–68, 2003.

[39] B. Safar, H. Kefi. *Apport d'une ontologie du domain pour affiner une requete a l'aide d'un treillis de Galois.* 2003.

[40] B. Safar, H. Kefi. *Domain Ontology and Galois Lattice Structure for Query Refinement.* in « International Conference on Tools with Artificial Intelligence », November, 2003.

[41] L. Segoufin. *Typing and querying XML documents: some complexity bounds.* in « ACM Conference on Principles of Database System », 2003.

[42] R. Thomopoulos, P. Bosc, P. Buche, O. Haemmerle. *Logical Interpretation of Fuzzy Conceptual Graphs.* in « 22nd International Conference NAFIPS 2003 », pages 173–178, July, 2003.

[43] R. Thomopoulos, P. Buche, O. Haemmerle. *Different Kinds of Comparisons between Fuzzy Conceptual Graphs.* in « 11th International Conference on Conceptual Structures, ICCS 03 », series LNAI, volume 2746, Springer, pages 54–68, July, 2003.

Internal Reports

[44] A. Muscholl, T. Schwentick, L. Segoufin. *Active Context-Free Games.* Technical report, Gemo, 2003.