INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team gemo

Management of Data and Knowledge Distributed Over the Web

Futurs

THEME SYM

Activity Report

2005

Table of contents

1. Team

Managers

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

Administrative Assistant

Stéphanie Meunier

INRIA personnel

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

University personnel

Philippe Chatalic [Assistant Professor, Univ. Paris 11]
Hélène Gagliardi [Assistant Professor, Univ. Paris 11]
François Goasdoué [Assistant Professor, Univ. Paris 11]
Nathalie Pernelle [Assistant Professor, Univ. Paris 11]
Chantal Reynaud [Professor, Univ. Paris 11]
Brigitte Safar [Assistant Professor, Univ. Paris 11]
Laurent Simon [Assistant Professor, Univ. Paris 11]
Véronique Ventos [Assistant Professor,Univ. Paris 11]

Scientific Advisors

Philippe Dague [Professor, Univ. Paris 11]
Dan Vodislav [Assistant Professor, CNAM]
Benjamin Nguyen [Assistant Professor, UVSQ]

Invited researchers and visitors

Tarek Melliti [Post-doc, since September]
Victor Vianu [Professor, U.C. San Diego, 3 months]
Melanie Weis [Ph.D. student, TU Berlin, since November]

Engineers

Gabriel Vasile [Since July]
Boris Vrdoljak [Until August]

Ph.D students

Philippe Adjiman [Allocataire MENRT, Paris 11]
Andrei Arion [Allocataire MENRT, Paris 11]
Bogdan Cautis [Allocataire MENRT, Paris 11]
Claire David [ENS Cachan]
Gloria Giraldo [ATER, Paris 11, until June]
Hassen Kefi [Allocataire MENRT, Paris 11]
Jonathan Mamou [part-time member between Tel Aviv U. and Paris 11]
Amar-Djalil Mezaour [ATER, Paris 11]
Antonella Poggi [joint European Label PhD, U. Roma, until March]
Nicoleta Preda [Allocataire MENRT, Paris 11]
Fatiha Sais [Contrat FTRD]
Mathias Samuelides [ENS Cachan]
Pierre Senellart [ENS Ulm]

2. Scientific Foundations

2.1. Scientific Foundations

Keywords: *Databases*, *Web services*, *World Wide Web*, *XML*, *change control*, *complexity*, *data integration*,

distributed query, *knowledge representation*, *logic*, *peer-to-peer (p2p)*, *query optimization*, *query language*, *semantic integration*, *semi-structured data*.

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 Gemo is to study fundamental problems that are raised by modern information and knowledge management systems, and propose novel solutions to solve these problems. A main theme is the integration of information, seen as a general concept, including the discovery of meaningful information sources or services, the understanding of their content or goal, their integration and the monitoring of their evolution over time.

Gemo works on environments that are both powerful and flexible to simplify the development and deployment of applications providing fast access to meaningful data. In particular, content warehouses and mediators offering a wide access to multiple heterogeneous sources provide a good means of achieving these goals.

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. A particularity of the group is to address data and knowledge management issues by combining techniques coming from artificial intelligence (such as classification) and databases (such as indexing).

3. Application Domains

3.1. Application Domains

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

Databases do not have specific application fields. As a matter of fact, most human activities lead today to some form of data management. In particular, all applications involving the processing of large amounts of data require the use of databases.

Technologies recently developed within the group focus on novel applications in the context of the Web, telecom, multimedia, enterprise portals, or information systems open to the Web. In the setting of the RNTL e.dot project, we have built a tool box of Web services that have been applied to integrate data extracted from the Web to an existing data warehouse on food risk assesment (developed by the INRA BIA group). More generally, the approach developed in e.dot seems appropriate for applications of (scientific, technological, economic) watch.

4. Software

4.1. Software

Some recent software developed in Gemo:

ActiveXML: a language and system based on XML documents containing Web service calls. ActiveXML is now in Open Source within the ObjectWeb Forge.

SomeWhere: a P2P infrastructure for semantic mediation.

KadoP: a peer-to-peer platform for warehousing of Web resources [18], [19].

TaxoMap: a prototype to automate semantic mappings between taxonomies

TreeFinder, Dryade: prototype systems that discover frequent tree patterns within a collection of XML data.

OntoMedia: a prototype for the automatic construction of ontology components, using DTDs, developed within the PICSEL2 project.

WebQueL: a multi-criteria filtering tool for Web documents, developed in the setting of the e.dot project.

Acware: a prototype of Web warehouse definition and construction, based on a declarative language, and implemented using ActiveXML

ULoad: a tool for creating and storing XML materialized views, and using them to answer XQuery queries [26]

5. New Results

5.1. Theoretical foundations

Keywords: *Semi-structured data, automata, query languages, verification.*

Participants: Serge Abiteboul, Luc Segoufin, Victor Vianu.

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 on-going research task. Current research for XML usually follows two main directions: The first one is the classical *offline* setting where the goal is to design models for XML and to study the complexity and the expressive power of the associated query languages. The second axis, which can be seen as the *online* setting, is the study of XML when used in the dynamic environment of the Web.

Most of the current proposals for the first task, understanding XML in a offline setting, 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 [9] 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 XPath and various fragments of XPath (a W3C standard which is in the core of many XML query languages).

In the *online* setting we have considered **Active XML (AXML)** which 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 5.4). Peers exchanging AXML data agree on a *data exchange schema* that specifies in particular which parts of the data are allowed to be intensional [11]. Before sending a document, a peer may need to *materialize* some of the data by calling the corresponding services in order to match the agreed data exchange schema. Previous works showed that the rewriting problem is undecidable in the general case and of high complexity in some restricted cases. In [20], we study a relevant practical setting that is (1) in the spirit of standard 1-unambiguity constraints imposed on XML schemas and (2) can be solved by a single pass over the document with a simple computational device.

As XML is used as an exchange format for data over the Web, systems using XML, such as Web services, must manipulate highly heterogeneous data formats. In order to reduce the risk of failure it is therefore important to be able to perform offline static analysis of the programs developed in such systems. Gemo has started studying problems related to *verification* of systems for XML.

In the offline setting we have preliminary results exhibiting relevant decidable logics over trees that allows negation, a limited use of joins, and some limited form of recursion. These preliminary results are important because having at the same time negation, recursion and joins usually immediately yields undecidability.

In the online setting, as part of the ASAX project (see Section 7.1), we studied the diagnosis of distributed telecommunication systems. In [15] we show that (i) the problem can be modeled using Datalog programs, and (ii) it can benefit from the large battery of optimization techniques developed for Datalog. In particular, we show some surprising relationships between techniques used for solving efficiently the diagnosis problem of telecommunication systems and a known Datalog optimization technique, namely Query-sub-query (QSQ). We adapt QSQ to a distributed setting. We show that a simple "generic" use of this extended QSQ achieves an optimization as good as that previously provided by the dedicated diagnosis algorithms. Furthermore, we show that it allows solving efficiently a much larger class of system analysis problems.

5.2. XML and Service Mediation

Keywords: *clustering, information extraction, ontology, search engine, semantic mediation*.

Participants: Gloria Giraldo, Chantal Reynaud, Marie-Christine Rousset, Michèle Sebag, Véronique Ventos.

5.2.1. Information extraction from semistructured data

In the continuation of our work on discovering frequent trees in huge and heterogeneous collections of tree data, we have designed, implemented and experimented a new tree mining algorithm, DRYADEPARENT [46], based on the hooking principle first introduced in DRYADE. In the experiments, we demonstrate that the branching factor and depth of the frequent patterns to find are key factors of complexity for tree mining algorithms. We have shown that DRYADEPARENT outperforms the current fastest algorithm, CMTreeMiner, by orders of magnitude on datasets where the frequent patterns have a high branching factor.

We have supervised the PhD work of Jonathan Mamou (part time member of Gemo) on XSearch, a search engine for XML combining structure and content, which has been defended in September 2005.

Those two works are related to the project ACIMDD, supported by the ACI "Masses de Données".

We have consolidated our work on flexible clustering based on Galois lattices by giving a better efficiency to the construction of the alpha Galois lattices [47], [14].

5.2.2. Automatic service mediation in PICSEL2

The thesis about automatic construction of ontologies from DTDs for semantic mediation which has been funded by France Telecom R&D in the setting of the PICSEL2 project has been defended [1] in January.

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

Keywords: *RDF, composition of resources, distributed mediation, inconsistency, peer-to-peer systems*.

Participants: Philippe Adjiman, Philippe Chatalic, François Goasdoué, Marie-Christine Rousset, Laurent Simon.

We have consolidated the SomeWhere platform which is the basis for the starting MEDIAD project of France Telecom R&D, whose topic is multi-scaled distributed mediation. The current version of SomeWhere is being patented. New research directions are under investigation: (i) extending the data model to RDF(S); (ii) detecting and handling inconsistencies; (iii) extending the SomeWhere platform and data model to the P2P composition of resources for the Semantic Web; (iv) combining the SomeWhere platform with other P2P lookup or indexing platforms (KadoP, PINS).

The vision of the Semantic Web promoted in SomeWhere has obtained a high visibility through several invited talks (e.g., [22]).

5.4. ActiveXML and Web Applications

Keywords: *Data integration, peer-to-peer, web services*.

Participants: Serge Abiteboul, Omar Benjelloun, Bogdan Cautis, Ioana Manolescu, Tova Milo, Benjamin Nguyen.

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 ActiveXML (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 [15] - see also http://activexml.net. 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. Other aspects of ActiveXML have been considered such as security and access control [16].

In the field of Web application design, conceptual modeling allows for declarative specification, easier correctness checks, and automatic deployment, from a high-level model to implemented code. Our complete approach developed in collaboration with the WebML team from Politecnico di Milano, Italy, for specifying and deploying Web applications including Web services, has appeared in [10].

5.5. Thematic Web Warehousing

Keywords: *Warehouse*, *alignment techniques*, *declarative specification*, *thematic information*.

Participants: Serge Abiteboul, Hélène Gagliardi, Gloria Giraldo, Ollivier Haemmerlé, Hassen Kefi, Ioana Manolescu, Amar-Djalil Mezaour, Benjamin Nguyen, Nathalie Pernelle, Nicoleta Preda, Chantal Reynaud, Marie-Christine Rousset, Fatiha Sais, Brigitte Safar.

5.5.1. Acquisition of data from the web

We have designed, implemented and experimented a focused crawler called WebCrawler. It uses automatic learning techniques for improving the search based on statistics that are computed (and updated) on the exploration graph. It is based on the query language WeQuel [39] which had been previously developed for filtering Web pages fetched by a generalist crawler. This work is the core of a PhD [2] which has been defended in June 2005.

5.5.2. Extraction and integration of web data driven by an ontology

In the setting of the e.dot project, we have developed and experimented a method [35], [36], [34], [42] for an automatic semantic enrichment of data tables that are extracted from Web documents by means of tags and values coming from a domain ontology. Thanks to such a semantic transformation of data tables found on documents, we can query in a uniform way relational data of an existing database conform to a given schema and possibly heterogeneous tabular data coming from the Web.

5.5.3. Mapping between ontologies

In the context of the e.dot project, we worked on the mappings between different taxonomies in order to access to several sources from a unique querying system. We explored some alignment techniques to generate semantic mappings automatically. The originality of the approach is to be a combination of terminological, structured and semantic techniques well-suited to the mapping of taxonomies which are schemas with very poor definitions of concepts, so mainly defined with reference to the terminology. Moreover, an original exploitation of the structure of the models (the taxonomies or additional models such as WordNet) has been proposed. A prototype, TaxoMap, finds mappings or suggests indicators to help users find mappings. Experiment results in various domains have been obtained. A publication of this work has just been accepted and will be published in January 2006 [37].

Other research work on mappings has been initiated in the setting of Picsel3. The aim is to integrate a new XML source to a data warehouse thanks to an ontology. We study generation of mappings as a hybrid process based on both schemas of the sources and data.

5.5.4. Acware

We are also developing 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 ActiveXML language. To this end, we have programmed a library of Web services useful in order to construct a Web warehouse.

5.6. XML query optimization

Keywords: *Query Optimization*, *Semi-structured Data*.

Participants: Ioana Manolescu, Andrei Arion.

The problem of XML query evaluation still poses significant challenges. In particular, the complexity of the XQuery language, standardized by the W3C, makes it very difficult to devise efficient storage and optimization strategies.

Our work on generic XQuery optimization focuses on the ability to support a storage model that is varied, and may be changing, e.g., due to the addition of a new index, or a new materialized view. The ability of an optimizer to cope with such structures is a crucial factor for the system performance. We have proposed a formalism for describing a wide family of storage structures, encompassing existing storage and indexing schemes [25]. Furthermore, we have developed the ULoad prototype [26], which relies on this formalism to rewrite XQuery queries. Work in this area continues, in cooperation with V. Benzaken (U. Paris XI) and Y. Papakonstantinou (UCSD).

A different but complementary work direction targeted the establishment of a common well-grounded algebraic formalism for XPath and XQuery queries. Such a formalism is essential as a basis for optimization. In cooperation with Y. Papakonstantinou (UCSD) we have presented a tutorial on this topic at the ICDE conference [38]; we continue the work on defining an unified algebra [49].

5.7. XML Warehousing in P2P

Keywords: *P2P*, *Warehouse*, *XML*.

Participants: Serge Abiteboul, Ioana Manolescu, Nicoleta Preda.

We have designed and implemented KADOP, a peer-to-peer platform for building and managing warehouses of Web resources. The KadoP system was demonstrated at the ICDE conference [18] and its general approach was published in [19]. KADOP relies on a Distributed Hash Table implementation (namely, FreePastry) to keep the network of peers connected, and to build a shared global resource index, and on the ActiveXML platform to store, query, and maintain the index. Furthermore, KADOP is able to process simple queries carrying over resources distributed in the whole network.

A main goal is to be able to index not only extensional XML data but also intensional one and in particular Web services.

6. Contracts and Grants with Industry

6.1. Industrial contracts

Gemo has had technical meetings in 2004 with several industrial partners, in particular France Telecom R&D, Xyleme, Mandrakesoft, eNetshare, and Exalead, as well as national organizations, in particular, Institut National de Recherche en Agronomie and Bibliothèque Nationale de France.

6.2. MediaD project with France Telecom

The MediaD project aims at designing a declarative environment, SomeWhere, for building peer-to-peer data management systems based on a simple data model: propositional logic. A peer-to-peer data management system is a valuable alternative to a centralized information integration system like a mediator when the number of sources that have to be integrated becomes huge: building a global mediated schema coping with all the sources peculiarities is hardly possible and inefficient.

The goal of MediaD project is to deploy very large applications that scales to thousands of peers. It is organized in two tracks. The first one is to study query answering possibly in the presence of inconsistency. The second one is to develop techniques for cooperative statement of mappings that relate the knowledge of the different peers within the peer-to-peer data management system.

6.3. PICSEL3 project with France Telecom

This project is the continuation of PICSEL2 on scaling up to the Web the mediator approach that has been implemented in PICSEL1.

The goal is twofold. It aims at automating the construction of wrappers which translate user queries into the query language accepted by each source and return answers from the sources in the language of the mediator. This work is concerned with mediation of ontologies. Furthermore, we are interested in reference reconciliation, i.e. identifying when different references in a data set correspond to the same real-world reality.

6.4. EC Edos Project

In the Edos Project, the Gemo group focuses on improving the data process of distribution of open source software, a challenging issue because of the scale of the distribution (large number of files and size), its dynamicity, the need for replication for better performance and the autonomy of actors. A clean and general API for distribution has been defined and the architecture of the distribution system specified.

The distribution system is based on the idea of exchanging XML data in a P2P environment. We have specified a description of the content metadata description in XML. We are using the KadoP system. For Edos, the most critical component of KadoP is a DHT (distributed hash table) system that connects the network of peers. Originally, KadoP used Pastry, one of the most popular one. Serious problems of scaling and robustness were encountered. The file system-based storage of index in Pastry was replaced by BerkeleyDB. The fix was not sufficient, and now alternatives such as JXTA are being tested.

In the context of Edos, the KadoP system allows to publish and search for software packages. Thanks to the underlying use of Active XML by KadoP, some information about packages may be given extensionally and other intensionally. In the manner, we do not have to copy all the information that is available and can tune the data that is exchanged to the particular needs of an actor.

6.5. RNTL Project e.dot

The goal of the e.dot project is to develop an XML warehouse for information concerning food safety risk analysis and management. It is composed of Gemo, the BIA Group of the Institut National de Recherche en Agronomie (INRA) and the Xyleme company.

One key point of this project 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, transform them into semantically enriched XML documents (using ontology terms as element tags). We started by specifying and implementing the different functionalities and modules that are necessary to meet the application needs and to take advantage of existing application knowledge (ontologies) and data. These modules have been packaged and integrated in an open, service-oriented architecture composed of e.dot services for data acquisition and semantic enrichment, an e.dot XML warehouse for data storage (ActiveXML and Xyleme), and a graphical query interface (MIEL++) integrating this warehouse with several other pre-existing data sources. The service integration and warehouse definition process has been guided by using the SPIN approach and we are currently studying a new and more performant solution for the storage and querying of ActiveXML documents based on an integration of the ActiveXML system and the Xyleme product.

6.6. WS-DIAMOND EU project

The DIAMOND project has started on Sept. 1st 2005, and is due to last until Feb. 29th, 2008. Its full title is “Web Services - DIAgnosability, MONitoring and Diagnosis”. The project is coordinated by the Torino University, and involves the Polytechnic University of Milano, the Vrije University of Amsterdam, the Vienna University, U. Klagenfurt (Germany), and from France the LAAS-CNRS lab, U. Rennes, and U. Paris 11. Participants from Gemo are Philippe Dague (site leader for U. Paris 11), Tarek Melliti, Philippe Chatalic, Franç ois Goasdoué and Laurent Simon.

7. Other Grants and Activities

7.1. National Actions

In France, close links exist with groups at Orsay (databases, V. Benzaken and N. Bidoit; bio-informatics, C. Froidevaux, C. Rouveirol; machine learning, M. Sebag), with the Cedric Group at CNAM-Paris; some INRIA groups (Atlas, P. Valduriez, DistribCom, A. Benveniste, at INRIA-Bretagne); the BIA group at INRA (O. Haemmerlé, P. Buche, C. Dervin), the LIRIS of the University of Lyon 1 (M. Hacid), and the LIRMM of the University of Montpellier (M. Chein, M-L. Mugnier).

7.1.1. ACI Project ACI-MDD

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

The goal of this project is to study fundamental problems raised by modern information retrieval and to determine novel solutions to solve these problems. In particular, we want to build tools for retrieving and extracting information, which fully and jointly exploit the structure and contents of the XML documents. 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.

7.1.2. ACI Project ACI-MDP2P

This project on Massive Data Management in Peer-to-Peer Systems is funded by the ACI Masses de Données. MDP2P 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 2005, Gemo has pursued the development of KadoP (*Knowledge and Data in Peer-to-Peer*), a platform for large-scale sharing of resources over many peers, based on a distributed hash table [18]. We have also studied the advantages of intensional information for the evaluation of XML queries in this setting [19]; this parf of the work is still ongoing.

7.1.3. ACI Project TraLaLa

TraLaLa stands for XML Transformation Languages: logic and applications. It is funded by the *ACI (Action Concertée Incitative) Masses de Données* and has started in September 2004. The setting is the integration and manipulation of massive data in XML format. We are interested more specifically in the programming and querying languages aspects: expressivity, typing, optimization. We are also interested in studying how this can be done in a context where documents are compressed or in a streaming scenario. The project is funded for three years. Its home page can be found at : http://www.cduce.org/tralala.html.

7.1.4. ACI Normes et Politiques Publiques

This project has started in 2005, as a collaboration with Benjamin Nguyen (University of Versailles) and with several political scientists (F.-X. Dudouet from University of Paris X). Our purpose is to investigate the process of drawing up Information Technology standards and regulations, to understand how the process proceeds, who are the actors involved, and which are the actual mechanisms for setting up a standard. We have designed an architecture for the construction of a semistructured data warehouse, used as a support for verifying the social scientists' hypothesis, and have performed a preliminary study of the interactions on the public mailing list of the standards committee [48], [33].

7.2. European Commission Financed Actions

In Europe, close links exist with University of Marburg (T. Schwentick), University of Athens (M. Vazirgiannis), University of Madrid (A. Gomez-Perez), University of Manchester (I. Horrocks), University of Rome (M. Lenzerini) and the Systems and Computer Engineering Research Institute of Lisbon (H. Galgardas).

Particular projects that we conduct are detailed next.

7.2.1. Procope

This year was the last year of a PAI-Procope project with the database group of Bernhard Seeger and Thomas Schwentick at Marburg University, Germany. The project will end in 2005. Its goal was to generate interactions between theory and practice in the context of systems for semi-structured data. It produced several join papers between the two groups.

7.2.2. XClean

This is the first year of the XClean PAI project, joint work with the database team from INESC-ID Lisbon (Portuguese Research Institute on Systems and Computer Engineering, team of Helena Galhardas and Pavel Calado). XClean is a three-years project. Its goal is to propose models, languages, and systems for XML data cleaning. The presence of multiple heterogeneous data collections creates the need to isolate single, correct descriptions of real-world objects out of multiple, conflicting, potentially dirty ones. While ad-hoc cleaning programs are usually written for every specific data set, scalability requires a formal approach based on algebraic operators. We investigate the integration of such operators in an algebra for XML queries. This work is mainly carried out as part of Melanie Weis' internship.

7.3. Bilateral International Relations

7.3.1. Cooperation with the Middle-East

Close links exist with the Hebrew University (C. Beeri) and the University of Tel-Aviv (T. Milo returned to Israel in 2004 after a long visit in the group).

7.3.2. Cooperation with North America

In the US, close links also exist with the Stanford University (J. Widom), AT&T (S.Amer-Yahia), University of Washington (A. Halevy), University of Rutgers (A. Borgida), University of Toronto (L. Libkin).

7.3.3. French-US team: GemSaD

Since 2003, 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 in Baltimore in June in parallel with the SIGMOD conference. Ioana Manolescu spent some time in San Diego and Victor Vianu is spending a sabbatical year in Paris. The home page of GemSaD can be found at http://www-rocq.inria.fr/~segoufin/GEMSAD/. GemSad is also supported by the National Science Foundation until 2006.

7.4. Visiting Professors and Students

This year the following professors visited Verso:

- Tova Milo, professor at the University of Tel-Aviv (in July)
- Neoklis Polyzotis, professor at the University of Southern California (in July)
- Victor Vianu, professor, UC San Diego (September to December)

The following students came for internships in the group: Alan Nash [UCSD; 1 month]; Antonella Poggi [U. Roma; 3 months; joint European PhD]; Emanuel Taropa [KAIST Seoul; 2 months, undegraduate internship] Ravi Vijay [IIT Bombay; 2 months, undegraduate internship] Wendy Wang [U. British Columbia; 3 months, PhD internship] Melanie Weis [Humboldt U. Berlin; 3 months, PhD internship]

8. Dissemination

8.1. Participation in Conferences

Serge Abiteboul was the Regional Coordination Committee Chair of the ACM SIGMOD/PODS 2005 Conference.

Marie-Christine Rousset has been nominated ECCAI Fellow 2005 (European Coordinating Committee for Artificial Intelligence).

Ioana Manolescu has joined the W3C XQuery Working Group in January 2005.

Members of the project have participated in program committees:

S. Abiteboul

- 6th International Workshop on Next Generation Information Technology and Systems (NGITS) 2005
- IEEE International Conference on Web Services (ICWS) 2005
- World Wide Web Conference (WWW) 2005
- International Conference on Very Large Databases (VLDB) 2005

I. Manolescu

- International Conference on Very Large Databases (VLDB) 2005
- International Conference on Data Engineering (ICDE) 2005
- International Conference on Service-Oriented Computing (ICSOC) 2005
- ACM International Conference on Information and Knowledge Management (CIKM) 2005
- WWW Conference 2005
- ACM Workshop on Web Information and Data Management (WIDM) 2005
- International Workshop on XQuery Implementation, Experience and Perspectives (XIME-P) 2005
- International Workshop on Databases and the Web (WebDB) 2005
- XML Symposium, in cooperation with VLDB 2005
- Indian Conference on Management of Data (COMAD) 2005
- Brazilian Database Symposium (SBBD) 2005
- Journées de Bases de Données Avancées (BDA) 2005

C. Reynaud

- 4th International Semantic Web Conference (ISWC), 2005.
- Atelier Modélisation des connaissances, EGC, Janvier 2005.
- Ingénierie des Connaissances, Mai-Juin 2005.

M-C. Rousset

- International Conference on Cooperative Information Systems (CoopIS 2005)
- 4th International Semantic Web Conference (ISWC 2005)
- 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)
- ACM Conference on Principles Of Database Systems (PODS 2005)
- 15th International Symposium on Methodologies for Intelligent Systems (ISMIS 2005),
- Symposium on Abstraction, Reformulation and Approximation (SARA 2005)

L. Segoufin

- Asian Computing Science Conference Data Management on the Web, Kunming, China, December 2005.
- ACM Conference on Principles Of Database Systems (PODS 2005)

L. Simon

- International Conference on Theory and Applications of Satisfiability Testing (SAT 2005)
- Internaional Workshop on Empirically Successful Classical Automated Reasoning (ESCAR 2005)
- Journées Francophones de Programmation par Contraintes (JFPC 2005)

V. Ventos

- Conférence d'Apprentissage (CAPS 2005)

8.2. Participation to the W3C Working Groups

In 2005, GEMO has actively participated in the World Wide Standardization Consortium. The W3C's objective is to develop interoperable technologies, by the creation of specifications, guidelines, software and tools, in the field of the World Wide Web. Since W3C is a forum for information, commerce, communication, and collective understanding, and that INRIA is a founding member, our participation attains the double objective of getting to know the world of standardization better, and bringing in a perspective from INRIA, a public European research institute, in groups normally driven by companies such as IBM, Oracle, Microsoft, HP etc.

GEMO has participated in two different activities (W3C has around 20 activities) : Extensible Markup Language and Semantic Web. More specifically inside these activities, GEMO members have been present in the XQuery Working Group, and the Semantic Web Best Practices and Development Working Group. Participation in these working groups is assured by the attendance to teleconferences and face-to-face meetings along with posting to a mailing list (a few hundred mails per month).

8.2.1. Semantic Web Best Practices

This working group, chartered for two years since March 2004, is due for a charter review in February 2006. Its goal was to provide hands-on support for developers of Semantic Web applications. This working group helps application developers by providing them with "best practices" in various forms, ranging from engineering guidelines, ontology / vocabulary repositories to educational material and demo applications. The Semantic Web Tutorials Page task force will possibly be integrated in the new Semantic Web Education and Outreach Working Group that may start next year.

B. Nguyen has participated to management of the Tutorials Page task force (since April 2005).

8.2.2. Extensible Markup Languages

The W3C working group on XML query has designed and refined, over the last five years, the XQuery language for querying XML. XQuery will get the “Candidate Recommendation” status before the end of 2005, which is an important achievement. One of the major current XQuery development concerns updates. A candidate syntax and a first draft of the semantics have been proposed in the last few months; the Update task force expects to publish a first Working Draft before the end of 2005.

I. Manolescu is one of the two editors of the XQuery Update Use Cases document, featuring sample modifications of an XML database. Such samples are meant to guide and provoke discussions on language design, and to inform potential users.

8.3. Invited Presentations

Marie-Christine Rousset has presented an invited tutorial on *Semantic Web Challenges* at the INFORSID conference, 2005.

Ioana Manolescu and Yannis Papakonstantinou (UCSD) presented a tutorial on *XQuery Midflight: Emerging Database-Oriented Paradigms and a Classification of Research Advances* at the International Data Engineering Conference, 2005.

8.4. Scientific Animations

Editors

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)

L. Simon

- Member of the Editorial Board of JSAT (the Journal on Satisfiability, Boolean Modeling and Computation)

9. Bibliography

Doctoral dissertations and Habilitation theses

[1] G.-L. GIRALDO-GOMEZ. *Construction automatisée de l'ontologie de systèmes médiateurs*, Ph. D. Thesis, Université Paris Sud, 2005.

[2] A.-D. MEZAOUR. *Recherche ciblée de documents sur le web*, Ph. D. Thesis, Université Paris Sud, 2005.

Articles in refereed journals and book chapters

[3] S. ABITEBOUL, R. AGRAWAL, P. A. BERNSTEIN, M. J. CAREY, S. CERI, W. B. CROFT, D. J. DEWITT, M. J. FRANKLIN, H. GARCIA-MOLINA, D. GAWLICK, J. GRAY, L. M. HAAS, A. Y. HALEVY, J. M. HELLERSTEIN, Y. E. IOANNIDIS, M. L. KERSTEN, M. J. PAZZANI, M. LESK, D. MAIER, J. F. NAUGHTON, H.-J. SCHEK, T. K. SELLIS, R. SILBERSCHATZ, M. STONEBRAKER, R. T. SNODGRASS, J. D. ULLMAN, G. WEIKUM, J. WIDOM, S. ZDONIK. *The Lowell database research self-assessment*, in "Communications of the ACM", vol. 48, n^o 5, 2005, p. 111-118.

[4] S. ABITEBOUL, B. NGUYEN, G. RUBERG. *Building an Active Content Warehouse*, in "Managing and Processing Complex Data for Decision Support", J. DARMONT, O. BOUSSAID (editors). , IDEA Group Publishing, 2005.

[5] S. ABITEBOUL, L. SEGOUFIN, V. VIANU. *Representing and Querying XML with Incomplete Information*, in "ACM TODS", 2005.

[6] P. ADJIMAN, P. CHATALIC, F. GOASDOUÉ, M.-C. ROUSSET, L. SIMON. *Distributed Reasoning in a Peer-to-Peer Setting: Application to the Semantic Web*, in "Journal of Artificial Intelligence Research", 2006.

[7] A. ARION, A. BONIFATI, I. MANOLESCU, A. PUGLIESE. *Un modèle de stockage de données XML basé sur les séquences*, in "Ingeniérie des Systèmes d'Information", vol. 2, Octobre 2005.

[8] D. L. BERRE, P. PURDOM, L. SIMON. *A phylogenetic tree for the SAT 2002 contest*, in "Annals of Mathematics and Artificial Intelligence (AMAI)", n^o 43, 2005.

[9] G. GOTTLOB, C. KOCH, R. PICHLER, L. SEGOUFIN. *The Parallel Complexity of XML Typing and XPath Query Evaluation*, in "J. ACM", vol. 52, n^o 2, 2005, p. 284–335.

[10] I. MANOLESCU, M. BRAMBILLA, S. CERI, S. COMAI, P. FRATERNALI. *Model-Driven Design and Deployment of Service-Enabled Web Applications*, in "ACM Transactions on Internet Technology", vol. 5, n^o 3, August 2005.

[11] T. MILO, S. ABITEBOUL, B. AMANN, O. BENJELLOUN, F. D. NGOC. *Exchanging Intensional XML Data*, in "ACM Transactions on Database Systems", vol. 30, n^o 1, March 2005, p. 1-40.

[12] A. MUSCHOLL, M. SAMUELIDES, L. SEGOUFIN. *Complementing deterministic tree-walking automata*, in "Information processing letters", 2005.

[13] C. REYNAUD, B. SAFAR, H. GAGLIARDI. *Une expérience de représentation d'une ontologie dans le médiateur PICSEL*, R. TEULIER, J. CHARLET, P. TCHOUNIKINE (editors). , L'Harmattan, 2005.

[14] V. VENTOS, H. SOLDANO. *Treillis de Galois Alpha*, in "Revue d'Intelligence Artificielle", n^o 19, 2005, p. 799–827.

Publications in Conferences and Workshops

[15] S. ABITEBOUL, Z. ABRAMS, S. HAAR, T. MILO. *Diagnosis of Asynchronous Discrete event systems. Datalog to the rescue!*, in "ACM Conference on Principle of Database Systems", 2005.

[16] S. ABITEBOUL, B. CAUTIS, A. FIAT, H. KAPLAN, T. MILO. *Secure Exchange of Modifiable Data and Queries*, in "Bases de Données Avancées (BDA)", 2005.

[17] S. ABITEBOUL, X. LEROY, B. VRDOLJAK, R. D. COSMO, S. FERMIGIER, S. LAURIÈRE, F. LEPIED, R. POP, F. VILLARD, J.-P. SMETS, C. BRYCE, K. R. DITTRICH, T. MILO, A. SAGI, Y. SHTOSSEL, E. PANTO. *EDOS: Environment for the Development and Distribution of Open Source Software*, in "Proc. of the 1st Int'l Conf. on Open Source Software Systems", 2005.

[18] S. ABITEBOUL, I. MANOLESCU, N. PREDA. *Constructing and Querying Peer-to-Peer Warehouses of XML Resources*, in "International Data Engineering Conference (ICDE), demo", IEEE Computer Society, April

2005, p. 1122-1123.

[19] S. ABITEBOUL, I. MANOLESCU, N. PREDA. *Sharing Content in Structured P2P Networks*, in "Bases de Données Avancées (BDA)", October 2005.

[20] S. ABITEBOUL, T. MILO, O. BENJELLOUN. *Regular Rewriting of Active XML and Unambiguity*, in "ACM Conference on Principles of Database Systems (PODS)", 2005.

[21] P. ADJIMAN, P. CHATALIC, F. GOASDOUÉ, M.-C. ROUSSET, L. SIMON. *Scalability Study of Peer-to-Peer Consequence Finding*, in "International Joint Conference on Artificial Intelligence", 2005.

[22] P. ADJIMAN, P. CHATALIC, F. GOASDOUÉ, M.-C. ROUSSET, L. SIMON. *SomeWhere in the Semantic Web*, in "International Workshop on Principles and Practice of Semantic Web Reasoning", 2005.

[23] L. AFANASIEV, I. MANOLESCU, P. MICHIELS. *MemBeR: A Micro-Benchmark Repository for XQuery*, in "International XML Database Symposium (XSym)", S. BRESSAN, S. CERI, E. HUNT, Z. IVES, Z. BELLAHSENE, M. RYS, R. UNLAND (editors). , LNCS, vol. 3671, Springer-Verlag, September 2005, p. 144-161.

[24] É. ALPHONSE, A. AMRANI, J. AZE, T. HEITZ, A.-D. MEZAOUR, M. ROCHE. *Préparation des données et analyse des résultats de DEFT'05*, in "Proceedings of DEFT'05 workshop of the national conference TA LN", june 2005.

[25] A. ARION, V. BENZAKEN, I. MANOLESCU. *XML Access Modules: Towards Physical Data Independence in XML Databases*, in "International Workshop on XQuery Implementation, Experience and Perspectives (XIME-P)", June 2005.

[26] A. ARION, V. BENZAKEN, I. MANOLESCU, R. VIJAY. *ULoad: Choosing the Right Storage for Your XML Application*, in "Very Large Databases Conference (VLDB), demo", K. BOHM, C. JENSEN, L. HAAS, M. KERSTEN, P.-A. LARSON, B.C. OOI (editors). , August 2005, p. 1330-1333.

[27] M. ATTNÄS, P. SENELLART, J. SENELLART. *Integration of SYSTRAN MT systems in an open workflow*, in "MT Summit X", September 2005.

[28] M. BANEK, Z. SKOCIR, B. VRDOLJAK. *Logical Design of Data Warehouses from XML*, in "Proc. of the 8th Int'l Conf. on Telecommunications ConTEL", vol. 1, 2005, p. 289-295.

[29] M. BENEDIKT, L. SEGOUFIN. *Regular tree languages definable in FO*, in "STACS", 2005.

[30] M. BENEDIKT, L. SEGOUFIN. *Towards a Characterization of Order-Invariant Queries over Tame Structures*, in "CSL", 2005.

[31] D. L. BERRE. *The SAT 2002 competition.*, in "Annals of Mathematics and Artificial Intelligence (AMAI)", 2005, p. 343–378.

[32] S. Cohen-Boulakia, S. Davidson, C. Froidevaux. *A User-centric Framework for Accessing Biological Sources and Tools*, in "DILS", 2005.

[33] F.-X. Dudouet, I. Manolescu, B. Nguyen, P. Senellart. *XML Warehousing Meets Sociology*, in "IADIS International Conference on the Web and Internet", October 2005.

[34] H. Gagliardi, O. Haemmerlé, D. Migliori, N. Pernelle, M.-C. Rousset, F. Sais. *Enriching a relational datawarehouse by integrating XML data : report ont the edot project*, in "Second Franco Japanese Workshop on Information Search Integration and Personnalization (ISIP)", 2005.

[35] H. Gagliardi, O. Haemmerlé, N. Pernelle, F. Sais. *A semantic enrichment of data tables applied to food risk assessment*, in "Discovery in Science", LNAI Springer Verlag, october 2005.

[36] H. Gagliardi, O. Haemmerlé, N. Pernelle, F. Sais. *An automatic ontology-based approach to enrich tables semantically*, in "Workshop AAAI on Context and Ontology", july 2005, p. 64–71.

[37] H. Kefi, B. Safar, C. Reynaud. *Alignement de taxonomies pour l'interrogation de sources d'information hétèrogènes*, in "Reconnaissance de Formes et Intelligence Artificielle", 2006.

[38] I. Manolescu, Y. Papakonstantinou. *XQuery Midflight: Emerging Database-Oriented Paradigms and a Classification of Research Advances*, in "International Data Engineering Conference (ICDE)", April 2005, p. 1143-1143.

[39] A.-D. Mezaour. *Filtering Web Documents for a Thematic Warehouse, Case Study : eDot a Food Risk Data (extended)*, in "New Trends in Intelligent Information Processing and Web Mining Conference (IIPWM", Springer Verlag series Advances in Soft Computing , June 13-16 2005, p. 269-278.

[40] D. Migliori, M.-C. Rousset, O. Haemmerlé. *Knowledge Management by Querying Relational Views of XML data: application to Microbiology*, in "IJCAI 2005 workshop on Knowledge Management", 2005.

[41] A. Poggi, S. Abiteboul. *XML data integration with identification*, in "Tenth International Symposium on Database Programming Languages (DBPL)", 2005.

[42] F. Sais, H. Gagliardi, O. Haemmerlé, N. Pernelle. *Enrichissement sémantique de documents XML représentant des tableaux*, in "Extraction et Gestion de Connaissances (EGC)", vol. 2, RNTI Cepadues ed., january 2005, p. 407–419.

[43] L. Segoufin, V. Vianu. *Views and Queries: Determinacy and Rewriting*, in "PODS", 2005.

[44] P. Senellart. *Identifying Websites with Flow Simulation*, in "International Conference on Web Engineering", M. G. David Lowe (editor). , Springer, July 2005, p. 124–129.

[45] P. Senellart, J. Senellart. *SYSTRAN Translation Stylesheets: Machine Translation driven by XSLT*, in "XML Conference and Exhibition", November 2005.

[46] A. TERMIER, M.-C. ROUSSET, M. SEBAG, K. OHARA, T. WASHIO, H. MOTODA. *Efficient mining of high branching factor attribute trees*, in "ICDM 05 (International Conference on Data Mining)", 2005.

[47] V. VENTOS, H. SOLDANO. *Alpha Galois Lattices: an overview*, in "International Conference in Formal Concept Analysis Lecture Notes on Computer Science", n^o 3403, 2005, p. 298-313.

Internal Reports

[48] F.-X. DUDOUET, I. MANOLESCU, B. NGUYEN, P. SENELLART. *XML Warehousing Meets Sociology... Introducing the W3C XQuery Working Group*, Technical report, Gemo, April 2005.

[49] I. MANOLESCU, Y. PAPAKONSTANTINOU. *An Unified Tuple-Based Algebra for XQuery*, Technical report, April 2005.

[50] P. SENELLART. *Identifying Websites with Flow Simulation*, Technical report, Gemo, April 2005.