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

2004

Table of contents

1. Team

Managers

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

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]
Ollivier Haemmerle [Assistant Professor, INA P-G]
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

Bernd Amann [Assistant Professor, CNAM; Professor, Paris 6 since Sept. 2004]

Invited researchers

Tova Milo [Professor, U. Tel Aviv; till Sept.]
Victor Vianu [Professor, U.C. San Diego, 3 months]

Engineers

Jérôme Baumgarten [till August]
Nicolaas Ruberg [Bank NDES of Brazil, till August]

Ph.D students

Philippe Adjiman [Allocataire MENRT, Paris 11]
Andrei Arion [Allocataire MENRT, Paris 11, from Sept.]
Omar Benjelloun [Allocataire MENRT, Paris 11, till June, Postdoc Stanford]
Bogdan Cautis [Allocataire MENRT, Paris 11, from Sept.]
Gloria-Lucia Giraldo [Paris 11]
Hassen Kefi [Allocataire MENRT, Paris 11]
Amar-Djalil Mezaour [Allocataire MENRT, Paris 11]
Benjamin Nguyen [Allocataire MENRT, Paris 11, till August, Assistant Professor UVSQ]
Antonella Poggi [joint European Label PhD, U. Roma, from Sept.]
Nicoleta Preda [Allocataire MENRT, Paris 11from Sept.]
Gabriella Ruberg [Federal University of Rio de Janeiro, till June]
Fatiha Sais [Contrat FTRD, from Sept.]
Mathias Samuelides [ENS Cachan]
Pierre Senellart [ENS Ulm, from Sept.]
Alexandre Termier [Allocataire MENRT, Paris 11, till June, Postdoc Japan]

2. 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. Introduction

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. They often consider data integration as in the application of food risk that we mention next.

3.2. A warehouse on food risk

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

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.

4. 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 [27][28].
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.
XQueC: a prototype for storing and querying compressed XML data.
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
STYX: definition and construction of a generic platform to integrate and query relevant XML resources concerning a Web community.
XSum: an free open-source XML visualization and analysis tool [26].
GeX: an XML database prototype [37]

5. New Results

5.1. Theoretical foundations

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

Participants: Serge Abiteboul, Tova Milo, Antonella Poggi, 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 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 further our results, first for XML and then for ActiveXML.

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 [54] 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).

ActiveXML (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 5.4). Each such call is typed using XML schemas. A formal foundation for this new generation of AXML documents and services is presented in [25]. This paper also study the fundamental issues they raise. The focus is on Web services that are (1) monotone and (2) defined declaratively as conjunctive queries over AXML documents. In this case it studies the semantics of documents and queries, the confluence of computations, termination and lazy query evaluation.

In the AXML setting, when a user requests 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. In [39] and [59] the general problem is shown to be non computable and an overview of many decidable cases is presented.

5.2. XML and Service Mediation

Keywords: *Semantic integration*, *clustering*, *ontologies*, *structure extraction*.

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

5.2.1. Information extraction from semistructured data

In the continuation of our work on discovering frequent trees in huge collections of tree data, we have proposed a novel algorithm (DRYADE) for discovering frequent trees, even in the presence of variations in the nesting of the labels of XML documents. The solution experimented in the Dryade algorithm [46][13] is to compute closed frequent trees, which has the advantage of providing a compact representation of frequent trees without loss of information. The performance of Dryade has been experimentally tested on artificial medium-size data and also on real data in the setting of the Xyleme project1.

We have extended our work implemented in the ZooM system for clustering semistructured data based on Galois lattices by the definition and the study of new Galois lattices, which we have called alpha Galois lattices, allowing the control of the number of nodes of the lattice and thus the adaptability to users and data. The theoretical foundations of alpha Galois lattices have been presented in [48][49].

5.2.2. Automating service mediation in PICSEL2

In the continuation of the work on the automatic construction of a ontology from a set of XML documents corresponding to standardized service specifications in a given domain, we have extended the prototype OntoMedia by functionalities for generating suitable user interfaces from the ontology and for generating wrappers from the XML service specifications [40].

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

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

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

5.3.1. Semantic peer-to-peer data management systems

We have designed the Somewhere PDMS whose data model complies with the W3C recommendations since it is captured by the propositional fragment of the OWL ontology language. Our work on distributed reasoning in propositional logic provides the query-answering algorithm of Somewhere since we have shown that in our data model query answering amounts to a proper prime implicant calculus [29][55]. This simple data model already scales up to more than a thousand peers in our first benchmarks. Moreover, it guarantees that the query-answering problem is decidable. Somewhere is the basis for the forthcoming MediaD project in collaboration with France Telecom R&D on efficient distributed mediation.

5.3.2. Mediation with end-users

In continuation of our work on query refinement based on the construction of a Galois lattice, we have made several experiments on the OntoRefiner prototype in order to analyze the number and the content of the clusters when there are a lot of answers, and we have implemented several optimizations [43]. Our heuristics exploit the taxonomy of concepts and allow to compute only approximate clusters, which are very close to the real clusters. The use of approximate clusters allow to avoid the closure computation which is the most costly step.

In the context of ActiveXML, we have addressed the problem of "dangling" service calls, i.e. service calls that return an error instead of a valid result. We have introduced the notion of *service call plan* which is composed of a service call and a set of possible alternative service calls fired by specific error events. A prototype has been implemented by Radu Pop during his internship and we are currently studying the semantic description and automatic generation of service call plans as a semantic extension of the ActiveXML model.

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 [20][12][11] - see also `http://activexml.net`. An overview of the project is now available [51]. 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.

The issue of query processing in this context is addressed in [22]. Other aspects of ActiveXML have been considered such as the flexibilily of query answering [24], security and access control [23]

We have performed a detailed performance analysis on the AXML peer implementation for PCs [42], identifying all elementary cost components associated to a service call activation. Furthermore, we have proposed a new execution model and an associated distributed optimization strategy for the problem of AXML document materialization [42].

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. In collaboration with the WebML team from Politecnico di Milano, Italy, we have extended the WebML (`www.Webml.org`) modeling language with support for Web services and workflow; we demonstrated WebRatio, a commercial tool implementing this approach, in the SIGMOD conference 2004 [33].

5.5. Thematic Web Warehousing

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

Participants: Serge Abiteboul, Omar Benjelloun, Gloria Giraldo, Hassen Kefi, Ioana Manolescu, Amar-Djalil Mezaour, Tova Milo, Benjamin Nguyen, Nathalie Pernelle, Nicoleta Preda, Chantal Reynaud, Marie-Christine Rousset, Brigitte Safar.

This theme has been driven in Gemo by the RNTL e.dot project. Thematic warehouses integrate data collected from multiple sources and relative to a same application domain described by an ontology. In contrast with mediators, which manage virtual data, warehouses store data extracted from information sources and then integrated with the data already present in the warehouse. In all our work on Thematic Warehouses, a domain ontology is used. It guides data acquisition, extraction and integration. Moreover, as XML is nowadays the new standard for the exchange of data, it has been chosen as the representation language of the warehouse. Our work is motivated by and developed in the context of the e.dot RNTL project which aims at enriching an existing data warehouse (in the domain of food risk assessment) with data from the Web (http://www.lri.fr/~cr/images/SchemaEdot.JPG). In e.dot, the domain ontology describes the contents of the warehouse in two parts: a relational database and data represented in the conceptual graph formalism. The ontology plays the role of the schema of the warehouse. It is a set of concepts organized in a hierarchy with their synonyms in French and their translation in English. The organisation and the materialisation of the warehouse is specified in a high-level model specific to thematic warehouses and with all operations invoked through Web services (ACWARE Active Content Warehouse). These specifications are executed by ActiveXML, a framework developed by Gemo to deal with embedded Web services calls inside Web documents [25].

5.5.1. Acquisition of data from the Web

Our work combines a search engine such as Google or a web crawler (such as that of Xyleme) with a filtering tool that can distinguish, among the possible thousands of web pages returned by Google or the Xyleme crawler, those that really contain useful data for the warehouse. In the first e.dot experiments, it was shown that guiding the search through the Web by keywords extracted from the domain ontology was not precise enough to guarantee that the returned Web pages were relevant to the topic of the warehouse. Our approach for designing a filtering tool is generic and declarative. We have defined and implemented a query language, called WebQL, which enables the combination of different criteria for specifying the Web pages of interest. Those criteria allow for combining content and structure of searched documents. For example, using WebQL, it is possible for the warehouse administrator to specify that he is interested in HTML pages containing tables with titles including a given keyword. The evaluation of a WebQL query does not provide a “yes or no” answer but a matching coefficient, which is the basis for ranking the pages that are filtered. We plan to integrate WebQL within a focused crawling tool.

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

We want to populate the thematic data warehouse with data found in tables of HTML or PDF documents because tables often contain relevant and synthetic data. We explicitly separate tasks that are specific to the format of a web source, e.g., HTML, from the tasks that are independent of any source but specific to the domain. Thus, we first automatically transform the various tables into a generic XML representation called XTab. Because of the semantic heterogeneity among sources, extracting data from Web pages is insufficient to support integration. Data has to be organized in a different way with a different vocabulary, i.e. we have to find an XML representation where most of the values and tags belong to the ontology. In our approach, we want this transformation to be as automatic and flexible as possible, only driven by the ontology and the way the data has been structured in the original table. Thus, we have defined a Document Type Definition named SML (Semantic Markup Language) that can automatically be generated using the ontology and can deal with additional or incomplete information in a semantic relation, ambiguities or possible interpretation errors. This transformation has been partly implemented and experimented on real data from the e.dot project. The extracted data will be exploited by an interrogation engine named MIEL++, which is developed by the INAPG team.

5.5.3. Mappings between ontologies

Our aim is to allow a user to interrogate with a single query several data sources related to the same topic but annotated with terms from multiple ontologies. In order to be able to query the other sources, mappings between terms of the different ontologies are needed. We propose to compute these mappings once and for all and to store them in the warehouse in order to be used by the query engine. The identification of the mappings combines both syntactic and semantic comparison techniques, i.e. name-based matches and structure-level matches. This research work is being applied in the e.dot project with two ontologies: Sym’Previus and Com’Base. A first tool supporting automatic syntactic mapping techniques has been developed.

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. We have finalized our work on the XQueC compressed XML database system [30], and we have re-designed and isolated a set of XML database functionalities into our GeX prototype [37]. GeX uses an efficient path-driven partitioning strategy, allowing it to process complex structural XML queries. Going beyond the framework of a single storage model, we have started to work on a generic XQuery optimization model, capable of coping with varied and changing persistent storage modules. This capacity is crucial in order to support efficient data access, e.g., indexes and materialized views. The MS internship and beginning of the PhD thesis of A. Arion focused on this topic [56]; a generic optimizer prototype is currently under development [52]. As part of our work in this area, we have organized, in cooperation with Y.Papakonstantinou (UCSD, USA) a workshop centered on XQuery, in cooperation with ACM SIGMOD [38].

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 [27][28]. 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.

As an application of the work around KadoP, we have applied our thematic Web warehousing approach to the study of the XQuery standardization process, within the framework of the ACI “Normes et Politiques Publiques”. We have devised a conceptual model for the actor interactions taking place on the public XQuery mailing list, loaded this list into our warehouse, and performed some preliminary analysis [53].

As another application, we are participating in the INEX (Initiative for the Evaluation of XML Information Retrieval), more precisely in its Heterogeneous track. This track focuses on IR-style querying over a corpus of heterogeneous XML sources from the domain of CS bibliographic data. We have applied our thematic Web warehousing approach to construct an unified conceptual schema, and a set of mappings from individual schemas to the unified one, which solve the schema heterogeneity problem [26].

6. Contracts and Grants with Industry

6.1. Introduction

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. PICSEL2 Project

This project is the continuation of PICSEL1, which focused on designing a declarative environment for building ontology-based mediators. It aims at scaling up to the Web the mediator approach that has been implemented in PICSEL1. 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 PICSEL1 for cooperative query answering.

6.3. EC Edos Project

The goal of the EDOS Project that started in 2004 is to boost quality and productivity in software development in the field of Open Source software. Typically, most recent Linux operating systems comprise thousands of individual packages. This makes putting together such a system a task of daunting difficulty; and the short release cycles traditionally practiced in Open Source software mean that this task has to be constantly repeated. And this is the case not just for Linux, but for any large, modular software project.

In this context, the Gemo group will focus on the management of distributed information (the software releases).

Participants include academic labs INRIA (Gemo and Cristal), the U. of Paris 7, Tel-Aviv, Geneva, Zürich and Torino; and private companies Edge-IT (a Mandrakesoft subsidiary), Nuxeo, Nexedi (France) and SOT (Finland).

6.4. 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.

7. Other Grants and Activities

7.1. National Actions

In France, close links exist with groups at Orsay (databases, 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. 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).

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. Our work in this project focused on harnessing the expressive power of ActiveXML to devise robust, large-scale platforms for sharing XML resources. Several of our results concerning distributed data management with ActiveXML are part of our work in the MDP2P project [22][21]. We have performed an extensive performance study of the ActiveXML basic platform for PCs [42], and have proposed novel execution methods and optimization algorithms for computations over many peers [42]. In parallel, we have designed and built KADOP, a platform for large-scale sharing of Web resources such as semi-structured documents and Web services [27][28]. KADOP combines the power of ActiveXML and that of a Distributed Hash Table implementation (FreePastry), to perform efficient resource lookup and querying.

7.1.3. ACI Project TraLaLa

TraLaLa stands for XML Transformation Languages: logic and applications. It is funded by the *ACI Masses de Données* and has just started. 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 just started, 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. Our task in this project is to design and prototype an architecture for the construction of a semistructured data warehouse, used as a support for verifying the social scientists' hypothesis. Thus, our participation in this contract draws on our experience with the SPIN and KADOP prototype. We have taken the W3C XQuery standardization working group as an example, and have performed a preliminary study of the interactions on the public mailing list of the standard committee [53].

7.2. European Commission Financed Actions

In Europe, close links exist with Tel Aviv University (T. Milo), 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 Politecnico di Milano (S. Ceri).

Particular projects that we conduct are detailed next.

7.2.1. Procope

This year is the second year of a PAI-Procope project with the database group of Bernhard Seeger and Thomas Schwentick at Marburg University, Germany. The project is expected to last 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.

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 (A. Mendelzon and L. Libkin).

7.3.3. 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 Paris in June for a three days workshop, XIME-P, co-organized by Ioana Manolescu and Yannis Papakonstantinou (UCSD). Moreover three seniors and one Ph.d students from UCSD came to visit Gemo in Paris and stayed from one week to three months. 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

This year the following professors visited Verso:

- Michael Benedikt, Lucent Research Lab (June)
- Tova Milo, professor at the University of Tel-Aviv (till June)
- Victor Vianu, professor, UC San Diego (June to August)

The following students came for internships in the group: Zoe Abrams [Stanford U.; 3 months]; Alan Nash [UCSD; 1 month]; Antonella Poggi [U. Roma; 2 months; joint European PhD]; Gabriella Ruberg [Federal University of Rio; 6 months]; Cristina Sirangelo [U. of Calabria; 8 months].

Nicolaas Ruberg also spent an internship in the group (6 months in 2004), sent by the Bank NDES of Brazil.

8. Dissemination

8.1. Participation in Conferences

Serge Abiteboul, Sophie Cluet, Michel Scholl and Vassilis Christophides are the 2004 recipients of **SIGMOD 2004 Test of Time Award** for their paper "From Structured Documents to Novel Query Facilities" in SIGMOD 94. Sophie (ex manager of Verso) is now Director of INRIA Rocquencourt; Michel (ex manager of Verso) is now Professor at CNAM-Paris; Vassilis (ex PhD student of Verso) in now Professor at the U. of Crete.

Marie-Christine Rousset has been **General Co-chair of FQAS 2004** (Flexible Query Answering Systems) and **Tutorial chair of ECAI 2004** (European Conference on Artificial Intelligence).

Ioana Manolescu has organized and been **Co-Program chair of XIME-P**, the First International Workshop on XQuery Implementation, Experience and Perspectives in cooperation with ACM SIGMOD, 2004.

Members of the project have participated in program committees:

S. Abiteboul

- Workshop on XML High performance, W3C, 2004
- Web and database workshop, 2004
- International Workshop on Web query language, 2004
- ACM Symposium on Applied Computing (ACM SAC), Cyprus, 2004
- World Wide Web Conference 2004, WWW2004
- World Wide Web Conference 2005, WWW2005
- IEEE International Conference on Web Services (ICWS ’05), Orlando 2005

I. Manolescu

- VLDB (International Conference on Very Large Databases) 2004
- ICDE (International Conference on Data Engineering) 2004
- International SIGMOD conference 2004

- International AIMSA (Artifficial Intelligence: Methods, Systems and Architectures) Conference 2004
- International ICDE/EDBT PhD workshop 2004
- Journees de Bases de Donnees Avancees" (BDA) 2004, the French database conference
- "Simposio Brasileiro de Banco de Dados" (SBBD) 2004, the Brasilian database conference
- ACM Workshop on Web Information and Data Management (WIDM) 2004

C. Reynaud

- International Conference on Knowledge Engineering and Management (EKAW), 2004.
- 3rd International Semantic Web Conference (ISWC), 2004.
- Workshop on Knowledge Management and the Semantic Web, International Conference on Knowledge Engineering and Management, 2004.
- Journée Web Sémantique Médical, Rouen, 2004.
- Atelier Modélisation des connaissances, EGC, Janvier 2005.
- Ingénierie des Connaissances, Mai-Juin 2005.

M-C. Rousset

- KR 2004 (International Conference on Principles of Knowledge Representation and Reasoning).
- SIGMOD 2004 (ACM International Conference on Management of Data).
- Congrés Francophone de Reconnaissances des Formes et Intelligence Artificielle (RFIA), 2004

L. Segoufin

- ACM International Symposium on Principles of Database Systems (PODS), Paris, France, June 2004

8.2. Invited Presentations

Serge Abiteboul has been invited to present *Distributed information management with XML and Web services* at ETAPS04, the European Joint Conferences on Theory and Practice of Software, in Barcelona. he has also been invited to present *ActiveXML and Active Query Answers* at FQAS 2004, International Conference on Flexible Query Answer, in Lyon. He has also been invited to present *ActiveXML, Security and Access Control* at SBBD 2004, XIX Brazilian Symposium on Databases, in Brasilia.

Ioana Manolescu presented an invited tutorial on *XML Query Processing: Storage and Query Model Interplay* at the International EDBT summer school in databases, and at the SBBD conference, 2004. She also presented an invited tutorial on *XML query processing and distributed applications* at the DRUIDE 2004 summer school, Domaine de Port-aux-Rocs, France, May 2004. She presented also this second tutorial at the AlgoDis summer school in Porquerolles, France, September 2004.

Chantal Reynaud has been invited to present *Semantic integration of Heterogeneous Data* at Topical session Workshop, IFIP World Congress (Toulouse, Août 2004).

Marie-Christine Rousset presented an invited tutorial on *Médiation sémantique dans un réseau Pair-à-Pair* at Summer School DistRibUtIon de Données à grande Echelle (DRUIDE), Le Croisic, France, May 2004, She also had an invited presentation on *Small Can Be Beautiful in the Semantic Web* at the International Web Semantic Conference (ISWC 2004), Hiroshima, Jaan, November 2004.

8.3. Scientific Animations

8.3.1. Editors

B. Amann

- Revue d'Intelligence Artificielle (RIA)
- Revue Ingénierie des Systèmes d'Information, Numéro Spécial "Les services web, théories et applications" (2005)

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)

9. Bibliography

Major publications by the team in recent years

[1] S. ABITEBOUL, A. BONIFATI, G. COBENA, I. MANOLESCU, T. MILO. *Dynamic XML documents with distribution and replicatio*, in "Proc. of the ACM SIGMOD Conf.", 2003.

[2] S. ABITEBOUL, R. HULL, V. VIANU. *Foundations of Databases*, Addison-Wesley, Reading-Massachusetts, 1995.

[3] S. ABITEBOUL, V. VIANU. *Computing with First-Order Logic*, in "Journal of Computer and System Sciences", 1994.

[4] V. CHRISTOPHIDES, S. ABITEBOUL, S. CLUET, M. SCHOLL. *From structured documents to novel query facilities*, ACML SIGMOD, Test of Time Award, 1994, p. 313–324.

[5] D. FLORESCU, A. LEVY, I. MANOLESCU, D. SUCIU. *Query optimization in the presence of limited access patterns*, in "Proc. of ACM SIGMOD Conf. on Management of Data", 1999, p. 311–322.

[6] F. GOASDOUE, C. REYNAUD. *Modeling Information Sources for Information Integration*, in "Knowledge Acquisition, Modeling and Management", 1999, p. 121-138.

[7] F. G. V. LATTES, M.-C. ROUSSET. *The Use of CARIN Language and Algorithms for Information Integration: The PICSEL System*, in "International Journal of Cooperative Information Systems", vol. 9, n° 4, 2000, p. 383-401.

[8] A. Y. LEVY, M.-C. ROUSSET. *CARIN: A Representation Language Combining Horn Rules and Description Logics*, in "European Conference on Artificial Intelligence", 1996, p. 323-327.

[9] T. MILO, S. ABITEBOUL, B. AMANN, O. BENJELLOUN, F. NGOC. *Exchanging Intensional XML Data*, in "Proceedings of the ACM SIGMOD International Conference on Management of Data", 2003.

[10] L. SEGOUFIN, V. VIANU. *Validating Streaming XML Documents*, in "Symposium on Principles of Database Systems", 2002, p. 53-64.

Books and Monographs

[11] S. ABITEBOUL, O. BENJELLOUN, I. MANOLESCU, T. MILO, R. WEBER. *Active XML: A Data-Centric Perspective on Web Services*, n° 3-540-40676-X, Book chapter in Web Dynamics, Levene, Mark; Poulovassilis, Alexandra (Eds.), 2004.

Doctoral dissertations and Habilitation theses

[12] O. BENJELLOUN. *Active XML: A data-centric perspective on Web services*, Ph. D. Thesis, September 2004.

[13] A. TERMIER. *Extraction d’arbres fréquents dans un corpus hétérogène de données semi-structurées*, Ph. D. Thesis, April 2004.

Articles in referred journals and book chapters

[14] 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.

[15] F. GOASDOUÉ, M.-C. ROUSSET. *Answering Queries using Views: a KRDB Perspective for the Semantic Web*, in "ACM Journal - Transactions on Internet Technology (TOIT)", vol. 4, n^o 3, 2004, 255,288.

[16] B. LUDASCHER, Z. IVES, I. MANOLESCU. *Reminiscences of Influential Papers*, in "SIGMOD Record", vol. 33, n^o 3, September 2004.

[17] A.-D. MEZAOUR. *Recherche ciblée de documents sur le web*, in "Revue des Nouvelles Technologies de l", vol. 2, January 2004, p. 491-502.

[18] S. DE AMO, N. BIDOIT, L. SEGOUFIN. *Order independent temporal properties*, in "Journal of Logic and Computation", vol. 14, n^o 2, 2004, p. 277–298.

Publications in Conferences and Workshops

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

[20] S. ABITEBOUL. *Distributed information management with XML and Web services*, in "European Joint Conferences on Theory and Practice of Software (ETAPS), in proc. FASE", Springer, LNCS, 2004.

[21] S. ABITEBOUL, B. ALEXE, O. BENJELLOUN, B. CAUTIS, I. FUNDULAKI, T. MILO, A. SAHUGUET. *An Electronic Patient Record on Steroids: Distributed, Peer-to-Peer, Secure and Privacy-conscious*, in "Demo; Intern. Conf. on Very Large Data Bases (VLDB)", 2004.

[22] S. ABITEBOUL, O. BENJELLOUN, B. CAUTIS, I. MANOLESCU, T. MILO, N. PREDA. *Lazy Query Evaluation for Active XML*, in "ACM SIGMOD Conference on Management of Data", June 2004.

[23] S. ABITEBOUL, O. BENJELLOUN, B. CAUTIS, T. MILO. *Active XML, Security and Access Control*, in "Simpósio Brasileiro de Banco de Dados (SBBD)", 2004.

[24] S. ABITEBOUL, O. BENJELLOUN, T. MILO. *Active XML and active query answers*, in "International Conference on Flexible Query Answer (FQAS)", L. N. IN COMPUTER SCIENCE (editor)., vol. 3055, Springer, 2004, p. 17–27.

[25] S. ABITEBOUL, O. BENJELLOUN, T. MILO. *Positive Active XML*, in "ACM SIGMOD-SIGACT- SIGART Symposium on Principles of Database Systems", June 2004.

[26] S. ABITEBOUL, I. MANOLESCU, B. NGUYEN, N. PREDA. *A Test Platform for the INEX Heterogeneous Track*, in "INEX 2004 Workshop", informal proceedings, 2004.

[27] S. ABITEBOUL, I. MANOLESCU, N. PREDA. *Constructing and querying peer-to-peer warehouses of XML resources*, in "Second International Workshop on Semantic Web and Databases (SWDB)", V. T. CHRIS

BUSSLER (editor)., Springer-Verlag, 2004.

[28] S. ABITEBOUL, I. MANOLESCU, N. PREDA. *Peer-to-peer warehousing of XML resources*, in "Bases de Donnees Avancees", 2004.

[29] P. ADJIMAN, P. CHATALIC, F. GOASDOUÉ, M.-C. ROUSSET, L. SIMON. *Distributed Reasoning in a Peer-to-peer Setting*, in "European Conference on Artificial Intelligence", 2004.

[30] A. ARION, A. BONIFATI, G. COSTA, I. MANOLESCU, A. PUGLIESE. *Efficient Query Evaluation over Compressed XML Data*, in "International Conference on Extending Database Technologies (EDBT)", March 2004.

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

[32] D. L. BERRE, L. SIMON. *Fifty-five solvers in vancouver: The sat 2004 competition*, in "Seventh International Conference on Theory and Applications of Satisfiability Testing (SAT2004)", LNAI, to appear 2004.

[33] M. BRAMBILLA, S. CERI, S. COMAI, M. DARIO, P. FRATERNALI, I. MANOLESCU. *Declarative Specification of Web Applications exploiting Web Services and Workflows*, in "ACM SIGMOD Conference on Management of Data", 2004.

[34] P. BUCHE, J. DIBIE-BARTHÉLEMY, O. HAEMMERLÉ, M. HOUHOU. *Towards flexible querying of XML imprecise data in a dataware house opened on the Web*, in "Flexible Query Answering Systems (FQAS)", Springer Verlag, june 2004.

[35] J. DIBIE-BARTHÉLEMY, O. HAEMMERLÉ, E. SALVAT. *Validation de graphes conceptuels*, in "Extraction et Gestion de Connaissances (EGC)", Cépaduès, Janvier 2004, p. 135–146.

[36] H. FOLCH, B. HABERT, M. JARDINO, N. PERNELLE, M.-C. ROUSSET, A. TERMIER. *Highlighting latent structure in documents*, in "International Conference on Language Resources and Evaluation (LREC)", vol. 4, 2004, 1131,1334.

[37] I. MANOLESCU, A. ARION, A. BONIFATI, A. PUGLIESE. *Path Sequence Based XML Query Processing*, in "Bases de Données Avancées (BDA)", october 2004.

[38] I. MANOLESCU, Y. PAPAKONSTANTINOU. *Report on the first XIME-P workshop*, in "SIGMOD Record", vol. 33, September 2004.

[39] A. MUSCHOLL, T. SCHWENTICK, L. SEGOUFIN. *Active Context-Free Games*, in "STACS", 2004.

[40] C. REYNAUD. *Building scalable mediator systems*, in "Topical Day in Semantic Integration of Heterogeneous Data, IFIP World Computer Congress", 2004.

[41] M.-C. ROUSSET. *Small Can Be Beautiful in the Semantic Web*, in "Third International Semantic Web Conference", F. V. H. S. MCILRAITH (editor)., vol. 3298, Springer (LNCS), 2004, p. 6–16.

[42] N. RUBERG, G. RUBERG, I. MANOLESCU. *Towards cost-based optimization for data-intensive Web service computations*, in "SBBD (Simposio Brasileiro de Bancos do Dados)", October 2004.

[43] B. SAFAR, H. KEFI, C. REYNAUD. *OntoRefiner, a user query refinement interface usable for Semantic Web Portals*, in "Applications of Semantic Web technologies to web communities, Workshop ECAI", 2004.

[44] F. SAIS, H. GAGLIARDI, O. HAEMMERLÉ, N. PERNELLE. *Enrichissement sémantique de documents XML représentant des tableaux*, in "Extraction et Gestion de Connaissances (EGC), to appear", 2005.

[45] L. SIMON, D. L. BERRE, M. NARIZZANO, A. TACCHELLA. *The second qbf solvers comparative evaluation.*, in "Seventh International Conference on Theory and Applications of Satisfiability Testing (SAT2004)", to appear 2004.

[46] A. TERMIER, M.-C. ROUSSET, M. SEBAG. *DRYADE: a new approach for discovering closed frequent trees in heterogeneous tree databases.*, in "International Conference on Data Mining (ICDM 2004)", 2004, p. 543–546.

[47] R. THOMOPOULOS, P. BUCHE, O. HAEMMERLÉ. *Sous-ensembles flous définis sur une ontologie*, in "Extraction et Gestion des Connaissances (EGC)", Cépaduès, Janvier 2004, p. 147–158.

[48] V. V., H. SOLDANO, T. LAMADON. *Treillis de Galois Alpha*, in "Conférence d'Apprentissage (CAP'04)", Presses Universitaires de Grenoble, June 2004, p. 175-190.

[49] VENTOS, V., H. SOLDANO, T. LAMADON. *Alpha Galois Lattices*, in "International Conference on Data Mining (ICDM'04)", IEEE (editor)., november 2004, p. 555-558.

[50] VENTOS, V., H. SOLDANO. *Alpha Galois Lattices: an overview*, in "International Conference in Formal Concept Analysis (ICFCA'05)", LNCS (editor)., Springer, february 2005, to appear.

Internal Reports

[51] S. ABITEBOUL, O. BENJELLOUN, T. MILO. *The Active XML project: an overview*, Technical report, Gemo, 10 2004.

[52] A. ARION, V. BENZAKEN, I. MANOLESCU. *ULoad: Choosing the Right Storage for your XML Application*, Technical report, Gemo, November 2004.

[53] F.-X. DUDOUET, I. MANOLESCU, B. NGUYEN, P. SENELLART. *Sociological Analysis of the W3C Standardization Process: XML Warehouse meets Sociology*, Technical report, Gemo, November 2004.

[54] G. GOTTLOB, C. KOCH, R. PICHLER, L. SEGOUFIN. *The Parallel Complexity of XML Typing and XPath Query Evaluation*, Technical report, Gemo, 2005.

Miscellaneous

[55] P. ADJIMAN, P. CHATALIC, F. GOASDOUÉ, M.-C. ROUSSET, L. SIMON. *Raisonnement distribué dans un environnement de type Pair-à-Pair*, 2004, Journées Nationales sur la résolution Pratique de Problèmes NP-Complets.

[56] A. ARION, V. BENZAKEN, I. MANOLESCU. *Vers un optimiseur générique des requêtes XQuery*, Technical report, Universite Paris SUD, 2004.

[57] F. GOASDOUÉ, M.-C. ROUSSET. *Intégration d'Information par Médiation*, 2004, Plein Sud - Spécial Recherche.

[58] I. MANOLESCU. *XML Query Processing: Storage and Query Model Interplay*, September 2004, EDBT Summer School.

[59] A. MUSCHOLL, T. SCHWENTICK, L. SEGOUFIN. *Active Context-Free Games*, 2004, invited.

[60] N. PREDA. *Construction d'un entrepot de donnees du Web en pair a pair*, september 2004.

[61] C. REYNAUD, B. SAFAR, H. GAGLIARDI. *Une expérience de représentation d une ontologie dans le médiateur PICSEL*, 2004.