INRIA

Project-Team gemo

Management of Data and Knowledge Distributed Over the Web

Futurs

THEME SYM

Activity Report

2006

Table of contents

1. Team

Managers

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

Administrative Assistant

Stéphanie Meunier [ITA]

INRIA personnel

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

University personnel

Philippe Chatalic [Assistant Professor, Univ. Paris 11]
Philippe Dague [Professor, Univ. Paris 11, HdR]
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, HdR]
Brigitte Safar [Assistant Professor, Univ. Paris 11]
Laurent Simon [Assistant Professor, Univ. Paris 11]
Véronique Ventos [Assistant Professor,Univ. Paris 11]

Scientific Advisors

Cluet Sophie [Department Director, MESR]
Dan Vodislav [Assistant Professor, CNAM]

Invited researchers and visitors

Daniele Braga [Post Doc, Politecnico di Milano, 1 month]
Tarik Mellati [Post Doc, until August, scientific advisor since then]
Neoklis Polyzotis [Assistant Professor, U.C. Santa Cruz, 1 month]
Cristina Sirangello [Post Doc fellowship]
Michalis Vazirgiannis [Marie-Curie fellowship, 6 months]
Victor Vianu [Professor, U.C. San Diego, 3 months]

Engineers

Gabriel Vasile
Evaldas Taroza [Since November]

Ph.D students

Nada Abdallah [Allocataire MENRT, Paris 11]
Philippe Adjiman [Allocataire MENRT, Paris 11, until September]
Andrei Arion [Allocataire MENRT, Paris 11]
François-Elie Calvier [Grant CNRS]
Bogdan Cautis [Allocataire MENRT, Paris 11]
Claire David [ENS Cachan]
Hassen Kefi [ATER, Paris 11, until April]
Charaf Laissoub [Contrat ANR]
Yingmin Li [European contract]
Bogdan Marinoiu [Grant BDI CNRS, Paris 11]
Antonella Poggi [PhD in cotutelle between U. di Roma and Paris 11]
Nicoleta Preda [Allocataire MENRT, Paris 11]
Cedric Pruski [PhD in cotutelle between Luxembourg U. and Paris 11]
Fatiha Sais [Contrat FTRD]
Mathias Samuelides [ENS Cachan]
Pierre Senellart [ENS Ulm]

Spyros Zoupanos [Cordi, since October]

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*, *ontology*, *peer-to-peer (p2p)*, *query optimization*, *query language*, *semantic integration*, *semi-structured data*.

See http://gemo.futurs.inria.fr/

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).

Some prospective work is presented in [12], [11]. The goal is to enable non-experts, such as scientists, to build *content sharing communities* in a true database fashion: declaratively. The proposed infrastructure is called a *data ring*.

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. For instance, in the setting of the EDOS EC Project, we are developing some software for the P2P management the data and metadata of Mandriva Linux distribution.

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.

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.

XTAB2SML: an automatic ontology-based tool to enrich tables semantically

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.

GUNSAT: a greedy local search algorithm for propositional unsatisfiability testing

5. New Results

5.1. Theoretical foundations

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

Participants: Serge Abiteboul, Bogdan Cautis, Luc Segoufin, Pierre Senellart, Victor Vianu.

One of the reasons for the success of the relational data model was probably its clean theoretical foundations. Obtaining such a clean foundation for the semistructured data model and XML is still an on-going research task. It make use of the tree structure of XML data and the connection between Monadic-Second-order (MSO) logic and automata on trees. A new challenge is the study of XML when used in the dynamic environment of the Web.

In [18] we have studied a sequential family of tree automata, which has the same expressive power than Transitive-Closure logic. Our results exhibit a hierarchy in term of expressive power and show its strict inclusion into MSO.

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 database context, where the power in term of computability is limited, the difficulty is to handle data values ranging in a possible infinite domain. In the offline setting we have exhibited in [17], [16], [33] relevant decidable logics over trees that could be used for static analysis. In [16] we show how to use those logics in the XML context for deciding XML-schema validation and XPath query inclusion.

It is also desirable that published or exchanged XML documents have specific restrictions on how they can be updated or transformed. We introduce in [19] this paper a family of constraints for describing how an XML document can evolve, namely XML update constraints. For these constrains, we studied the implication problem, giving algorithms and complexity results for constraints of varying expressive power.

The common assumption about digital signatures is that they disallow any kind of modification on signed data. However, a more flexible approach is often needed and has been advocated lately, one in which some restricted modifications may still occur, without invalidating the data. In [9] we consider digital signatures of insert-only collections, in which element insertions are accepted but no removals should occur. Newly inserted elements do not have to be signed or known by the initial signer. We propose two techniques: one which transposes the insert-only problem into a delete-only one (which is already solved), and another technique based on zero-knowledge proofs. We also give performance measures and discuss applications.

We present in [13] a new model for representing probabilistic information in a semi-structured (XML) database, based on the use of probabilistic event variables. This work is motivated by the need of keeping track of both confidence and lineage of the information stored in a semi-structured warehouse. For instance, the modules of a (Hidden Web) content warehouse may derive information concerning the semantics of discovered Web services that is by nature not certain. Our model, namely the fuzzy tree model, supports both querying (tree pattern queries with join) and updating (transactions containing an arbitrary set of insertions and deletions) over probabilistic tree data. We are currently analyzing its expressive power and working on implementation issues.

In the online setting we have considered streaming XML data with limited memory resources. In this context, we considered in [34] the DTD validation problem: checking whether a XML document conforms a DTD.

5.2. Information Extraction

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

Participants: Chantal Reynaud, Brigitte Safar, Véronique Ventos.

5.2.1. *Information extraction from semi-structured data*

In the continuation of our work on integration in PICSEL 3, we have proposed an approach to build wrappers automatically. These wrappers are able to extract data from a XML source in order to populate a RDF data warehouse. The approach exploits mappings between (1) an ontology represented in RDF-S and (2) the source to be integrated and its DTD. The approach has been developed in the picsel 3 framework allowing a uniform access to both data stored in distributed and autonomous sources and data locally stored in a data warehouse. It is generic, applicable whatever the XML source of the application domain is.

5.2.2. *Information extraction and Galois Lattices*

Galois lattices (or concept lattices) are exhaustive representations of the concepts embedded in a data set since they explicitly represent every subset of instances distinguishable according to the language used to describe clusters. In previous work, we defined Alpha Galois lattices which are coarsest concept lattices where the coarseness is linked to a partition of the universe of instances into basic classes. We extend this study with theoretical and experimental results about the incremental construction of Alpha Galois lattices. First we define a general Merge operator of concept lattices then we show that when applying such a Merge operator on two alpha lattices, we obtain a alpha lattice linked to the union of the local universes and partitions. This leads to a efficient class-incremental construction of a Alpha lattice by merging the lattices corresponding to each basic class. A strategy for a general update of the Alpha lattice when new basic classes or new instances are provided has been defined.

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: Nada Abdallah, Philippe Adjiman, Philippe Chatalic, François Goasdoué, Gia Hien Nguyen, Marie-Christine Rousset, Laurent Simon.

The SomeWhere [3] platform, allowing to perform distributed reasonning tasks in a peer-to-peer architectures, has been consolidated. It is a building block of the MEDIAD project with France Telecom R&D, as well as one of the components being integrated in the platform to be produced by the WebContent project, whose topic concern multi-scaled distributed mediation. The vision of the Semantic Web promoted in SomeWhere has obtained a high visibility through several invited talks (e.g., [30]). The code of SomeWhere has been patented and several new extensions have been investigated recently.

We have shown that RDF(S) may be used an alternative data model to formalize distributed sets of ontologies in SomeWhere. This has been made possible thanks to a method able to propositionalize peer theories expressed in such a way, as well as to translate back answers produced by SomeWhere into first order logic [4]. The resulting framework is called SomeRDFS. Current work further explore the possibility to use additional constraints in order to be able to perform query rewriting in term of views.

Because in peer-to-peer architectures each peer has the same abilities one cannot prevent the global theory of a SomeWhere network (i.e. to the union of all peer theories) to be inconsistent. We have designed an method able do detect all such inconsistencies, as new mappings are added between consistent theories. Furthermore, we have proposed a method for still performing well founded reasonning, despite the presence of such inconsistencies [20]. Further work investigates how the notion of thrust may be used to for preferential reasonning in such inconsistent theories.

Enventually, we study in the WebContent project, in what respect the SomeWhere platform may be integrated with other P2P lookup or indexing platforms (KadoP, PINS).

5.4. Thematic Web Warehousing

Keywords: *Warehouse*, *ontology alignment techniques*, *thematic information*.

Participants: Hélène Gagliardi, Hassen Kefi, Nathalie Pernelle, Chantal Reynaud, Marie-Christine Rousset, Fatiha Sais, Brigitte Safar.

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

We have developed a method and a tool called Xtab2SML for an automatic semantic enrichment of data tables that are extracted from Web documents by means of tags (relations and terms) coming from a domain ontology. This work has been initiated in the setting of the e.dot project. As we want this tool to be completely automatic, the interpretation of each lines of a data table cannot be precise and complete for each kind of table. Thus, we have defined an adaptative method which allows keeping unidentified columns of the table in the recognized instances of the semantic relations of the ontology. When the semantically enriched data tables are queried, these additional values can be showed to the user in order to facilitate its interpretation of the answers. Besides, the existence of a semantic relation between values of the same lines is kept even if the system is not able to identify any semantic relation. The advantages of these two kind of flexibility have been showed in [21]. We are now working on the reference reconciliation problem. It consists in deciding whether different identifiers refer to the same data, i.e. correspond to the same world entity (the same hotel, the same person, ...). This work is done in the setting of the Picsel3 project.

5.4.2. Mapping between ontologies

This work has been initiated in the setting of the e.dot project. 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. A prototype, TaxoMap, finds mappings or suggests indicators to help users find mappings. Experiment results in the field of predictive microbiology have been obtained [22]. We have continued the development of the TaxoMap system, focusing on original structural techniques suitable with a dissymmetry in the structure of the mapped taxonomies, one taxonomy being deep whereas the other one is flat. We have worked on two techniques published in [27] and in [29]. The first one is performed on the taxonomy which is supposed to be the deepest. We exploit the location of the mapping candidates, considering that their proximity in the graph of the taxonomy is a semantic proximity. Then we use WordNet, exploiting its structure and its semantic relations. WordNet is not considered here simply as a source of synonyms, hypernyms or hyponyms. It provides a structural support exploited to detect relations between concepts. We propose an efficient strategy which does not require the computation of many similarity measurements. We pursued also our experiments on test taxonomies which are not structurally dissymmetric and which cover a large domain [37]. The application conditions of the techniques were not achieved but our objective was to sketch some ideas to do improvements and to widen the scope of the approach. This work is the core of a PhD which has been defended in March 2006.

Other research works on mappings are done in the setting of Picsel3. We studied the integration of a new XML source to a data warehouse thanks to an ontology. The generation of mappings is viewed as a hybrid process based on both schemas of the sources and data. This work has been published in [28].

5.5. XML query optimization

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

Participants: Serge Abiteboul, Andrei Arion, Asmae Fahoum, Ioana Manolescu, Thai-Tho Nguyen.

5.5.1. *XML Access modules*

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 propose a formalism for describing a wide family of storage structures, encompassing existing storage and indexing schemes - "The XML Access Modules" (XAM). Current XML materialized view proposals are based on tree patterns, since query navigation is conceptually close to such patterns. However, the existing algorithms for extracting tree patterns from XQuery do not detect patterns across nested query blocks. Thus, complex, useful tree pattern views may be missed by the rewriting algorithm. In [14] we present a novel tree pattern extraction algorithm from XQuery queries that is based on the XAMs and is able to identify larger patterns than previous methods. We fully characterize the complexity of XAM containment and rewriting under the constraints expressed in the structural summary, whose enhanced form also entails integrity constraints, in [23].

5.5.2. *Materialized view management*

We also continue our work on ULoad - a prototype of a materialized view management system which relies on the XAM formalism to rewrite XQuery queries. Further work has been done on extending Uload with new features: Asmae Fahoum implemented the building blocks for XQuery updates and Thai Tho Nguyen devised and implemented a model for cardinality and selectivity cost estimation that uses the XSketch synopses.

5.5.3. *XML compression*

In the XQueC (XQuery Processor and Compressor) project we address the problem of embedding compression into XML databases without degrading query performance. Since the setting is rather different from relational databases, the choice of compression granularity and compression algorithms must be revisited. Query execution in the compressed domain must also be rethought in the framework of XML query processing, due to the richer structure of XML data. In [5] we present how XQueC covers a wide set of XQuery queries in the compressed domain relaying on a workload-based cost model to perform the choices of the compression granules and of their corresponding compression algorithms.

5.5.4. *Self tuning*

We started some collaborative work with UCSC and U.Tel Aviv on a framework for Continuous On-Line Tuning [32], a novel self-tuning framework that continuously monitors the incoming queries and adjusts the system configuration in order to maximize query performance. The key idea behind Colt is to gather performance statistics at different levels of detail and to carefully allocate profiling resources to the most promising candidate configurations. Moreover, Colt uses effective heuristics to self-regulate its own performance, lowering its overhead when the system is well tuned and being more aggressive when the workload shifts and it becomes necessary to re-tune the system. We considered the design of the generic Colt system, and its specialization to the important problem of selecting an effective set of indices for a relational query load. We developed an implementation of the proposed framework in the PostgreSQL database system and evaluated its performance experimentally. Our results validate the effectiveness of Colt in self-tuning a relational database, demonstrating its ability to modify the system configuration in response to changes in the query load. Moreover, Colt achieves performance improvements that are comparable to more expensive off-line techniques, thus verifying the potential of the on-line approach in the design of self-tuning systems.

5.5.5. *Other works*

XML path summaries are useful tools for access path selection. We report in [15] the advantages of using XML path summaries in conjunction with a path-partitioned store and we present an efficient method for complex tree reconstruction, with much lower memory needs than existing alternatives.

Another research direction targets a better understanding of the drawbacks and advantages of XPath/XQuery implementation strategies. An important step towards such a better understanding is the development of a microbenchmark [24], that helps to reveal detailed information regarding the performance of implementations at isolated language features.

5.6. XML Warehousing in P2P

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

Participants: Serge Abiteboul, Ioana Manolescu, Nicoleta Preda.

We have worked on the optimization of KADOP, a peer-to-peer platform for building and managing warehouses of Web resources. 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.

5.7. Monitoring and Web services

Keywords: *Web services*, *diagnosability*, *formal models*, *model-based diagnosis*, *monitoring*, *repair*, *repairability*, *self-healing.*

Participants: Serge Abiteboul, Philippe Dague, Bogdan Marinoiu, Tarak Melliti, Yingmin Li.

5.7.1. *Error diagnosis and self-healing*

This work, that began at the end of 2005, is carried out in the framework of the European project WS-DIAMOND. It is well-known that self-healing software is one of the challenges for IST research. This project aims to take a step in this direction by developing a framework for self-healing Web Services. The goal is to produce:

- an operational framework for self-healing service execution of conversationally complex Web services, where monitoring, detection and diagnosis of anomalous situations, due to functional (in particular semantic) or non-functional errors (e.g., Quality of Service), is carried on and repair/reconfiguration is performed, thus guaranteeing reliability and availability of Web services;
- a methodology and tools for service design that guarantee effective and efficient diagnosability/repairability during execution;
- demonstration of these results on real applications.

Our main involvement in this project is about model-based diagnosis of cooperative Web services, i.e. apply to P2P distributed software systems the techniques developed in Artificial Intelligence and successfully applied to engineered centralized hardware systems. Our two other contributions concern formal models for Web services, as the method rests entirely on the existence of adequate behavioral models to which actual observations are compared, and study of diagnosability at the design stage, which is the common trend to diagnosis activities in all branches of industry.

During this first year, the following work has been achieved:

- Developing an observation and data log platform for basic Web services.

An extension of the Web service deployment specification (WSDD file) is defined, allowing the developer to specify for each operation what are the informations to log and the privacy police of their accessibility. The standard AXIS deployment platform is enriched by an observation handler generator and an information Web service generator. Each time a basic WS is invoked, its associated information WS is invoked too and records in databases (via an interface with MySQL) all its inputs, outputs and error messages specified in the WSDD extension, with the given privacy policy. This can be applied to the information WS itself, which is thus self-observed. All these extensions and log capabilities were implemented in Java. The logged information will be used by the diagnosis task to identify the primary cause(s) of a detected symptom.

- Modeling BPEL Web services for diagnosis.
  A method to generate automatically a diagnosis model, in the form of data dependency relations, for orchestrated complex Web services has been developed. BPEL (Business Process Execution Language) basic and structured activities are first modeled with Petri nets, places being used to represent data and transitions to represent activities. For that, control places, in charge of transmitting activation, are added to data places (in particular an input and output activation places) and reading arcs (along which tokens are not propagated) are added to normal arcs. Operational dependency between the transitions executions is thus captured. In order to capture data dependency (which is essential for diagnosis of semantic faults), each transition of the Petri net is enriched with a set of basic data dependency relations expressing that an input is just forwarded to output, or that an output is created by the operation, or that an output is elaborated from one or several inputs. In order to aggregate such enriched Petri nets, composition rules of these basic relations are defined, for different modes (sequential, alternative, hierarchical through data structures). Based on these rules, an algorithm is designed that builds the data dependency model of an orchestrated BPEL service from the analysis of its BPEL code and the models exposed by the private services it invokes. Such a model can then be used by a diagnosis algorithm.

This year our work has been published in [25], [26], [36]. Further work will include: implementing the algorithm of construction of the diagnosis model of a BPEL Web service; extending it to the case of choreographed Web services; designing and implementing a decentralized diagnostic architecture in the form of diagnostic Web services: both supervised and completely distributed approaches will be studied; studying in the same way repair or reconfiguration task; defining diagnosability and recoverability criteria and a methodology of designing Web services applications that respect these criteria.

5.7.2. *P2P Monitoring*

We also started some work on the monitoring P2P systems. Alerters placed on peers, are in charge of the surveillance of particular events they monitor (e.g., web service calls, database updates, web pages changes). They produce streams of (Active)XML data. Our system implements an algebra over data streams. We worked on modules to filter or combine streams. The results may be published as Web pages, RSS streams, continuous Web services or emails. An application based on RSS feeds has been developed.

6. Contracts and Grants with Industry

6.1. Industrial contracts

Gemo has had technical meetings in 2006 with many industrial partners, in particular France Telecom R&D, Xyleme and Mandrakesoft, as well as national organizations, in particular, Institut National de Recherche en Agronomie.

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

EDOS (http://www.edos-project.org/) is a research project funded by the European Commission as a STREP project under the IST activities of the 6th Framework Programme. The project involves universities (Paris 7, Tel Aviv, Geneva), INRIA (Gemo and Cristal teams) and private companies (Caixa Magica, Nexedi, Nuxeo, Edge-IT and CSP Torino). It is centered around the software management and more particularly, of Mandriva Linux distribution.

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.

The goal is to build a P2P distribution system, that improves the classical approach based on hierarchies of mirrors, by providing a better sharing of resources. The system is based on the idea of exchanging XML data in a P2P environment, in our case metadata about the software modules to be distributed. Metadata describes the name, version, size, composition, etc of the software modules, but also location of replicas in the P2P network.

We defined the P2P system architecture, based on three categories of actors: Publisher (that introduces new content in the system), Mirrors (trusted peers) and Clients (end users). Peers are organized in two sub-networks: the indexing network, composed of trusted peers (Publisher and Mirrors), storing the distributed index on metadata, and the distribution network, composed of all the peers, storing content replicas. We also defined a general API for the distribution system, describing its basic functionalities: publishing of new content, metadata indexing and querying, subscription to thematic distribution channels and event notification, download in flash-crowd (one source, many requests at the same time) and off-peak (many sources, content updates) situations, etc.

The main effort in the last period of time was devoted to the implementation of the distribution system. The first version consists of the system provides a set of JSP web applications (one for each peer type), built on top of a Java implementation of EDOS distribution API. The distribution system uses four main software components: (i) KadoP for distributed indexing and querying of XML metadata, (ii) ActiveXML for metadata representation and storage, (iii) IDiP (University of Tel Aviv) for content dissemination in flash-crowd situations, based on clustering of content and multi-cast, and (iv) BitTorrent for efficient download of files.

Next steps concern extensive testing, inclusion of security issues (key-based authentication, firewall/NAT handling), improvements in the system usability. In the EDOS context, several necessary extensions of KadoP were revealed; some of them were already implemented (replication in query processing, management of very long postings), other are to be realized in the next period of time (mass publication optimization, extensions in the query language).

6.5. RNTL Project WebContent

Gemo (S. Abiteboul, M.-C. Rousset) and the CEA (C. Fluhr) have been at the initiative of the WebContent project (http://www.webcontent.fr/). WebContent groups together 19 partners from academic research and industry (EADS, Thalès, Exalead, Xyleme, NewPhenix). WebContent has recently started (kickoff meeting has been hold on April 24th 2006). The goal of WebContent is to build a flexible and generic platform for content management and to integrate Semantic Web technologies in order to show their effectiveness on real applications with strong economic or societal stakes. Gemo is strongly involved into the working groups Lot0 (Management of the project), Lot3 (Semantic enrichment of ontologies and documents) and Lot5 (Management and interrogation of large volumes of contents in P2P).

6.6. WS-DIAMOND EU project

WS-DIAMOND ("Web Services - DIAgnosability, MONitoring and Diagnosis") is a FP6 European project (FET Open Strep) which started on Sept. 1st 2005 and will last until Feb. 29th, 2008. EU funding for University Paris-Sud is 188 kEuros. The project is coordinated by the University of Turin, and involves the Polytechnic University of Milan, the Vrije University of Amsterdam, the University of Vienna, the University of Klagenfurt, and from France the LAAS-CNRS, the University of Rennes 1, and the University of Paris-Sud. Participants from Gemo are Philippe Dague (site leader for U. Paris-Sud), Tarak Melliti (post-doc from Oct. 1st 2005 to Aug. 31th 2006, assistant professor at U. of Evry from Sep. 1st 2006), and Yingmin Li (master internship from April 1st 2006 to Sept. 30th 2006, Ph.D. student from Oct. 2006).

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), the LIRMM of the University of Montpellier (M. Chein, M-L. Mugnier), and the the LI of the University of Tours (G. Venturini).

7.1.1. ACI Project ACI-MDP2P

MDP2P (http://www.sciences.univ-nantes.fr/lina/ATLAS/MDP2P/, Massive Data in Peer-to-Peer)) is a joint research project funded by the ACI Masses de Données of the French Ministry of Research. The project completed in 2006. The main areas covered were data replication and large-scale load balancing, large-scale indexing and retrieval of text and multimedia documents, and massive data management in P2P systems. The contribution of Gemo was primarily the KadoP system and optimization techniques around it.

7.1.2. ACI Project TraLaLa

TraLaLa stands for XML Transformation Languages: logic and applications. It is funded by the *ACI (Action Concertée Incitative) Masses de Données*, has started in September 2004 and will end in September 2007. 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 home page of the project can be found at: http://www.cduce.org/tralala.html.

7.1.3. *ACI Normes Pratiques et Régulations des Politiques Publiques*

This ACI started in 2005 and is projected to last three years. This ACI is a collaboration between Benjamin Nguyen (University of Versailles), and François-Xavier Dudouet (CNRS, Laboratoire IRISES). During this year we have analyzed in detail standardization mailing lists (from the W3C) and conducted a sociological analysis. This analysis is based on the use of XQuery based tools and models. We are in the process of joining this ACI with the ANR-JCJC "WebStand" who has some concurrent goals. Our research results have been used by sociologists in a publication in the Revue François de Science Politique.

7.1.4. *ANR JCJC WebStand*

The objective of this ANR, that started in 2006, is to analyze the theoretical problems surrounding the use of semi-structured databases in social sciences. This ANR regroups both computer science and sociology laboratories: Gemo (I. Manolescu), LRI (D. Colazzo), PRiSM (B. Nguyen) and LEST (A. Vion). The goal of this project is to create a sociology platform combining traditional sociology tools with novel XML based models, storage and computing. Our current research will be presented in January 2007 in the EHESS (Ecole des Hautes Etudes en Sciences Sociales) Marlowe seminary.

7.1.5. *ARC ASAX*

ASAX (http://gemowiki.futurs.inria.fr/twiki/bin/view/Gemo/AsaxWeb) is a cooperative research action headed by the DistribCom team (INRIA Rennes), in cooperation with Gemo (S. Abiteboul), the LIAFA/Paris, and Tel-Aviv university. ASAX's purpose is the analysis of Active XML systems (http://activexml.net) and Web services. Currently, only a fragment of AXML, called "positive AXML", such as systems having monotonic answers to queries, can be given a deterministic behavioral semantics; the goal of ASAX is to break this limitation, and provide a formal semantics and analysis algorithms for AXML systems. Asax is finishing. The work should continue in a future project name docFlow.

7.2. European Commission Financed Actions

In Europe, close links exist with University of Dortmund (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. *XClean*

In the XClean project we devise an algebraic framework for cleaning XML data. The algebraic operators can be expressed by customizable XQuery expressions and functions and in this way we can use any standard XQuery processor to do the cleaning. We have also devised a declarative language specialized for XML cleaning, that can be used even for non-expert users, and can be compiled to XQuery. This approach is described in [38] and we implemented it in a prototype [35]

7.2.2. *Marie Curie Fellowship NGWeMiS*

NGWeMiS (Next Generation Web Mining and Searching) is a project lead by M. Vazirgianis (U. Athens). The project lies in the area of knowledge extraction and management from the massive and heterogeneous document collections on the World Wide Web. The main objective of the proposed project is the design guidelines and prototypes development for next generation web mining and searching techniques based on the P2P paradigm. The innovation lies in the usage of P2P paradigm in the various levels of web content management and searching, and the study and development of novel similarity measures among web documents that take into account multple facets including structure and semantics iii. clustering the web data and meta data taking into account their P2P organization paradigm.

7.3. Bilateral International Relations

7.3.1. Cooperation within Europe

Procope

This year started a PAI-Procope project with the database group of Thomas Schwentick at Dortmund University, Germany. The project will end in 2008. Its goal is to work on verification and queries in the presence of data values. It produced already several join papers between the two groups.

Polonium

This year started a PAI-Polonium project with the group of Lasota Slavomir at Warsaw University, Poland. The project will end in 2007. Its goal is to work on verification and queries in the presence of data values. It produced already several join papers between the two groups.

Van-Gogh

This year started a PAI-Van-Gogh project with the group of Lasota Slavomir at Warsaw University, Poland. The project will end in 2007. Its goal is to work on expressive power and performances of XML query languages.

7.3.2. Cooperation with the Middle-East

Close links exist with University of Tel-Aviv (T. Milo).

7.3.3. Cooperation with North America

Close links also exist with UC Santa Cruz (N. Polyzotis), U. of Rutgers (A. Borgida), U. of Toronto (L. Libkin), Google Research (O. Benjelloun),

7.3.4. 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 till 2009. The two groups met in Chicago in June 2006 in parallel with the SIGMOD conference. Bogdan Cautis spent two months in San Diego and Victor Vianu spent most of the 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 end of 2006.

7.4. Visiting Professors and Students

This year the following professors visited Verso:

- Tova Milo, professor at the University of Tel-Aviv (in February)
- Neoklis Polyzotis, professor at the University of Southern California (in September)
- Nicole Schweikardt, assistant professor at the Humbolt University of Berlin (in July)
- Victor Vianu, professor, UC San Diego (January to August)

The following PhD students came for internships in the group: Melanie Weis [Humboldt U. Berlin; 2 weeks, PhD internship], Karl Schnaitter [UC Santa Cruz; 1 month, PhD internship].

8. Dissemination

8.1. Participation in Conferences

Serge Abiteboul has been chair of the ACM SIGMOD Awards Committee. I. Manolescu has been co-chair of the First International Workshop on Performance and Evaluation of Data Management Systems (EXPDB), in cooperation with ACM SIGMOD/PODS 2006.

Members of the project have participated in program committees:

S. Abiteboul

- ACM-SIGMOD International Conference on the Management of Data (SIGMOD06)
- International Conference on Extending Data Base Technology (EDBT06)
- International Workshop on Web and Databases, (WebDB06)
- World Wide Web Conference (WWW07)

Ph. Dague

- 17th International Workshop on Principles of Diagnosis (DX) 2006
- 20th International Joint Conference on Artificial Intelligence (IJCAI) 2007

I. Manolescu

- International Conference on Extending Database Technology (EDBT06)
- EDBT Ph.D. Workshop 2006
- The 15th International World Wide Web Conference (WWW06)
- The 22nd International Conference on Data Engineering (ICDE 2006)
- The 4th International Conference on Service Oriented Computing (ICSOC 2006)
- 13th International Conference on Management of Data. (COMAD 2006)
- International Workshop on Web and Databases, (WebDB 2006)
- 3rd Int'l Workshop on XQuery Implementation, Experience and Perspectives (XIME-P 2006)

C. Reynaud

- 5th International Semantic Web Conference (ISWC) 2006.
- International Workshop on Contexts and Ontologies: Theory, practise and Applications (ECAI) 2006.
- 15ème congrès francophone de Reconnaissance des Formes et d'Intelligence Artificielle, (RFIA) 2006.
- Journées Francophones d'Ingénierie des Connaissances, (IC) 2006.
- Atelier Fouille du Web, (EGC) 2006.
- Atelier Modélisation des connaissances, (EGC) 2006.

M-C. Rousset

- 21th National Conference on Artificial Intelligence (AAAA 2006)
- 3rd European Semantic Web Conference (ESWC 2006)
- 29th International Joint Conference on Artificial Intelligence (IJCAI 2007)
- 16th International Confeence on the Principles of Knowledge Representation and Reasoning (KR 2006)
- 16th International Symposium on Methodologies for Intelligent Systems (ISMIS 2006),
- 15th international conference on knowledge engineering and knowledge management (EKAW 2006),
- First International Conference on Knowledge Science, Engineering and Management (KSEM 2006),
- Workshop on Information Integration on the Web (IIWeb 2006),

L. Segoufin

- International Symposium on Theoretical Aspects of Computer Science (STACS 2006)

L. Simon

- European Conference on Artificial Intelligence (ECAI 2006)
- International Conference on Theory and Applications of Satisfiability Testing (SAT 2007)
- International Workshop on Empirically Successful Classical Automated Reasoning (ESCoR 2006)
- Journées Francophones de Programmation par Contraintes (JFPC 2006)
- Journées Francophones de Programmation par Contraintes (JFPC 2007)

M. Vazirgiannis

- International Conference on User Modeling (UM 2007)
- Workshop on Information Retrieval in Peer-to-Peer Networks (P2PIR 2006)
- International Conference on Web Information Systems Engineering (WISE 2006)
- ACM International Workshop on Web Information and Data Management (WIDM 2006)
- European Conference on Machine Learning - European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2006)

8.2. Participation to the W3C XQuery Working Group

We participated to the XQuery Working Group (XQuery WG) as a part of the "XQuery Update Task Force". This task force has the goal of devising an update language for XML that will be an extension of the XQuery language (W3C recommendation since november 2006). The Gemo group has organised and hosted in june a joint W3C meeting in which took part the XQuery, XMLSchema and XSL working groups.

I. Manolescu is also one of the editors of the "XQuery Update Facility Use Cases" [39] a W3C recommendation that specifies usage scenarios for the XQuery Update Facility.

8.3. Invited Presentations

Serge Abiteboul presented invited talks at:

- *European Conference on Object-Oriented Programming* [8];
- *East-European Conference on Advances in Databases and Information Systems* [11];
- *Journée francophone sur les Entrepôts de Données*;
- *Conférence sur les Approches Formelles dans l'Assistance au Développement de Logiciels*;
- *Conférence sur les Bases de Données Avancées* (tutorial on Peer-to-peer Data Management); and
- INRIA meetings (20 years of INRIA-Lorraine; 10 years of INRIA-Liama, INRIA-SOPHIA Colloquium).

Marie-Christine Rousset presented invited talks at:

- *International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2006)*;
- *Second International Workshop on Exchange and Integration of Data (EID 2006)*; and
- *5th International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2006)*.

Luc Segoufin has presented an invited paper at the conference Computer Science Logic, 2006.

Laurent Simon has presented an invited talk, on *Some Results and Lessons from the SAT Competitions* at the Second International Workshop on Constraint Propagation and Implementation, held in conjunction with Eleventh International Conference on Principles and Practice of Constraint Programming (CP'2005).

8.4. Scientific Animations

Editors

C. Reynaud

- Journal Electronique d'IA de l'AFIA (JEDAI).

- Revue Information - Interaction - Intelligence (RI3).

- Revue des Nouvelles Technologies de l'Information, Special issue "Fouille du Web" (RNTI).

M-C. Rousset

- Interstices (revue electronique de vulgarisation sur la recherche en informatique): http://interstices.info/

- 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)

- Guest Editor of a Special Issue of JSAT on SAT 2006 Competitions and Evaluations.

9. Bibliography

Year Publications

Articles in refereed journals and book chapters

[1] S. ABITEBOUL, S. ALSTRUP, H. KAPLAN, T. MILO, T. RAUHE. *Compact Labeling Scheme for Ancestor Queries.*, in "SIAM J. Comput.", vol. 35, n° 6, 2006, p. 1295-1309.

[2] S. ABITEBOUL, L. SEGOUFIN, V. VIANU. *Representing and querying XML with incomplete information.*, in "ACM Trans. Database Syst.", vol. 31, n° 1, 2006, p. 208-254.

[3] 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", vol. 25, 2006, 269,314.

[4] P. ADJIMAN, F. GOASDOUÉ, M.-C. ROUSSET. *SomeRDFS in the Semantic Web* , in "Journal on Data Semantics", vol. 8, 2006.

[5] A. ARION, A. BONIFATI, I. MANOLESCU, A. PUGLIESE. *XQueC: A Query-Conscious Compressed XML Database (to appear)*, in "ACM TOIT", 2007.

[6] C. LE DUC, N. LE THANH, M.-C. ROUSSET. *A Compact Representation for Least Common Subsumers in the description logic ALE.* , in "AI Communications", vol. 19, n⁰ 3, 2006, p. 239–273.

[7] A. LEGER, J. HEINECKE, L. J. NIXON, P. SHVAIKO, J. CHARLET, P. HOBSON, F. GOASDOUÉ. *The Semantic Web from an Industrial Perspective* , in "Reasoning Web (LNCS, Tutorial Lectures)", 2006.

Publications in Conferences and Workshops

[8] S. ABITEBOUL. *Turning the Network into a Database with Active XML*, in "European Conference on Object-Oriented Programming", 2006.

[9] S. ABITEBOUL, B. CAUTIS, A. FIAT, T. MILO. *Digital Signatures for Modifiable Collections.*, in "ARES", 2006, p. 390-399.

[10] S. ABITEBOUL, I. MANOLESCU, E. TAROPA. *A Framework for Distributed XML Data Management.*, in "EDBT", 2006, p. 1049-1058.

[11] S. ABITEBOUL, N. POLYZOTIS. *Data Ring: Let Us Turn the Net into a Database!*, in "ADBIS", 2006.

[12] S. ABITEBOUL, N. POLYZOTIS. *The Data Ring: Community Content Sharing*, in "Conference on Innovative Database Systems Research", 2007.

[13] S. ABITEBOUL, P. SENELLART. *Querying and Updating Probabilistic Information in XML.*, in "EDBT", 2006, p. 1059-1068.

[14] A. ARION, V. BENZAKEN, I. MANOLESCU, Y. PAPAKONSTANTINOU, R. VIJAY. *Algebra-Based Identification of Tree Patterns in XQuery*, in "FQAS", 2006.

[15] A. ARION, A. BONIFATI, I. MANOLESCU, A. PUGLIESE. *Path summaries and path partitioning in modern XML databases(poster)*, in "WWW", 2006.

[16] M. BOJAŃCZYK, C. DAVID, A. MUSCHOLL, T. SCHWENTICK, L. SEGOUFIN. *Two-variable logic on data trees and XML reasoning*, in "ACM Symposium on Principles of Database Systems (PODS)", 2006.

[17] M. BOJAŃCZYK, A. MUSCHOLL, T. SCHWENTICK, L. SEGOUFIN, C. DAVID. *Two-Variable Logic on Words with Data*, in "IEEE Symposium on Logic in Computer Science (LICS)", 2006.

[18] M. BOJAŃCZYK, M. SAMUELIDES, T. SCHWENTICK, L. SEGOUFIN. *Expressive Power of Pebble Automata*, in "International Colloquium on Automata, Languages and Programming (ICALP)", 2006.

[19] B. CAUTIS, S. ABITEBOUL, T. MILO. *Reasoning about XML Update Constraints*, in "BDA", 2006.

[20] P. CHATALIC, G.-H. NGUYEN, M.-C. ROUSSET. *Reasoning with Inconsistencies in Propositional Peer-to-Peer Inference Systems.* , in "ECAI 2006 (European Conference on Artificial Intelligence)", August 2006, p. 352–357.

[21] H. GAGLIARDI, O. HAEMMERLE, N. PERNELLE, F. SAIS. *Decouverte de relations candidates à l'enrichissement d'entrepot thematiques* , in "COSI (Colloque sur l'optimisation et les systemes d'information), également présenté à l'atelier Fouille du web de EGC06", june 2006.

[22] H. KEFI, B. SAFAR, C. REYNAUD. *Alignement de taxonomies pour l'integration de sources d'information heterogenes* , in "Reconnaissances des Formes et Intelligence Artificielle", 2006.

[23] I. MANOLESCU, V. BENZAKEN, A. ARION, Y. PAPAKONSTANTINOU. *Structured Materialized Views for XML Queries*, in "BDA", 2006.

[24] I. MANOLESCU, C. MIACHON, P. MICHIELS. *Towards micro-benchmarking XQuery*, in "ExpDB", 2006.

[25] T. MELLITI, C. BOUTROUS-SAAB, L. MOKDAD. *Performance evaluation for mobile access to composite Web services*, in "Proc. of International Conference on Internet and Web Applications and Services (ICIW'06), Guadeloupe, French Caribbean", IEEE Computer Society, 2006.

[26] T. MELLITI, C. BOUTROUS-SAAB, S. RAMPACEK. *Verifying correctness of Web services choreography*, in "Proc. of 4th IEEE European Conference on Web Services, Zurich, Switzerland", IEEE Computer Society, 2006.

[27] C. REYNAUD, B. SAFAR, H. KEFI. *Structural techniques for alignment of structurally dissymmetric taxonomies* , in "15th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2006", H. S. PINTO, M. LABSKY (editors). , Zeithamlova Milena, Ing. - Agentura Action M, October 2006, p. 39-40.

[28] C. REYNAUD, B. SAFAR. *Mappings pour l'integration de documents XML* , in "Atelier Modélisation des connaissances - EGC 2006", Janvier 2006.

[29] C. REYNAUD, B. SAFAR. *When usual structural alignment techniques don't apply* , in "The ISWC'06 workshop on Ontology matching (OM-06)", 2006.

[30] M.-C. ROUSSET, P. ADJIMAN, P. CHATALIC, F. GOASDOUÉ, L. SIMON. *Somewhere in the Semantic Web.* , in "Invited talk, SOFSEM 2006(Int. Conf. on Current Trends in Theory and Practice of CS)", also in Proceedings of PPSWR 2005 (International Workshop on Principles and Practice of SW Reasoning, January 2006.

[31] M.-C. ROUSSET, P. ADJIMAN, P. CHATALIC, F. GOASDOUÉ, L. SIMON. *SomeWhere: a scalable P2P infrastructure for querying distributed ontologies.* , in "ODBASE 2006 (5t International Conference on Ontologies Databases and Applications of Semantics", R. MEERSMAN, Z. TARI, ET AL (editors). , Springer, October 2006, p. 698–703.

[32] K. SCHNAITTER, S. ABITEBOUL, T. MILO, N. POLYZOTIS. *COLT: continuous on-line tuning (demo)*, in "SIGMOD Conference", 2006, p. 793-795.

[33] L. SEGOUFIN. *Automata and logics for words and trees over an infinite alphabet*, in "Conference for Computer Science Logic (CSL)", 2006.

[34] C. SIRANGELO, L. SEGOUFIN. *Constant-memory validation of streaming XML documents against DTDs*, in "International Conference on Database Theory (ICDT)", 2007.

[35] M. WEIS, I. MANOLESCU. *XClean in action(Demo)*, in "CIDR", 2007.

Internal Reports

[36] Y. LI. *Modeling BPEL Web services for diagnosis: towards self-healing Web services*, Technical report, 2006.

[37] C. REYNAUD, B. SAFAR. *Structural techniques for Alignment of taxonomies: experiments and evaluation*, Technical report, Gemo, June 2006.

[38] M. WEIS, I. MANOLESCU. *Declarative XML Data Cleaning with XClean*, Technical report, Gemo, 2006.

Miscellaneous

[39] W3C XQUERY WORKING GROUP. *XQuery Update Facility Use Cases (W3C recommendation). Available at , http://www.w3.org/TR/xqupdateusecases/*, 2006.