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

Table of contents

1. Team

INRIA personnel

Serge Abiteboul [DR-INRIA, HdR]
Ioana Manolescu [CR-INRIA]
Luc Segoufin [DR-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]

Administrative Assistant

Stéphanie Meunier [ITA, HdR]
Marie Domingues [ITA, HdR]

Scientific Advisors

Sophie Cluet [Department Director, MESR]
Tarek Melliti [Assistant Professor, Univ. Evry Val d'Essonne]
Dan Vodislav [Assistant Professor, CNAM Paris]
Philippe Rigaux [Professor, U. Paris Dauphine, on sabbatic]
Marie-Christine Rousset [Professor, Univ. Grenoble, HdR]

Invited researchers and visitors

Laura Brandan Briones [Post Doc fellowship]
Farid Nouioua [Post Doc fellowship]
Neoklis Polyzotis [Assistant Professor, U.C. Santa Cruz, 1 month]
Hassan Shraim [Post Doc fellowship]
Michalis Vazirgiannis [Professor, U. Athens, Marie-Curie fellowship]
Victor Vianu [Professor, U.C. San Diego, 2 months]
Yuhong Yan [Research Officer, NRC Canada, IIT, Fredericton, 1 month]
Haïfa Zargayouna [Post Doc fellowship]

Engineers

Omar Aaouatif [3 months internship]
Ali Aharbil [3 months internship]
Anca Ghitescu [from November]
Mohamed Ouazara [from September]
Evaldas Taroza [till October]
Gabriel Vasile

Ph.D students

Nada Abdallah [Allocataire MENRT, Paris 11]
Andrei Arion [Allocataire MENRT, Paris 11]
Vincent Armant [Allocataire MENRT, Paris 11]
Pierre Bourhis [ENS Cachan, since September]
François-Elie Calvier [Grant BDI CNRS, Paris 11]
Bogdan Cautis [Allocataire MENRT, Paris 11, till September]
Claire David [ENS Cachan]
Charaf Laissoub [Contrat ANR, till April]

Yingmin Li [European contract]
Gia Hien Nguyen [Grant MENRT, Grenoble 1]
Bogdan Marinoiu [Grant BDI CNRS, Paris 11]
Antonella Poggi [PhD in cotutelle between U. di Roma and Paris 11, till April]
Nicoleta Preda [Allocataire MENRT, Paris 11]
Radu Pop [Cifre with Mandriva Software]
Cedric Pruski [PhD in cotutelle between Luxembourg U. and Paris 11]
Fatiha Sais [Contrat FTRD]
Mathias Samuelides [ENS Cachan]
Pierre Senellart [ENS Ulm]
Mouhamadou Thiam [PhD in cotutelle between Gaston Berger U. and Paris 11]
Lina Ye [Allocataire MENRT, Paris 11]
Spyros Zoupanos [CORDI]

2. Overall Objectives

2.1. Introduction

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.

2.2. Highlights of the year

A lot of work has been devoted to the ANR Project WebContent.

Serge Abiteboul has been the recipient of the EADS Award in Computer Science, that is selected by the Academy of Science.

3. Scientific Foundations

3.1. 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*.

A main theme of the team 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 [19]. 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*.

4. Application Domains

4.1. 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.

5. Software

5.1. 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.
SomeWhere+: a P2P infrastructure tolerant to inconsistency.
KadoP: a peer-to-peer platform for warehousing of Web resources.
OptimAX: an algebraic cost-based optimizer for ActiveXML.
TaxoMap: a prototype to automate semantic mappings between taxonomies
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.
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
LN2R: a logical and numerical tool for references reconciliation

6. New Results

6.1. 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.

With XML documents, data may be extracted, queried, or used in navigation, because of its association with a position in a document, rather than because of its actual content. It is thus believed that those foundations will be based on tree automata and on Monadic-Second-Order (MSO) logic making use of the tree structure of XML data. Towards this direction we studied in [38] some complexity issues related to a sequential family of tree automata, which has the same expressive power than unary Transitive-Closure logic. But of course data values cannot be completely ignored. In [7] we show how to use decidable logics over infinite alphabet in the XML context for deciding XML-schema validation and XPath query inclusion.

By essence XML is used in an Internet based environment. On the Web, one may have to process on the fly, heavy streams of information, to support the surveillance of rapidly changing data sources. Also, by the nature of the web, the information is imprecise, incomplete, inconsistent, of uneven quality. The answer to a Web query may include huge number of results (see Google search) and it is typically as important to rank these results than to obtain them.

We have considered streaming XML data with limited memory resources. In this context, we considered in [44] the DTD validation problem: checking whether a XML document conforms a DTD.

It is often desirable that a user only has a partial access to a database and several users see different parts of the databases. The subpart that is seen by a user is called a *views*. When a user specify its query, this query has to be rewritten according the real database and then evaluated. In [32] we study the language necessary for rewriting a Conjunctive Queries.

We continued in [43] work on probabilistic semi-structured models. We give complexity results about the probabilistic tree model (based on trees where nodes are annotated with conjunctions of probabilistic event variables) that was previously introduced. We identify a very large class of queries for which queries and elementary updates are tractable. We also consider other theoretical issues, as the equivalence of probabilistic trees or the validation of a probabilistic tree against a DTD.

A new challenge is the study of XML when used in the dynamic environment of the Web. 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.

6.2. Ontology-Based Information Retrieval

Keywords: *information retrieval*, *ontology evolution*, *query enrichment*, *resource annotation*.

Participants: Nathalie Pernelle, Cedric Pruski, Chantal Reynaud, Mouhamadou Thiam.

6.2.1. Ontology-Based Queries Enrichment

In order to improve Web Information Retrieval using Ontologies, we proposed an extension and an implementation of OWL for Web Queries Enrichment. This work has been done in the setting of the O3 approach designed by Cedric Pruski in Luxembourg during his Master of Science. O3 uses the WordNet linguistic tool in order to optimize, in terms of relevance, the returned documents when searching the Web. Its main idea consists in enriching, following well-defined rules, the query constructed by users by extracting from WordNet the appropriate vocabulary that characterizes best the search domain. O3 has been formalized using first-order logic and graph theory. This formal framework permitted the rigorous definition of query expansion rules. In parallel, the standardization of OWL has hastened the quick and massive development of OWL ontologies across the Web. This is why, to benefit from both O3 and OWL ontologies, we decided to make O3 compatible with OWL. We studied the possibilities offered by OWL that cope with O3 as well as an extension of the language. We implemented the so called extended OWL through query expansion rules and we made an experimental validation using the TARGET tool 2007 [27].

6.2.2. Supporting Ontology Evolution

First, we surveyed techniques for ontology evolution 2007 [5]. After identifying the different kinds of evolution the Web is confronted with, we detailed the various existing languages and techniques devoted to Web data evolution, with particular attention to Semantic Web concepts, and how these languages and techniques can be adapted to evolving data in order to improve the quality of Web Information Systems Applications. Second, we proposed a set of modelling features for ontology evolution 2007 [28]. These features have been defined after the rigorous study of the evolution of a particular domain (the domain defined by the WWW series of conference topics) over a ten years period of time. The results of this study lead directly to the definition of the various kinds of evolution that can appear. They allowed the proposition of modelling features that aims at designing evolving ontologies. Indeed, these features will allow us to understand the evolution of ontologies and will aid to predict future versions of ontologies. We highlighted the contribution of such ontologies through an example implementing ontology-based query expansion techniques to improve the relevance of documents when searching the web.

6.2.3. Semantic Annotation

Where data sources are numerous and heterogeneous, a data integration system needs automatic tools to annotate and query semi-structured documents. We propose an automatic approach for semantic annotation of HTML or XML documents [45]. It relies on the model describing the domain of interest. The difficulty lies in the heterogeneous structure of the documents and in that a document contains both structured and unstructured parts. To overcome this problem, we have defined a first set of annotation rules using SWRL. That rules take into account both the semantic relations defined in the model and the heterogeneity of documents structure. The resulting annotated documents are represented in RDF language according to the semantic RDF Schema model which is extended to the annotation task from the domain description. Since october 2007, this work is done in the setting of the SHIRI project which is supported by the DIGITEO Foundation.

6.3. Peer-to-Peer Inference Systems

Keywords: *RDF*, *distributed diagnosis*, *distributed reasoning*, *inconsistency*, *ontology alignment*, *propositional logic*.

Participants: Nada Abdallah, Vincent Armant, François-Elie Calvier, Philippe Chatalic, Philippe Dague, François Goasdoué, Gia Hien Nguyen, Chantal Reynaud, Marie-Christine Rousset, Laurent Simon.

Peer-to-Peer Inference Systems (P2PISs) are made of autonomous peers (i.e., built and managed independently) that can communicate in order to perform an inference task at the P2PIS level (e.g., consequence finding or query answering). For that purpose, communication rules between peers are modeled by *mappings* that define semantic relationships between their knowledge.
A crucial aspect of that new setting is that peers are equivalent in functionalities and no actor has a global view of a P2PIS, i.e., there is no centralized control or hierarchical organization in the system. Each peer only knows the knowledge it manages and its mappings with some other peers. This raises exciting non trivial algorithmic issues since, in the literature, reasoning algorithms have been designed with the assumption that the knowledge on which inferences have to be performed is given as an input. New decentralized algorithms have to be designed with the idea that only a subset of the global knowledge is available to a peer as an input (i.e., the peer's knowledge and mappings), but the algorithms have still to be sound and complete for the inference task w.r.t. the global knowledge of the P2PIS (i.e., the knowledge and mappings of all the peers). The SomeWhere platform has been developed for experimenting with such distributed reasoning tasks. 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.

6.3.1. Consequence Finding

Many challenging Artificial Intelligence (AI) tasks like common sense reasoning, diagnosis, or knowledge compilation can be stated in terms of *consequence finding*. That key inference basically consists in deriving theorems of interest that are intentionally characterized within a logical theory. Such theorems can be those in terms of a fixed language, those resulting from some incoming knowledge in the theory, etc.

Recently, we have designed the first peer-to-peer inference systems for consequence finding, in which each peer manages a clausal theory of propositional logic in terms of its own set of propositional variables. A peer establishes (or suppresses) a mapping by adding to (or removing from) its theory a clause made of some of its variables and some variables from other peers, those peers being notified of the operation. For those systems, we have proposed the DEcentralized Consequence finding Algorithm (DECA) that performs a decentralized resolution procedure in order to compute clausal implicates (i.e., consequences) of a clause submitted to a peer w.r.t. a P2PIS, including *all* the *proper prime ones* (i.e., the strongest consequences).

A key point in the design of the above P2PISs is that mappings are *undirected*, i.e., any peer involved in a mapping can use it to propagate knowledge to the other peers participating in the mapping. Therefore, such systems model autonomous components that communicate through interfaces that are both input and output.
We have recently proposed an alternative design of P2PISs in which mappings are *directed*. A mapping is stated between two peers, but only one of them can use the mapping to propagate knowledge to the other. From a practical viewpoint, a mapping from a peer to another specifies some knowledge that the former peer has to observe and the knowledge it must notify to the latter peer if the observed knowledge holds. Such new P2PISs are of great interest in order to apply AI reasoning because they can model many real applications in which autonomous components communicate through interfaces that are either input or ouput, like distributed functions in Automotive Engineering, distributed control systems for industrial machinery or processes in Automation, etc. For those systems, we have proposed a new DEcentralized Consequence finding Algorithm for directed mappings (DECA_K) that computes clausal implicates of a clause submitted to a peer w.r.t. a P2PIS, including *all* the proper prime ones.
In P2PIS retaining the classical semantics for mappings (i.e. undirected view), the ability of each peer to freely add new mappings with other peers may have affect the consistency of the global resulting theory. This cannot be avoided because of the decentralised nature of the architecture. In order to prevent the trivialisation of the reasoning in such cases, we have designed a method able to detect incrementally all possible minimal causes of inconsistency and to store them in a distributed way in the P2PIS. Furthermore, we have proposed a new distributed consequence finding algorithm (WFDECA) able to perform well founded reasoning despite the presence of possible inconsistencies. These algorithms have been implemented and an experimental evaluation is underway. One noticeable feature of WFDECA is that different consequences, though all well founded, may have different supports that are not necessarily consistent with each other. In such cases, it is up to user's responsibility to choose between consequences having incompatible supports. One possible choice criteria is to prefer the most *trustable* consequences. We are currently investigating on different trust models that have been proposed for P2P file sharing system and consider their possible adaptation the task of distributed consequence finding.

6.3.2. Distributed Diagnosis

The logical theory of consistency-based diagnosis has been worked out in the eighties in the centralized case. It starts from a model, assumed to be given, of the behavior of the (component based-) system in consideration (correct behavior and possibly some faulty behaviors if known by advance) and aims at maintaining consistency between the current hypotheses of behavioral modes of the components (correct or faulty) and the observations (e.g. sensors measurements). It is stated in a logical framework, where the model SD - for System Description - and the observations OBS are expressed in first order logic, the mode of each element in COMPS being explicitly represented thanks to the predefined Ab (for Abnormal) predicate (so, $\neg$ Ab(c) means that component c is correct and Ab(c) that component c is faulty). Diagnostic reasoning is a typical example of non monotonic reasoning: initially all components are assumed to be correct, up to the moment this becomes inconsistent with observations. Then consistency between the model and the observations is restored by changing some component mode assignment from correct to faulty (in general, a principle of parsimony is applied and we are interested only in minimal - for cardinality or for set inclusion - sets of faulty components). Technically, from a logical inference point of view, computation of the diagnoses (complete components modes assignments consistent with observations) relies on calculus of (prime) implicates and implicants of SD $\cup$ OBS in terms of the target language built from the Ab(c), for c in COMPS. This diagnostic activity can be done off line - from a given set of observations - or on line in a

general monitoring framework where new observations occur along time, and where the real (unknown) mode of each component can itself vary along time (from correct to faulty but also from faulty to correct in case of transient faults). Assuming centralized system, centralized model, centralized diagnostic algorithm is a severe restriction for several real case applications: the system can be "naturally" distributed (telecommunication networks, Web services, etc.), the system can be too huge or complex to have a unique storable and accessible global model, privacy issues can prevent the existence of such a global model, the diagnostic algorithm can take advantage to perform decentralized local diagnosis and its implementation to be decentralized on several control units. This is why decentralized diagnosis receives a growing interest from some years. The work that has been initiated is an attempt to design, implement and test distributed consistency-based diagnosis algorithms in a logical framework, relying on previous work conducted inside Gemo on P2PISs, in particular what concerns consequence finding and handling of inconsistencies. In this P2P framework, each peer represents a subsystem and its local theory is the propositionalized subsystem description, the mappings (shared variables) expressing connections between subsystems. Observation peers (sensors) have a local theory limited to a propositional symbol expressing the measurement's value. The algorithm currently developed for generating minimal diagnoses relies on a distributed computation of (restrictions to the target language of) implicants of the global (unknown as a whole) theory [52]. Several problems will have to be addressed in the future: incrementality w.r.t. increasing asynchronous observations; characterization of all diagnoses in presence of fault models; on line monitoring and diagnosis with observations varying asynchronously along time; repair by reconfiguring the system (changing mappings); open world (addition or suppression of peers), etc.

6.3.3. Mapping distributed ontologies

In the setting of the MediaD project we address the problem of discovering mappings between distributed ontologies in the setting of SomeRDFS, a peer data management system (PDMS)derived from SomeWhere. Since the setting of PDMS is particular, we proposed techniques to take advantage of SomeRDFS reasoning in order to help discovering mappings between the knowledge of each peer, i.e. ontologies, that can be mapping shortcuts or new mappings 2007 [24]. The aim of the proposed techniques is to discover elements that are relevant to be mapped. These elements will then be aligned applying usual alignment techniques. The implementation of this work is in progress.

6.4. Thematic Web Warehousing

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

Participants: François-Elie Calvier, Hélène Gagliardi, Nathalie Pernelle, Chantal Reynaud, Marie-Christine Rousset, Fatiha Sais, Brigitte Safar, Haïfa Zargayouna.

6.4.1. Reference reconcilation

We are 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, ...). We have developped a logical and numerical approach named LN2R (L2R + N2R) which is automatic and guided by the semantic of an RDFS+ schema. L2R is logic-based [40], [39]. In the N2R method, the semantics of the schema is exploited by an informed similarity measure which is used by a numerical computation of the similarity of reference pairs. This numerical computation is expressed in an equation system that is non linear. We have shown on one benchmark dataset that we can obtain better results than supervised approach. We have also studied the scalability of such approaches [34], [41]. This work is done in the setting of the Picsel3 project.

6.4.2. Mapping between ontologies

This work has been initiated in the setting 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, structural 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 2007 [6].We continue our work on TaxoMap in the setting of the WebContent project. First, we investigated techniques which rely on an additional source, called background knowledge. We made a comparative analysis of works using background knowledge 2007 [36]. We studied the difficulties encountered when using Wordnet 2007 [37] and we showed how the Taxomap system can avoid these difficulties 2007 [35]. Further work has been done on adapting TaxoMap for the Ontology Alignment Evaluation Initiative (OAEI 2007) campaign. So we participated in the OAEI 2007 campaign 2007 [51] which consists of applying matching systems to ontology pairs and evaluating their results. Moreover, TaxoMap has been tested and evaluated together with OLA jointly developed by the teams at Diro, university of Montreal and at INRIA Rhône-Alpes (EXMO group) in the setting of the WebContent project. The following corpora have been chosen: a corpus delivered by EADS in the aeronautics field, OAEI Benchmark test and AGROVOC-NAL, two very rich thesauri used in the "food" corpus in the OAEI 2007 campaign. These experiments have shown the complementary nature of the two tools and have emphasized two main difficulties: the alignment of very large ontologies and the evaluation of the results when no reference mappings are provided.

6.5. XML query optimization

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

Participants: Serge Abiteboul, Andrei Arion, Ioana Manolescu, Spyros Zoupanos.

6.5.1. Materialized views for XML queries

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 proposed a new language for describing materialized XML views, which can be used to speed up the processing of XML queries. We have devised associated algorithms for rewriting XQuery queries based on this rich view language [21].

While materialized views can speed up query processing, their practical applicability requires several developments. First, they have to be maintained in the event of updates applied to the underlying documents. The internship of Abhipreet Das (IIT Bombay) has focused on proposing algorithms for incrementally propagating updates to the materialized views. Second, view selection may be cumbersome to the user, therefore automated view selection mechanisms are needed. The internship of Nikhil Pandey (IIT Bombay) has lead to some work in this area, however the problem was not fully solved.

6.5.2. Algebraic optimization for ActiveXML

The ActiveXML language (AXML in short) allows describing complex distributed data manipulation tasks. Each such task could be executed in many ways producing the same results but with very different performance. We have made important progress in laying out an algebraic formalism for optimizing AXML document evaluation, more precisely on specifying a small set of special Web services dedicated to distributed evaluation and on their usage within the optimizer. The first prototype of an AXML optimizer, OptimAX, has been developed and demonstrated [15]. The optimizer is integrated with a new version of an AXML peer, developed mostly this year by E. Taroza.

6.5.3. Performance evaluation methodology

Performance evaluation is a natural component in many data-oriented works such as those carried on in Gemo. However, the complexity of the languages we target, such as XQuery, and the complexity of settings in which our techniques are deployed, such as peer-to-peer systems, make the task of performance evaluation very complex. For instance, in a peer-to-peer XML data management setting, one has to distinguish the impact of the underlying peer network from that of data indexing, from that of query evaluation algorithms, and finally from the optimizer quality. Benchmarks are essential tools for performance evaluation. We have proposed a benchmark for XML data management in P2P, named P2PTester [23], designed to ease and systematize the

task of performance evaluation. Performance evaluation in the large raises lively discussion; a panel organized in the VLDB conference on this topic received significant attention [30]. Participants agreed on the need for a more thorough procedure both for performing performance evaluation and for ensuring such evaluations are repeatable.

6.5.4. Self tuning

We started some collaborative work with UCSC and U.Tel Aviv on a framework for Continuous On-Line Tuning [42], 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 developped 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.

6.6. XML Warehousing in P2P

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

Participants: Serge Abiteboul, Ioana Manolescu, Nicoleta Preda, Melanie Weis.

6.6.1. XML indexing: KadoP

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.

A recent development of the system includes two techniques meant to handle efficiently long posting lists exchanged during query processing. The first technique relies on a distributed search structure that parallelizes the transfer of long posting lists, while the second enables to reduce the transferred lists at the expense of some precision. These techniques are described in [14].

We have also participated to the development of a prototype for measuring the performance of P2P queries [23].

6.6.2. XML data cleaning: XClean

In the context of XML data warehousing, it often happens that different XML representations of a same object appear in the sources. In this context, it becomes necessary to identify common entities in the XML sources and propose a consolidated version thereof. We have proposed the XClean framework for declaratively specifying data cleaning processes, which are then compiled into XQuery queries [48]. M. Weis has developed a prototype implementing this framework, which has been demonstrated [49].

6.7. Monitoring and Web services

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

Participants: Serge Abiteboul, Laura Brandan Briones, Pierre Bourhis, Philippe Dague, Yingmin Li, Bogdan Marinoiu, Tarek Melliti, Lina Ye.

6.7.1. 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, up to mid 2008. 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 the two first years, 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 algorithm 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 (analogous to dynamic slicing methods in software debugging), 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. This data dependency model is expressed as a set of propositional Horn clauses that will be used by the diagnosis algorithm.

The enriched Petri net generator, which takes as input a BPEL code and produces as output its enriched Petri net model in the form of an xml file, and the diagnostic model compiler, which takes as input an enriched Petri net model and produces as output its associated diagnostic knowledge base as a set of causal rules expressed as logical Horn clauses, both in the form of xml files, have been implemented in Java and tested on examples (in particular the Foodshop service used in the WDS-DIAMOND project).

- Developing a decentralized diagnostic algorithm
  A decentralized on line diagnostic algorithm for BPEL orchestrated Web services has been designed, that relies on the local diagnostic models of each Web service built off line as explained above and on the observations stored on line by the data log platform. A local diagnoser is provided to each BPEL service, that performs local consistency-based diagnosis thanks to the local diagnostic model of the service (initially a diagnostic session is triggered when a local diagnoser is awakened by an exception raised in its associated Web service). The local output diagnosis is made up of possible local faults as input data from users or faulty internal basic Web services (among those invoked by the BPEL service), or of input variables coming from shared variables in another composite Web service. These local diagnosers communicate (in both ways) with a coordinator, in charge of building global diagnoses by merging local ones. This coordinator does not initially have any information about the individual Web services except the shared variables between them, which are obtained off line and are at interface level, satisfying thus privacy issues. The coordinator tries to prolong each local diagnosis containing a suspected input variable coming from another service by invoking the local diagnoser of this service. At the end, global diagnoses thus generated are made up of input faulty data from users, faulty internal basic Web services or faulty interfaces between two Web services (these last ones being able often to be checked for confirmation through logged observation). In fact, the local diagnosers and the coordinator are regarded also as Web services communicating via WSDL messages, thus WSDL standard can be used to describe the diagnosis operation offered by a diagnostic Web service. Up to now, the local diagnosers and the coordinator have been implemented as Java objects, thus basic Web services, and interfaced with the data log platform, and are currently tested on applications, such as the Foodshop service.

In 2007, this work has been published in [29], [50], [53]. Direct continuation of this work will include: implementing the diagnostic coordinator as a BPEL Web service; extending the diagnostic architecture to the case of choreographed Web services and testing the whole on real examples. Notice that the thesis work just set about by Vincent Armant about distributed diagnosis in a peer-to-peer framework is expected to be later tested with the local diagnostic knowledge bases of Web services produced here, in order to provide a completely distributed monitoring and diagnostic platform for Web services. Another connected work that just begins is the study of diagnosability (and recoverability) properties at design stage. The aim is to define formal properties of a discrete-event model, together with a predefined set of faulty non observable events and a predefined set of observable events, expressing that a given fault will always be detectable or that two given faults will always be discriminable, and then to design algorithms to check off line these properties on the model. These criteria and checking methods will be adapted for study of Web services diagnosability and recoverability and a methodology for designing Web services applications that respect these criteria will be developed.

6.7.2. P2P Monitoring

We have worked on the conception and implementation of tools for monitoring Peer to Peer Systems.

A system named P2PMonitor has been developed for this purpose. It is a P2P system itself, with peers exchanging messages by Web service calls. This system is based on alerters, that are software modules placed on monitored peers, in charge of the surveillance of particular types of events (e.g. web service calls, database updates etc.).They produce streams of (Active)XML data. Our system implements an algebra over data streams. A declarative language allows the user to specify the complex events of interest and the ways the notifications about these events should be created and sent to her. The system is in charge of choosing the best execution plan and of placing the processors on peers. This work has been published in [17] and [16].

A subject related to monitoring is view maintenance over active documents. Indeed, the monitoring problem can be seen as aggregating streams into an active document and incrementally evaluating a tree-pattern query over this active document. We have developed algorithmic datalog-based foundations for such an incremental query processing and this work has been published in [10].

A paper presenting a demonstration scenario for the monitoring system integrating the view maintenance for active documents as a way of defining complex monitoring tasks, has been published in [18].

6.8. Web and Graph Mining

Keywords: *Graph mining*, *Web mining*, *similarity*.

Participants: Serge Abiteboul, Pierre Senellart, Michalis Vazirgiannis.

We introduce in [33] a new method for finding nodes semantically related to a given node in a hyperlinked graph, namely the Green method, based on classical Markov chains. It is generic, adjustment-free and easy to implement. We test it in the case of the hyperlink structure of the English version of an on-line encyclopedia, namely Wikipedia. We present an extensive comparative study of the performance of our method compared to several other classical methods. The Green method is found to have both the best average results and the best robustness.

In [8], we review a number of classical text mining approaches to synonym extraction over different kinds of corpora. We also introduce a graph mining technique that discovers related words in a monolingual dictionaries, closely inspired by Kleinberg's hubs and authorities, and discuss the more profound relations between classical text mining problems and graph mining.

7. Contracts and Grants with Industry

7.1. 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.

7.2. 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.

7.3. 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 entity.

7.4. EC Edos Project

EDOS 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), research centers (CSP Torino) and private companies (Mandriva, Caixa Magica, Nexedi, Nuxeo, Edge-IT). 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 process of data 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 combines the functionalities of content (software) distribution with the idea of exchanging XML data in a P2P environment, in our case metadata about the software modules to be distributed. Metadata includes identifiers (name, version), static (size, license, summary, etc) and dynamic properties of software modules (composition, replica locations, statistics about the distribution process, etc).

We defined the P2P system architecture, based on three categories of actors: Publishers (that introduce 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 (Publishers and Mirrors), storing the distributed index on metadata, and the distribution network, composed of all the peers, storing content replicas. The system's software architecture is based on a Java API implementing content distribution functionalities at several abstraction levels: 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 situations (many sources, content updates).

The project was successfully ended in September 2007. The effort in the last period has been directed to the consolidation of the system, to several optimizations, to the integration of security mechanisms, to the development of an advanced GUI and to an evaluation on the Grid'5000 platform. The EDOS content distribution system has been published as an open source project on the INRIA Gforge site (http://gforge.inria.fr/projects/edos-cds/) and a first version is included in the Mandriva distribution. The system has represented a real world application for the Gemo P2P software (KadoP, ActiveXML) and led to many improvements in these modules. The work on the EDOS distribution system has been presented to several conferences: VLDB 2007 [13] (a demonstration of the system), OSS 2007 [11] and FOSDEM 2007 [12] (architecture and functionalities) and BDA 2007 [20] (evaluation on Grid'5000).

7.5. RNTL Project WebContent

The WebContent project (http://www.webcontent.fr) has completed its first year in July 2007. 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 activity in WebContent this year has been manifold. In the architecture group (Lot 0), we have secured an agreement on the usage of Web services as means of interconnecting the project components. In the peer-to-peer group (Lot 5), we have completed a total rehaul of the AXML peer developed by Gemo. E. Taroza has proposed and implemented a new peer, more robust, and more modular; for instance, the XML storage services provided by the AXML peer have been isolated as a separated component and delegated to the eXist system. M. Ouazara is currently moving this storage component to MonetDB, the system which was retained by the WebContent consortium. At the same time, we have worked on extending the KadoP system to support XML namespaces. The development of OptimAX, the algebraic optimizer for AXML, also contributes to WebContent. Finally, we are currently extending the SomeRDFS prototype with the ability to translate from SPARQL to XQuery, as required by the WebContent integrated platform. In the semantic enrichment of ontologies and documents group (Lot 3), we evaluated our alignment tool TaxoMap both on corpora delivered by WebContent partners and on tests provided by the 2007'International OAEI campaign.

Results of the algorithms used by TaxoMap have been analyzed and comparisons with results obtained by OLA (developed by INRIA Rhône-Alpes, EXMO group) have been provided. Adaptations of TaxoMap have been made for the evaluation, specially to link the application to the Alignment API. The experiments highlighted the need to process large-scale ontologies. We are currently working in this direction.

7.6. 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), Tarek Melliti (post-doc from Oct. 1st 2005 to Aug. 31th 2006, assistant professor at U. of Evry from Sep. 1st 2006), Yingmin Li (master internship from April 1st 2006 to Sept. 30th 2006, Ph.D. student from Oct. 2006), Lina Ye (master internship from March 19th 2007 to Sept. 18th 2007, Ph.D. student from end of Sept. 2007), Laura Brandan Briones (post-doc from May 2007) and Omar Aaouatif (engineer internship from March 5th 2007 to June 4th 2007).

8. Other Grants and Activities

8.1. National Actions

In France, close links exist with groups at Orsay (databases, V. Benzaken and N. Bidoit; bio-informatics, C. Froidevaux; machine learning, M. Sebag), with the Cedric Group at CNAM-Paris; some INRIA groups (Atlas, P. Valduriez, DistribCom, A. Benveniste, at INRIA-Bretagne, Exmo, J. Euzenat, at INRIA Rhone-Alpes, Mostrare at INRIA Futurs Lille); the BIA group at INRA (P. Buche, C. Dervin), the GRIMM of the University of Toulouse Le Mirail (O. Haemmerlé), the LIRIS of the University of Lyon 1 (M. Hacid), the LIRMM of the University of Montpellier (M. Chein, M-L. Mugnier), the LI of the University of Tours (G. Venturini), and the UMPA at École normale supérieure de Lyon (Y. Ollivier).

8.1.1. DocFlow

DocFlow is a research project supported by the ANR Masses de données (2007-2009) with the Distribcom team at INRIA-Rennes (Albert Benveniste) and the Méthodes Formelles group at Labri-Bordeaux (Anca Muscholl). The topic is the analysis, monitoring, and optimization of Web documents and services. It builds on Active XML, a formalism for data exchange across peers developed by Gemo. The project aims at achieving a convergence of data and workflow management over the Web through the concept of active peer-to-peer documents.

8.1.2. 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 ended during the summer 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 could be found at: http://www.cduce.org/tralala.html.

8.1.3. 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). The project has completed this year, but the work carried on has been merged (and continues through) the WebStand project (see below).

8.1.4. ANR JCJC WebStand

The objective of this ANR, that started in 2006, is to analyze the problems surrounding the use of semi-structured databases in social sciences. This ANR regroups both computer science and sociology laboratories. Work done in Gemo which contributes to WebStand includes XML data cleaning [48], [49] and work on automatic selection and maintenance of materialized XML views. The joint work of the consortium has lead to a publication in a social sciences conference [54].

8.1.5. SHIRI Digiteo project

SHIRI is a research project funded by the Ile de France region as a Digiteo project which started on Oct. 1st 2007 and will last until Sept. 30th, 2011. It involves two partners of Digiteo, Supelec and the University of Paris-Sud. The aim of SHIRI is to design an annotation system to improve the relevance of the search on the Web when resources contain both semi-structured and textual data.

8.2. 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).

Particular projects that we conduct are detailed next.

8.2.1. 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.

8.3. Bilateral International Relations

8.3.1. Cooperation within Europe

Procope

Gemo has a PHC-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

Gemo has a PHC-Polonium project with the group of Lasota Slavomir at Warsaw University, Poland. The project will stop at the end of 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

Gemo has a PHC-Van-Gogh project with the group of Maarten Marx at Amsterdam University, The Netherlands. The project will stop at the end of 2007. Its goal is to work on expressive power and performances of XML query languages.

TARGET

Gemo started a cooperation with the Luxembourg University in November 2005 which lead to a PhD in co-tutelle with Paris-Sud university. The PhD project is TARGET for opTimal Adaptive infoRmation manaGemEnT over the web. It aims at improving web information retrieval by integrating web data evolution, users knowledge evolution and search domain evolution. The PhD student is Cedric Pruski.

University of Oxford

Gemo started a collaboration with Georg Gottlob from University of Oxford on the topic of the definition of the *Match* operator in data exchange. This collaboration led to a three-month stay of Pierre Senellart at University of Oxford.

8.3.2. Cooperation with Senegal

Gemo started a cooperation with the Gaston Berger University last year: a PhD in co-tutelle with Paris-Sud university started in december 2006. The subject of the thesis is the integration of semi-structured data for information retrieval. The PhD student is Mouhamadou Thiam.

8.3.3. Cooperation with the Middle-East

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

8.3.4. Cooperation with North America

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

8.3.5. 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 end 2008. Victor Vianu and Ravi Vijay, a Ph.d student from UCSD spent 3 months in Gemo this summer. Bogdan Cautis spent 1 week in San Diego. The home page of GemSaD can be found at http://www-rocq.inria.fr/~segoufin/GEMSAD/. GemSad is also partially supported by the National Science Foundation.

8.4. 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)
- Victor Vianu, professor, UC San Diego (July to September)
- Yuhong Yan, research officer, NRC Canada, IIT, Fredericton (in December)

The following PhD students came for internships in the group: Ravi Vijay [UCSD, USA; 2 months, PhD internship].

9. Dissemination

9.1. Thesis

The following PhD thesis were defended in 2007:

- Andrei Arion, XML Access Modules: Towards Physical Data Independence in XML Databases.
- Bogdan Cautis, Signing and Reasoning about Tree Updates.
- Antonella Poggi (with Universita' degli Studi di Roma "La Sapienza"), Structured and Semi-structured Data Integration.
- Fatiha Saïs, Semantic Data Integration guided by an Ontology.
- Mathias Samuelides, Tree walking automata with pebbles.
- Pierre Senellart, Understanding the Hidden Web.

9.2. Participation in Conferences

Gemo project members have co-chaired scientific events:

- S. Abiteboul has co-chaired the International Workshop on Data and Service Integration (SDIS'07), in cooperation with VLDB.
- I. Manolescu has co-chaired the 10th International Workshop on the Web and Databases (WebDB 2007), in cooperation with ACM SIGMOD.

Members of the project have participated in program committees:
S. Abiteboul

- World Wide Web Conference (WWW07)
- International Conference on Very Large Databases (VLDB'07)
- International Workshop on Web Information and Data Management (WIDM'07)
- World Wide Web Conference (WWW08)
- ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2008)
- Journées Francophones de Bases de Données Avancées 2007

P. Chatalic

- Journées Francophones de Programmation par Contraintes (JFPC 2007)

Ph. Dague

- 20th International Joint Conference on Artificial Intelligence (IJCAI) 2007
- 18th International Workshop on Principles of Diagnosis (DX) 2007
- 21th International Workshop on Qualitative Reasoning (QR) 2007

I. Manolescu

- 33rd Very Large Databases Conference (VLDB 2007)
- Conference on Information and Knowledge Management (CIKM 2007)
- Web Information and Data Management workshop (WIDM 2007), in cooperation with the CIKM conference
- Experimental Evaluation in Databases (ExpDB 2007) workshop, in cooperation with the ACM SIGMOD conference
- Journées Francophones de Bases de Données Avancées 2007

F. Goasdoué

- 16èmes congrès francophone Reconnaissance des Formes et Intelligence Artificielle (RFIA08)

C. Reynaud

- Third workshop on Context and Ontology Representation and Reasoning (C&O:RR-2007)
- 16èmes congrès francophone Reconnaissance des Formes et Intelligence Artificielle, member of the editorial board (RFIA08)
- Conférence Extraction et Gestion des Connaissances (EGC07)
- 17èmes Journées Francophones d'Ingenierie des connaissances (IC07)
- 1ères Journées Francophones sur les ontologies (JFO2007)
- Atelier Modélisation des connaissances (EGC07)
- Atelier Ontologies et Gestion de l'Hétérogénéité Sémantique (OGHS'07)
- Atelier Ontologies et Textes (TIA07)

M-C. Rousset

- International Joint Conference on Artificial Intelligence 2007
- International Semantic Web Conference 2007
- Atelier Modélisation des connaissances (EGC07)
- Atelier Modélisation des connaissances (EGC08)
- European Semantic Web Conference 2008

Fatiha Sais

- Manifestation des Jeunes Chercheurs en Sciences et Technologies de l'Information et de la Communication (MajecSTIC 2007)

L. Segoufin

- ACM Symposium on Principles of Database Systems (PODS'07)
- EACSL Conference for Computer Science Logic (CSL'07)

P. Senellart

- Text Mining Workshop 2007

L. Simon

- International Conference on Theory and Applications of Satisfiability Testing (SAT 2007)
- Journées Francophones de Programmation par Contraintes (JFPC 2007)

M. Vazirgiannis

- International Conference on User Modeling (UM 2007)

D. Vodislav

- Journées Francophones de Bases de Données Avancées 2007

9.3. Invited Presentations

Serge Abiteboul has been invited speaker at the Symposium on Theoretical Aspects of Computer Science, STACS 2007 [9]. He has been invited speaker at the PhD Student Workshop of SIGMOD 2007 where he spoke on "Life in Academia". He has been also invited at the Dagstuhl-Seminar on Programming Paradigms for the Web (2007).

Marie-Christine Rousset presented a tutorial at BDA 2007 on "Building scalable semantic peer-to-peer data management systems: the SomeWhere approach". She presented a lecture talk at the ACAI 2007 Summer School on "Logic-based techniques for information integration".

9.4. Scientific Animations

Editors

F. Goasdoué

- Guest editor of a special issue of Technique et Science Informatiques (TSI) on the Semantic Web, Hermès-Lavoisier.
- Member of the reading committee of the book Semantic Web Methodologies for E-Business Applications: Ontologies, Processes and Management Practices, Idea Group Publishing (scheduled for publication in 2008).

I. Manolescu

- Guest editor of a special issue of the Elsevier Journal of Information Systems on Performance Evaluation in Database Systems.

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.

10. Bibliography

Year Publications

Articles in refereed journals and book chapters

[1] P. ADJIMAN, F. GOASDOUÉ, M.-C. ROUSSET. *SomeRDFS in the Semantic Web* , in "Journal on Data Semantics", LNCS, vol. 8, 2007, p. 158-181.

[2] A. ARION, A. BONIFATI, I. MANOLESCU, A. PUGLIESE. *XQueC: A Query-Conscious Compressed XML Database*, in "ACM TOIT", vol. 7, n^{o} 2, 2007.

[3] C. DOULKERIDIS, V. ZAFEIRIS, K. NØRVÅG, M. VAZIRGIANNIS, E. A. GIAKOUMAKIS. *Context-based caching and routing for P2P web service discovery*, in "Distributed and Parallel Databases", vol. 21, n^{o} 1, 2007, p. 59-84.

[4] M. EIRINAKI, M. VAZIRGIANNIS. *Web site personalization based on link analysis and navigational patterns*, in "ACM Trans. Trans. Internet Techn.", vol. 7, n^{o} 4, 2007.

[5] N. GUELFI, C. PRUSKI, C. REYNAUD. *Towards the Adaptive Web using Metadata Evolution*, in "Handbook of research on Web Information Systems Quality", 2007.

[6] C. REYNAUD, B. SAFAR. *Techniques structurelles d'alignement pour portails web*, in "RNTI, Revue des Nouvelles Technologies de l'Information", 2007.

[7] L. SEGOUFIN. *Static Analysis of XML Processing with Data Values*, in "ACM Sigmod Record", vol. 36, n^{o} 1, 2007.

[8] P. SENELLART, V. D. BLONDEL. *Automatic discovery of similar words*, in "Survey of Text Mining: Clustering, Classification and Retrieval", M. W. BERRY, M. CASTELLANOS (editors), Second Edition, Springer-Verlag, January 2008.

Publications in Conferences and Workshops

[9] S. ABITEBOUL. *A Calculus and Algebra for Distributed Data Management*, in "Symposium on Theoretical Aspects of Computer Science (STACS)", 2007.

[10] S. ABITEBOUL, P. BOURHIS, B. MARINOIU. *Incremental View Maintenance for Active Documents*, in "National Conference, Bases de Données Avancées", 2007.

[11] S. ABITEBOUL, I. DAR, R. POP, G. VASILE, D. VODISLAV. *EDOS Distribution System: a P2P architecture for open-source content dissemination*, in "Int. Conf. on Open Source Systems", 2007.

[12] S. ABITEBOUL, I. DAR, R. POP, G. VASILE, D. VODISLAV. *Snapshot on the EDOS Distribution System*, in "Free and Open Source Software Developers' European Meeting", 2007.

[13] S. ABITEBOUL, I. DAR, R. POP, G. VASILE, D. VODISLAV, N. PREDA. *Large Scale P2P Distribution of Open-Source Software (demo)*, in "International Conference on Very Large Databases", 2007.

[14] S. ABITEBOUL, I. MANOLESCU, N. POLYZOTIS, N. PREDA, C. SUN. *XML Processing in DHT networks*, in "International Conference on Data Engineering", also in National Conference, Bases de Données Avancées 07, 2008.

[15] S. ABITEBOUL, I. MANOLESCU, S. ZOUPANOS. *OptimAX: optimizing distributed continuous queries (demo)*, in "National Conference, Bases de Données Avancées", 2007.

[16] S. ABITEBOUL, B. MARINOIU. *Distributed Monitoring of Peer to Peer Systems*, in "ACM Int.'l workshop on Web Information and Data Management", 2007.

[17] S. ABITEBOUL, B. MARINOIU. *Monitoring Peer to Peer Systems*, in "National Conference, Bases de Données Avancées", 2007.

[18] S. ABITEBOUL, B. MARINOIU, P. BOURHIS. *Distributed Monitoring of Peer to Peer Systems (demonstration)*, in "International Conference on Data Engineering", To appear, 2008.

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

[20] S. ABITEBOUL, R. POP, G. VASILE, D. VODISLAV. *Scalability Evaluation of a P2P Content Dissemination System*, in "National Conference, Bases de Données Avancées", 2007.

[21] A. ARION, V. BENZAKEN, I. MANOLESCU, Y. PAPAKONSTANTINOU. *Structured Materialized Views for XML Queries*, in "Very Large Databases Conference", 2007, p. 87-98.

[22] K. BERBERICH, S. J. BEDATHUR, G. WEIKUM, M. VAZIRGIANNIS. *Comparing apples and oranges: normalized pagerank for evolving graphs*, in "WWW 2007", ACM, May 2007, p. 1145-1146.

[23] B. BUTNARU, F. DRAGAN, G. GARDARIN, I. MANOLESCU, B. NGUYEN, R. POP, N. PREDA, L. YEH. *P2PTester: a tool for measuring P2P platform performance*, in "International Conference on Data Engineering", 2007, p. 1501-1502.

[24] F.-E. CALVIER, C. REYNAUD. *Découverte de correspondances entre ontologies distribuées*, in "Atelier Ontologies et Gestion de l'Hétérogenéite Sémantique, Plate-Forme AFIA 2007", 2007, p. 31-40.

[25] B. CAUTIS, S. ABITEBOUL, T. MILO. *Reasoning about XML Update Constraints*, in "ACM Conf. on Principles of Database Systems", 2007.

[26] C. DOULKERIDIS, A. VLACHOU, Y. KOTIDIS, M. VAZIRGIANNIS. *Peer-to-Peer Similarity Search in Metric Spaces* , in "VLDB 2007", VLDB Endowment, September 2007, p. 986-997.

[27] N. GUELFI, C. PRUSKI, C. REYNAUD. *Les ontologies pour la recherche ciblée d'information sur le Web*, in "18èmes Journées Francophones d'Ingénierie des Connaissances, IC'2007", 2007, p. 61-72.

[28] N. GUELFI, C. PRUSKI, C. REYNAUD. *Understanding and Supporting Ontology Evolution by Observing the WWW Conference*, in "Int. Workshop on Emergent Semantics and Ontology Evolution associated to ISWC2007", 2007.

[29] Y. LI, T. MELLITI, P. DAGUE. *Modeling BPEL Web services for diagnosis: towards self-healing Web services*, in "Proc. of the 3rd International Conference on Web Information Systems and Technologies (WEBIST'07), Barcelona, Spain", March 2007, p. 297-304.

[30] I. MANOLESCU, S. MANEGOLD. *Performance Evaluation and Experimental Assessment - Conscience or Curse of Database Research? (panel)*, in "VLDB", 2007, p. 1441-1442.

[31] D. MAVROEIDIS, M. VAZIRGIANNIS. *Stability Based Sparse LSI/PCA: Incorporating Feature Selection in LSI and PCA* , in "Machine Learning: ECML 2007", Springer, September 2007, p. 226-237.

[32] A. NASH, L. SEGOUFIN, V. VIANU. *Determinacy and Rewriting of Conjunctive Queries Using Views: A Progress Report*, in "International Conference on Database Theory (ICDT)", 2007, p. 59-73.

[33] Y. OLLIVIER, P. SENELLART. *Finding Related Pages Using Green Measures: An Illustration with Wikipedia*, in "Proc. AAAI, Vancouver, Canada", July 2007, p. 1427–1433.

[34] N. PERNELLE, F. SAÏS. *Passage à l echelle de la reconciliation de concepts et de la reconciliation de references : quelques points de comparaisons.*, in "Workshop DECOR : 'Passage à l'échelle des techniques de découverte de correspondances' of EGC'2007, Namur (Belgium)", 2007.

[35] C. REYNAUD, B. SAFAR. *Exploiting WordNet as Background Knowledge*, in "International ISWC'07 Ontology Matching (OM-07) Workshop, Busan, Corea", 2007.

[36] C. REYNAUD, B. SAFAR. *Utilisation de connaissances supplémentaires pour la découverte de mappings dans le système TaxoMap*, in "Atelier DECOR, EGC 2007", 2007.

[37] B. SAFAR, C. REYNAUD, F.-E. CALVIER. *Techniques d'alignement d'ontologies basées sur la structure d'une ressource complémentaire*, in "1ères Journées Francophones sur les Ontologies", 0ctober 2007, p. 21-35.

[38] M. SAMUELIDES, L. SEGOUFIN. *Complexity of Pebble Tree-Walking Automata*, in "Fundamentals of Computation Theory (FCT)", 2007, p. 458-469.

[39] F. SAÏS, N. PERNELLE, M.-C. ROUSSET. *Approche logique pour la réconciliation de références*, in "Actes of Extraction et Gestion des Connaissances (EGC 2007),Belgium", 2007, p. 623-634.

[40] F. SAÏS, N. PERNELLE, M.-C. ROUSSET. *L2R: a Logical method for Reference Reconciliation*, in "Proceedings of the Twenty-second AAAI Conference on Artificial Intelligence (AAAI-07)", 2007.

[41] F. SAÏS, N. PERNELLE, M.-C. ROUSSET. *Reconciliation de references : une approche adaptee aux grands volumes de données.*, in "Proceedings of the fifth Conference on Optimization and Information Systems (COSI), Algeria", 2007.

[42] K. SCHNAITTER, S. ABITEBOUL, T. MILO, N. POLYZOTIS. *On-Line Index Selection for Shifting Workloads*, in "International Workshop on Self-Managing Database Systems", 2007.

[43] P. SENELLART, S. ABITEBOUL. *On the Complexity of Managing Probabilistic XML Data*, in "Proc. PODS, Beijing, China", June 2007, p. 283–292.

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

[45] M. THIAM, N. PERNELLE, F. SAÏS. *WebdocEnrich : enrichissement semantique flexible de documents semi-structurés*, in "Actes of Extraction et Gestion des Connaissances (EGC 2007),Belgium", 2007.

[46] G. TSATSARONIS, M. VAZIRGIANNIS, I. ANDROUTSOPOULOS. *Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri* , in "IJCAI 2007", February 2007, p. 1725-1730.

[47] A. VLACHOU, C. DOULKERIDIS, Y. KOTIDIS, M. VAZIRGIANNIS. *SKYPEER: Efficient Subspace Skyline Computation over Distributed Data* , in "ICDE 2007", IEEE Computer Society, May 2007, p. 416-425.

[48] M. WEIS, I. MANOLESCU. *Declarative XML Data Cleaning with XClean*, in "Conference on Advanced Information Systems Engineering", 2007, p. 96-110.

[49] M. WEIS, I. MANOLESCU. *XClean in action (demo)*, in "Conference on Innovative Database Systems Research", 2007, p. 259-262.

[50] Y. YAN, P. DAGUE. *Monitoring and Diagnosing Orchestrated Web Service Processes*, in "Proc. of 5th IEEE International Conference on Web Services (ICWS'07), Salt Lake City, Utah, USA", IEEE Computer Society, July 2007, p. 51-59.

[51] H. ZARGAYOUNA, B. SAFAR, C. REYNAUD. *TaxoMap in the OAEI 2007 alignment contest*, in "Ontology Alignment Evaluation Inititiative (OAEI) 2007 Campaign - Workshop on Ontology Matching", 2007.

Internal Reports

[52] V. ARMANT. *Diagnostic distribué à base de modèles dans un système pair-à-pair*, Technical report, Computer Science Master report, University Paris-Sud, Orsay, 2007.

[53] L. YE. *Cooperative diagnosis for BPEL Web services*, Technical report, Computer Science Master report, University Paris-Sud, Orsay, 2007.

Miscellaneous

[54] D. COLAZZO, F.-X. DUDOUET, I. MANOLESCU, B. NGUYEN, P. SENELLART, A. VION. *Traiter des corpus d'information sur le Web. Vers de nouveaux usages informatiques de l'enquête*, 2007, Neuvième Congrès de l'Association Francaise de Sciences Politiques (AFSP).