INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Project-Team gemo*

# *Management of Data and Knowledge Distributed Over the Web*

*Saclay - Île-de-France*

Theme : Knowledge and Data Representation and Management

Activity Report

2009

# Table of contents

# 1. Team

**Research Scientist**

Serge Abiteboul [ DR-INRIA, HdR ]<br>
Ioana Manolescu [ CR-INRIA, HdR ]<br>
Laurent Romary [ DR-INRIA, HdR ]

**Faculty Member**

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 ]
Fatiha Saïs [ Assistant Professor, Univ. Paris 11 ]
Laurent Simon [ Assistant Professor, Univ. Paris 11 ]
Véronique Ventos [ Assistant Professor, Univ. Paris 11 ]

**Technical Staff**

Jesùs Camacho-Rodriguez [ from October ]
Julien Leblay [ July-September ]
Nobal Niraula
Mohamed Ouazara [ until September ]
Alin Tilea

**PhD Student**

Nada Abdallah [ Allocataire MENRT till September then ATER, Paris 11 ]
Vincent Armant [ Allocataire MENRT, Paris 11 ]
Fadia Azaiez [ Digiteo grant ]
Michel Batteux [ CIFRE with Sherpa Engineering, Paris 11 ]
Pierre Bourhis [ ENS Cachan ]
François Calvier [ Grant BDI CNRS till September then ATER, Paris 11 ]
Alban Galland [ X-Telecom, Paris 11 ]
Martin Goodfellow [ PhD in U. Strathclyde, UK, 6 months ]
Fayçal Hamdi [ ANR grant ]
Konstantinos Karanasos [ ANR grant ]
Asterios Katsifodimos [ Allocataire MENRT, Paris 11, from October ]
Evgeny Kharlamov [ Univ. Bolzano, 1 month ]
Wael Khemiri [ Allocataire MENRT, Paris 11 ]
Yingmin Li [ ANR grant ]
Bogdan Marinoiu [ Grant BDI CNRS, Paris 11, until June ]
Yassine Mrabet [ Digiteo contract ]
Cédric Pruski [ PhD in cotutelle between Luxembourg U. and Paris 11 ]
Mouhamadou Thiam [ PhD in cotutelle between Gaston Berger U. and Paris 11 ]
Lina Ye [ Allocataire MENRT, Paris 11 ]
Nadjet Zemirline [ Contract, Paris 11 ]
Spyros Zoupanos [ CORDI till October, then ATER, Paris 9 ]

**Post-Doctoral Fellow**

Gauvain Bourgne [ Post Doc fellowship, until February ]
Fanny Chevalier [ Digiteo grant ]
Yannis Katsis [ ERC grant, since September ]
Othman Nasri [ Post Doc fellowship ]

**Visiting Scientist**

Tarek Melliti [ Assistant Professor, Univ. Evry Val d'Essonne, scientific advisor ]
Angela Bonifati [ Researcher, CNR Italy, 1.5 month ]
George Katsirelos [ NICTA, visiting post-doc since September ]
Amélie Marian [ Assistant Professor, Rutgers U., 2 months ]
Neoklis Polyzotis [ Assistant Professor, U.C. Santa Cruz, 1 month ]
Yuhong Yan [ Assistant Professor, U. Concordia, Montreal, 1.5 month ]
Philippe Rigaux [ Professor, Univ. Paris 9, 12 months ]
Marie-Christine Rousset [ Professor, Univ. Grenoble, 12 months ]

**Administrative Assistant**

Marie Domingues [ ITA, until August ]
Céline Halter [ ITA, since September ]

# 2. Overall Objectives

## 2.1. Introduction

Gemo is a joint project with Laboratoire de Recherche en Informatique (UMR 8623 CNRS-University Paris-Sud), located in Orsay.

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

Chantal Reynaud, Brigitte Safar, Fayçal Hamdi and Haïfa Zargayouna have been awarded the EGC-Application award 2009 for the paper *Partitionnement d'ontologies pour le passage à l'échelle des techniques d'alignement*. The SAT solver developed by L. Simon and external colleagues has won the industrial UNSAT category of the 2009 SAT competition.

The team management has changed during the year. Until April, S. Abiteboul was team manager and I. Manolescu co-team manager. After that, I. Manolescu became manager by interim and C. Reynaud co-team manager.

# 3. Scientific Foundations

## 3.1. Scientific Foundations

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).

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

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 WebContent RNTL Project, we have developed a platform for the P2P management of warehouses of Web documents.

# 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.

AlignViz a visualization tool for alignments between ontologies.

GUNSAT a greedy local search algorithm for propositional unsatisfiability testing

Glucose a modern CDCL SAT solver that predicts learnt clause usefulness

KadoP a peer-to-peer platform for warehousing of Web resources.

LN2R a logical and numerical tool for reference reconciliation.

MESAM a plug-in for Protege 2000 to merge generic and specific models.

OptimAX an algebraic cost-based optimizer for ActiveXML.

PAP an alignment-based ontology partitioning system.

ReaViz a reactive, declarative workflow enactment tool.

SHIRI-Annot a tool to annotate semantically documents elements by exploiting both content and structure.

SHIRI-Querying a tool to approximate user's queries according to the SHIRI annotation design.

SomeWhere a P2P infrastructure for semantic mediation.

SomeWhere+ a P2P infrastructure tolerant to inconsistency.

SpyWhere a generator of mapping candidates for enriching peer ontologies.

TARGET a framework and a tool to search the Web by using adaptive ontologies.

TaxoMap a prototype to automate semantic mappings between taxonomies.

ViP2P a P2P XML data management platform based on materialized views.

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

# 6. New Results

## 6.1. Ontology-Based Information Retrieval

**Participants:** Nathalie Pernelle, Cédric Pruski, Chantal Reynaud, Mouhamadou Thiam, Yassine Mrabet.

### *6.1.1. Adaptive Ontologies for Web Information Retrieval*

We address the problem of taking knowledge evolution for improving Web search in the sense of the relevance of the returned results. The advocated solution is based on the use of ontologies, cornerstone of the Semantic Web, for representing both the domain targeted by the query and the profil of the user who submits the query. Ontologies are considered as knowledge that is evolving over time. In consequence, the ontology evolution problem has to be tackled as regards the evolution of the target domain but also with respect to the evolution of the user's profile. This work is the core of a PhD [3] which was defended in April 2009. We introduced a new paradigm : adaptive ontology as well as a process for making adaptive ontologies smoothly follow evolution of the domain. The so-defined model relies on the adaptation of ideas developed in the field of psychology and biology to the knowledge engineering field. We proposed an approach exploiting adaptive ontologies for improving Web Information Retrieval. To this end, we first introduced data structures, WPGraphs and W3Graphs, for representing Web data. We the introduce the ASK query language tailored for the extraction of relevant information from these structures. We also propose a set of query enrichment rules based on the exploitation of ontological elements as well as adaptive ontologies characteristics of the ontology representing the domain targeted by the query and the one representing the view of the user on the domain. Lastly, we have devised a tool for managing adaptive ontologies and for searching relevant information on the Web as well as experimental validation of the introduced concepts. We based our validation on the definition of a realistic case study devoted to the retrieval of scientific articles published at the International World Wide Web series of conference.

### *6.1.2. Semantic Annotation*

SHIRI is an ontology-based system for integrating semi-structured documents related to a specific domain. The system¿s purpose is to allow users to access to relevant parts of documents as answers to their queries. SHIRI uses RDF/OWL for representation of resources and SPARQL for their querying. It relies on an automatic, unsupervised and ontology-driven approach for extraction, alignment and semantic annotation of tagged elements of documents. We have developed and tested the SHIRI-Extract component, which exploits a set of named entity and term patterns to extract term candidates to be aligned with the ontology. It proceeds in an incremental manner in order to populate the ontology with terms describing instances of the domain and to reduce the access to extern resources such as Web. The results obtained on a HTML corpus related to call for papers in computer science show how that the number of terms (or named entities) aligned directly with the ontology increases as the method is applied [32]. For the *SHIRI*-querying, we have defined an order relation on the queries and validated it on two corpora [27].

## 6.2. Peer-to-Peer Inference Systems

**Participants:** Nada Abdallah, Vincent Armant, François Calvier, Philippe Chatalic, Philippe Dague, François Goasdoué, Chantal Reynaud, Laurent Simon.

A P2P inference system (P2PIS) is made of autonomous agents called peers. Each peer models its application domain using a knowledge base (KB) and peers having similar interests can establish semantic correspondences between their KBs called mappings. Mappings play a central role since on the one hand they define how the KBs of some peers integrate and on the other hand they give rise to a semantic network, the decentralized KB of the P2PIS, in which it becomes possible to reason. However, reasoning in the distributed logical setting of a P2PIS is not simple. Indeed, the challenge is to design decentralized reasoning algorithms with the purpose to reduce an inference task to perform on the KB of a P2PIS to a decentralized calculus among the peers, while none of them has a comprehensive view of the global KB.

### 6.2.1. Consequence Finding

In the last years, we have investigated the basic AI task of consequence finding in (possibly inconsistent) propositional P2PISs. Consequence finding consists of deriving theorems of interest that are intentionally characterized within a logical theory. It proves useful in many composite AI tasks such as common sense reasoning, diagnosis, or knowledge compilation. Most of our results have been implemented in the SomeWhere platform, the scalability of which has been demonstrated on synthetic data (up to a thousand of peers). It is worth noticing that our results allow one for the very first time to perform a truly decentralized consequence finding calculus in a distributed theory of propositional logic, i.e., without having a global view of that theory.

### 6.2.2. SAT Solving

Following our previous work, we identified a way of predicting learnt clause usefulness in modern SAT solvers. This work [16] allowed us to propose a new version of Minisat that won the industrial UNSAT category in the 2009 SAT Competition. This contest category is highly competitive.

### 6.2.3. P2P conservative extension checking

We have pointed out that the notion of non conservative extension of a knowledge base (KB) is important to the distributed logical setting of propositional P2PIS. It is useful to a peer in order to detect/prevent that a P2PIS corrupts (part of) its knowledge or to learn more about its own application domain from the P2PIS [5]. That notion is all the more important since it has connections with the privacy of a peer within a P2PIS and with the quality of service provided by a P2PIS. We have therefore studied the following tightly related problems from both the theoretical and decentralized algorithmic perspectives: (i) deciding whether a P2PIS is a conservative extension of a given peer and (ii) computing the witnesses to the corruption of a given peer's KB within a P2PIS so that we can forbid it.

### 6.2.4. P2P consistency checking and query answering

We have investigated a decentralized data model and associated algorithms for peer data management systems (PDMSs) based on the DL-LITE$_{\mathcal{R}}$ description logic [9], [10]. That logic is a fragment of the forthcoming W3C recommendation for the Semantic Web: OWL2. Our approach relies on reducing query reformulation and consistency checking for DL-LITE$_{\mathcal{R}}$ into reasoning in propositional logic. This enables a straightforward deployment of DL-LITE$_{\mathcal{R}}$ PDMSs on top of SomeWhere, our scalable peer-to-peer inference system for the propositional logic. We have also shown how to answer queries using views – predefined queries – in DL-LITE$_{\mathcal{R}}$ in the centralized and decentralized cases, by combining the query reformulation algorithm of DL-LITE$_{\mathcal{R}}$ and the state-of-the-art query rewriting algorithm: MiniCon.

### 6.2.5. Tools for experimental evaluation of P2PIS architectures

Because of the distributed nature of SomeWhere like P2PIS and their assynchronous mode of communication, performing large scale experiments with such architectures is a complex tasks. In order to alleviate this task, we are developping a set of tools developed in order to the automated deployment of such P2PIS on the Inria Grid'5000 plateform. SWTOOLS contains a generator of P2PIS instances, with random (but convincing) local theories and mappings. It allows for multi-cluster node reservation on the grid, the automatic deployement of peers on these nodes, query generation and dispatching on the network and automatic results collection on the peers.

### 6.2.6. Distributed Diagnosis

Research on consistency-based distributed diagnosis, set up in the framework of propositional P2PISs, pursued in 2009. It is in some sense a dual problem of consequence finding, as the algorithm developed is based on the distributed computation of prime implicants of the (unknown) global theory (so, not relying on a preliminary computation of conflicts as most of the diagnosis algorithms in the centralized case). We improved our first algorithm, which incrementally returns diagnoses by dynamically building a tree throughout the network. by taking advantage of a jointree structure. We have implemented an automated benchmark generator which builts peer-to-peer inference systems structured by social network topologies. Experimentations and comparison of the approaches are an on going work, as well as addressing scalability issues. We handle also privacy respect of the peers by allowing agents to reason “as much as possible” together while keeping their secret secret. Some privacy of the final diagnostic results is also studied. Up to now, the network is considered as static, i.e. the acquaintances of each peer are fixed. Next step will be to study dynamicity of the network, i.e. addressing departures and arrivals of peers.

This research is one of the topics of the submitted proposal of associated INRIA team Smarties with INRIA Rennes Dream group and NICTA Canberra (Australia).

### *6.2.7. Mapping distributed ontologies*

Our work takes place in the setting of the peer data management system (PDMS) SomeRDFS. Ontologies are the description of peers data. Peers in SomeRDFS interconnect through mappings which are semantic correspondences between their own ontologies. Thanks to its mapping a peer my interact with the others in order to answer a query. A crucial aspect in SomeRDFS is that peers are equivalent in functionalities. No peer has a global view of the data management system. Each peer has its own ontology, its own mappings and its own data. It ignores the ontology, the mappings and the data of the other peers. In this setting, our work aims at increasing the mappings of the peers in order to increase the quantity and the quality of the answers of the whole data management system. Previously, we proposed an approach to identify two kinds of mappings: mapping shortcuts corresponding to a composition of pre-existent mappings and mappings which can not be inferred from the network but yet relevant. We focused on the identification of the second kind of mappings. We updated our algorithms identifying relevant mapping candidates and proposed new filtering criteria to limit the matching process to a restricted set of elements. These criteria are relative to the peers involved in the mappings, the kind of mappings which are looked for (generalization or specialization mappings)or to the quality of the mappings. Then we focused on the alignment step. We proposed alignment techniques suitable to our context. We adapted terminological alignment techniques, proposed to integrate mechanisms based on query answering and techniques using the PDMS as an external source. Finally we investigated a methodology to help validating discovered mappings. These new propositions have been implemented in SpyWhere and experiments are currently conducted. This work is the core of a PhD which will be defended next year.

## 6.3. Thematic Web Warehousing

**Participants:** Serge Abiteboul, Hélène Gagliardi, Alban Galland, Fayçal Hamdi, Nobal Niraula, Nathalie Pernelle, Chantal Reynaud, Fatiha Saïs, Brigitte Safar.

### *6.3.1. Reference Reconciliation*

The reference reconciliation problem consists in deciding whether different identifiers refer to the same data (same person, same conference, ...). The logical and numerical approach named LN2R that we have developed has been detailed in [47]. This approach allows computing a set of reference pairs that (1) refers to the same data, (2) does not refer to the same data and (3) may refer to the same data. In addition to the reconciliation and no reconciliation decisions, the logical method L2R allows inferring the semantic equivalence of heterogeneous basic values that are stored in a dictionary. We are studying how this dictionary can be automatically refined in order to improve the reconciliation results in the settings of collaboration with THALES (in the HEDI project). In [45] we have shown how the reference reconciliation is used in a data warehouse building process, where data is extracted from the original sources, transformed in order to conform with the ontology and then reconciled by using the ontology. In order to enhance the user confidence in the results of data reconciliation methods that are numerical, global and ontology driven, we have proposed an explanation approach. In this approach, the explanations are computed and represented in colored Petri Nets.

### 6.3.2. Reference Fusion

The issue of data fusion arises once reconciliations have been determined. The objective of the fusion is to obtain a unique representation of the real world entity. We have proposed a fusion approach which deals with the uncertainty in the values associated with the attributes thanks to a formalism based on belief functions whose shapes are based on a set of criteria, using evidence theory formalism [31] has been developed. The aim now is to build a flexible querying approach of the fused data where the user preferences are taken into account.

### 6.3.3. Mapping between ontologies

We pursue our work on TaxoMap in the setting of the WebContent and Geonto projects. Following several issues previously investigated as the use of support knowledge published this year [8], we focused on two main points, alignment of very large ontologies and mapping refinement.

Very large ontologies have been built in some domains such as medecine or agronomy and the challenge now lays in scaling up alignment techniques that often perform complex tasks. We proposed two partitioning methods which have been designed to take the alignment objectives into account in the partitioning process as soon as possible. These methods transform the two ontologies to be aligned into two sets of blocks of a limited size. Furthermore, the elements of the two ontologies that might be aligned are grouped in a minimal set of blocks and the comparison is then enacted upon these blocks. We performed experiments with the two methods on various pairs of ontologies and results are promising. This work obtained the best application paper award at EGC2009 [19] and has been selected for an english book chapter [43].

We investigated mapping refinement because current ontology matchers are not efficient for all application domains or ontologies and very often the quality of the results could be improved by considering the specificities of the ontologies domain. We proposed an environment, called TaxoMap framework, based on TaxoMap, which helps an expert to specify treatments based on alignment results. The aim is to refine these results or to merge, restructure or enrich ontologies [28]. Currently, this approach has been applied to mapping refinement in the topographic field with the ANR project, GEONTO.

At the same time, developments on TaxoMap have been pursued. Terminological techniques have been improved with a better morpho-syntactic analysis and we introduced new structural techniques. This allowed us to participate for the third time in the international alignment contest OAEI2009 [18] which consists of applying matching systems to ontology pairs and evaluating their results. We took part to five tests and experimented our algorithm on large multilingual ontologies (English, French, German). Our participation in the campaign allows us to test the robustness of TaxoMap, our partitioning algorithms and new structural techniques.

We have also worked on the alignment of generic and specific models in the setting of Adaptive Hypermedia (AH) in order to help AH creators' models to be reused in a platform only made up of generic components. We developed a Protege Plug-in assisting designers to specialize generic models using their own models [36]. The plug-in includes two parts. A knowledge part gathers the meta-model based on the OWL meta-model and deduction rules. The processing part is made of components performing interactions with an inference engine (Jess) and the OWL Protege editor. We use the OWL protege API to manipulate OWL models and the SWRL Jess bridge to execute SWRL rules using the jess inference engine.

Finally, we initiated a work about global comparison of ontologies. The aim is to be able to assess and compare the points of view behind each particular conceptualization of the world. We studied which insights could be discovered from ontology matching, introducing initial ideas towards a global comparison of ontologies [48].

### 6.3.4. Integration of web resources

We investigated the integration of resources available on the Web in Adaptive Hypermedia Systems (AHS). More and more metadata describing resources are available on the Web using Semantic Web languages, and can be reused. Our aim is to build an open corpus AHS by, on one hand, reusing AHS technologies, particularly the adaptation engine which is the heart of these systems and, on the other hand, reusing resources and their descriptions which are available on the Web. Moreover, we want to allow the creator of an adaptive system not only to reuse adaptation strategies that come with the system, but to also be able to specify his own ones. For that, we propose a pattern-based approach to express adaptation strategies in a semi-automatic and simple way. It allows the creator of an adaptive system to define elementary adaptations by using and instantiating adaptation patterns. These elementary adaptations can then be combined, allowing to specify adaptation strategies in an easy and flexible manner. We distinguish adaptive navigation according to two main criteria: the selection operations performed in order to obtain resources being proposed to the user and the elements of the domain model involved in the selection process. We validated our approach using the GLAM adaptation engine. We show that the GLAM rules can be automatically generated from pattern-based adaptations.

In a separate research work, we have carried work on Liquid Queries, a new paradigm for querying combined Web information systems. A liquid query provides to the user an interface containing a certain number of attributes, but which can be modified if the user requires e.g. more attributes by joining with other, dynamically identified, data sources, aggregates the content on some criteria, omits some columns etc. Behind liquid queries stands the SeCo execution engine, a database-like system for joining independent Web information sources. This work is carried in collaboration with Politecnico di Milano [42].

### *6.3.5. Viewpoints corroboration*

We consider a set of views stating possibly conflicting facts. Negative facts in the views may come, e.g., from functional dependencies in the underlying database schema. We want to predict the truth values of the facts. Beyond simple methods such as voting (typically rather accurate), we explore techniques based on “corroboration”, i.e., taking into account trust in the views. We introduce three fixpoint algorithms corresponding to different levels of complexity of an underlying probabilistic model. They all estimate both truth values of facts and trust in the views. We present experimental studies on synthetic and real-world data. This analysis illustrates how and in which context these methods improve corroboration results over voting methods. We believe that corroboration can serve in a wide range of applications such as source selection in the semantic Web, data quality assessment or semantic annotation cleaning in social networks. This work sets the bases for a wide range of techniques for solving these more complex problems.

### *6.3.6. Social networks*

Use of the web to share personnal data is increasing rapidly with the emergence of Web 2.0 and social networks applications. However, users have yet to trust all the different hosts of their data and face difficulty with updates. To overcome this problem, we are studying on a model of distributed knolwedge base with access control and cryptographic functionalities,. The model allows exchanging documents, access control statements, keys and instructions in a distributed setting. We are considering different implementations of this model that can be used to leverage technologies such as DHT or Gossiping.

In such a social network, participants may bring conflicting opinions. We have studied the problem of trying to corroborate information coming from a very large number of participants. We have proposed and evaluated various algorithms towards this goal.

## 6.4. XML Query Optimization

**Participants:** Ioana Manolescu, Martin Goodfellow, Konstantinos Karanasos, Spyros Zoupanos.

### *6.4.1. XML storage and query optimization*

Work in this area is winding down. The main results concern contributions to the Springer Encyclopedia of Database Systems, published in 2009. Our contribution in this area include an algebra for XML query processing [44], a survey of XML storage techniques in relational databases [41].

### 6.4.2. Performance evaluation methodology

Performance evaluation is a natural component in many data-oriented works such as those carried on in Gemo. Intense discussion emerged within the international scientific data management community since 2006 as to what is the standard of proof for performance claims made in research papers, and how best to foster rapid prototyping and improvement of research platforms by sharing or at least facilitating the process of experimenting with somebody else's code. Gemo has been at the forefront of this effort, since I. Manolescu has been the first SIGMOD Repeatability chair (2008). This involvement has continued in 2009 where I. Manolescu has been the SIGMOD Repeatability and Workability co-chair (with S. Manegold from CWI Amsterdam). K. Karanasos has participated to the SIGMOD 2009 Repeatability and workability PC. Insights obtained from this effort have been made available to the community via a joint paper [51] and a tutorial [26].

### 6.4.3. Incremental maintenance of XML materialized view

Work has started together with A. Bonifati on the issue of incrementally maintaining tree pattern materialized views. The visit of M. Goodfellow in our group is centered around this topic.

## 6.5. XML Warehousing in P2P

**Participants:** Serge Abiteboul, Jesùs Camacho-Rodriguez, Ioana Manolescu, Konstantinos Karanasos, Asterios Katsifodimos, Alin Tilea, Spyros Zoupanos.

### 6.5.1. XML materialized views in P2P

This activity has strongly picked up in volume of work and in the number of people involved. Significant prototyping and experimentation was carried on within the VIP2P project (http://vip2p.saclay.inria.fr). ViP2P allows wide-scale sharing of XML documents based on a distributed hash table (or DHT). VIP2P peers independently publish (share) XML documents that others may query. Moreover, each peer may choose to materialize a given set of views, described by XML tree patterns. A query posed in the network of peers is thus first rewritten with the help of the views available in the network, then executed by exploiting these views [38], [39].

Many developments are ongoing around the ViP2P platform, as part of the ANR CODEX project. First, we consider the extension of the platform to handle XML documents with RDF annotations. The problem is much more complex than when considering XML documents only, since it requires detecting distributed value joins over XML tree pattern queries. Our approach to solve it is described in a technical report [49]. Second, we started to tackle the issue of adapting the materialized views in the ViP2P network to the needs (queries) issued by the various peers [50].

### 6.5.2. P2P XML indexing

We have pursued the work on the ViP2P peer-to-peer platform for building and managing warehouses of Web resources. Our previous work addressed the issue of indexing extensional XML data (trees). We are currently working on indexing also graph data, more precisely, XML documents with references to other documents or including function calls.

## 6.6. Monitoring and Web services

**Participants:** Serge Abiteboul, Michel Batteux, Gauvain Bourgne, Pierre Bourhis, Philippe Dague, Yingmin Li, Bogdan Marinoiu, Tarek Melliti, Othman Nasri, Lina Ye.

### 6.6.1. Error diagnosis and self-healing

The work devoted to self-healibility of Web services continued. Our initial model of conversationally complex Web services as Petri nets, with control and data places, enriched by data dependencies, has been extended in order to model directly semantic faults and their propagation inside the Petri net model itself. For this, faults inside places and transitions are introduced and their propagation is represented by using colored tokens (colors represent normal, faulty and unknown status of data values) and a color propagation function [21]. This ECPN (Enriched Colored Petri Net) model gives birth to a set of algebraic linear equations describing its behavior through the dynamic evolution of the markings. It is turned into a set of linear symbolic inequalities the solutions of which, in terms of color variables, express the minimal diagnoses. An effective backwards propagation algorithm has been designed to compute these solutions. This has been done first in the centralized case [20] and then extended to the decentralized case for cooperating choreographed BPEL Web services, where global diagnosis is achieved by a coordinator dialoguing with each local diagnoser [22]. A complete implementation has been realized, where the ECPN is automatically generated from the BPEL code and both local diagnosers and global coordinator are implemented as Web services too. Extension to a purely distributed framework achieving diagnosis by direct dialog between local diagnosers without coordinator will be studied.

### 6.6.2. Diagnosability

The aim of diagnosability is to ensure that a given partially observable system has the property that any fault (taken from a set of faults given a priori) will be detectable and identifiable in a bounded time after its occurrence. Work on diagnosability is led in the framework of discrete-event systems and has been conducted along two directions. First, in the DIAFORE project, we formalized diagnosability analysis, usually expressed in terms of automata in the literature, in terms of Input-Output Symbolic Transition Systems (IOSTS), allowing both the representation of the interaction of the system with its environment and a concise representation of the system's model. This study allowed the adaptation and use of the CEA Agatha tool of symbolic verification of conformity to formal specifications for checking diagnosability [17]. Second, as one of our objective is, as for diagnosis, to tackle diagnosability analysis for distributed systems where the global model is not known, distributed diagnosability of a given pattern (rational language defined by an automaton that describes the situation, the diagnosability of which we want to analyze, which is more general that a simple faulty transition) is being studied for distributed systems modeled as communicating local labeled transition systems [35], [33], [34]. This work is intended to be applied in particular to conversational Web services. The next step will consist in considering also distributed observation. We have also continued our work on computing minimal prefixes of Petri nets unfoldings for verifying diagnosability [24].

Obvious relationships exist between diagnosability analysis and verification and model checking. A collaboration with colleagues of the ForTesSE group of LRI is being launched in order to investigate the complementarity and combination of these methods, in particular between passive testing and diagnosability analysis, between the results of this analysis and the automatic generation of on-line diagnoser, and relationship with reconfiguration actions by automatic composition in the context of services. A co-supervised thesis will begin in January 2010.

All this is also part of the objectives of the Smarties associated team's proposal.

### 6.6.3. 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. We focused on a problem closely related to monitoring: 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 [12]. We have also addressed interesting issues that appeared in this context, like query satisfiability over active documents and stream relevance for considered queries [13].

# 7. Contracts and Grants with Industry

## 7.1. Industrial contracts

**Participant:** Serge Abiteboul.

Serge Abiteboul is a member of the advisory board of Thomson.

## 7.2. HEDI project

**Participants:** Fatiha Saïs, Nathalie Pernelle, Ioana Manolescu.

HEDI–Heterogeneous Electronic Data Integration project is a new collaboration agreement between Gemo team of INRIA Saclay and Thales Corporate Services started in September 2009. This project aims at designing a data reconciliation tool for electronic component descriptions.

## 7.3. RNTL Project WebContent

**Participants:** Serge Abiteboul, Fayçal Hamdi, Ioana Manolescu, Bogdan Marinoiu, Nobal Niraula, Mohamed Ouazara, Chantal Reynaud, Brigitte Safar, Spyros Zoupanos.

The WebContent project (http://www.webcontent.fr) ends in December 2009. 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. components. In the semantic enrichment of ontologies group (Lot 3) we tested the partitioning tool. We worked on a methodological guide for a joint use of the partitioning tool able to manage large ontologies and matching tools integrated in the WebContent platform. Experiments were also performed on the ontologies provided by AgroParistech, showing the capacity of TaxoMap, our alignment tool, to align large ontologies composed of more than 20000 concepts. In the peer-to-peer group (Lot 5), we have finalised the development of the peer-to-peer storage and query processing tool, and we have integrated the caching functionality developed by our partners. We have tested and fine-tuned our modules in particular on the EADS application data set.

## 7.4. ANR-PREDIT DIAFORE project

**Participants:** Gauvain Bourgne, Philippe Dague, Othman Nasri.

The DIAFORE ("DIAgnostic de FOnctions REparties") ADEME/ANR/PREDIT project which started on February 2006 in the framework of System@tic Paris-Région cluster, ended on September 2009. The project was coordinated by CEA LIST and involved Renault Trucks (Volvo group), Serma Ingénierie, UTC and University Paris-Sud. We pursued two objectives inside this project: embedded diagnosis of distributed electronic functions inside a vehicle (with the LFSE group of CEA) and diagnosability analysis (with the LISE group of CEA). The case study was the Smart Distance Keeping function and faults considered were sensor faults (wheels speed sensors, engine transmission speed sensor, radar). A demonstrator of diagnosability analysis on an abstract IOSTS model of SDK function was done with the CEA Agatha tool using the theoretical results of our study.

## 7.5. ANR-PCI COSMAT project

**Participant:** Laurent Romary.

COSMAT ("Service collaboratif de traduction automatique pour textes scientifiques", "Collaborative Machine Translation service for Scientific texts") is an ANR PCI ("Programme Contenus et Interactions") project which started on Oct. 1, 2009 and will end on Sept. 09, 2012. Funding for INRIA-Gemo is 77 kEuros. The project is coordinated by University of Maine and also involves Systran. The main objective for Gemo is to design an appropriate document model for scholarly papers based on the TEI (Text Encoding Initiative) guidelines, which may be used to exchange precise content between HAL (the French national publication repository) and the automatic translation tools that the two other partners will deploy for scientific texts. The textual model will be implemented as an export module for HAL, allowing also to provide a long-term archiving perspective for its content.

## 7.6. EU-eContent+ PEER project

**Participant:** Laurent Romary.

PEER ("Publishing and the Ecology of European Research ") is an EU eContent+ project which started on Sept. 1, 2008 and will last 36 months. Funding for INRIA is 247 kEuros. The project is coordinated by STM (International Association of Scientific, Technical and Medical Publishers) and involves the European Science Foundation, Göttingen State and University Library, and the Max Planck Society. The overall aim of the project is to develop an Observatory to monitor the effects of systematic archiving on publishing and the ecology of European research (PEER). INRIA (through Gemo and IT support group SEISM) is in charge of specifying and implementing the gateway through which scientific publishers deposit metadata and full text content of scholarly papers to be redirected to several trusted publication repositories. This involves in particular the mapping of publishers' data onto a standardized XML representation.

# 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; information visualization, J.-D. Fekete), with the Cedric Group at CNAM-Paris; some INRIA groups (Dahu, L. Segoufin, at INRIA-Saclay, Atlas, P. Valduriez and DistribCom, A. Benveniste, at INRIA-Bretagne; Exmo, J. Euzenat, at INRIA Rhone-Alpes; Mostrare at INRIA-Nord-Europe); 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. ANR CODEX*

Codex is a research project supported by the ANR Domaines Emergents call (2009-2012), coordinated by I. Manolescu (http://codex.saclay.inria.fr). The partners are: the MOSTRARE group of INRIA Lille (J. Niehren), the Database group from LRI (D. Colazzo, Université Paris Sud-11), the PPS Laboratory from University of Paris 7 (G. Castagna), the University of Blois (M. Halfeld-Ferrari), the University of Paris 1 (F. Gire), the WAM group of INRIA Rhône-Alpes, and the Innovimax start-up (M. Zergaoui). The project studies optimization, distributed processing and coordination, and adaptation techniques to cope with XML document dynamicity. Within CODEX we have spent significant effort developing and testing ViP2P, an XML content sharing platform which scales up to hundreds of peers [38], [38].

### 8.1.2. ANR Dataring

The DataRing project (2008-2011) lead by INRIA-Sophia (Patrick Valduriez) includes also U. Grenoble (Marie-Christine Rousset) and Teleco-ParisTech (Pierre Senellart). The project addresses the problem of P2P data sharing for online communities, by offering a high-level network ring across distributed data source owners. Users may be in high numbers and interested in different kinds of collaboration and sharing their knowledge, ideas, experiences, etc. Data sources can be in high numbers, fairly autonomous, i.e. locally owned and controlled, and highly heterogeneous with different semantics and structures. What we need then is new, decentralized data management techniques that scale up while addressing the autonomy, dynamic behavior and heterogeneity of both users and data sources.

### 8.1.3. ACI 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. It is finishing this year.

### 8.1.4. ANR GEONTO project

The objective of this ANR MDCO project (2008-2011), is to make data in the geographic domain inter-operate. We focus on two main goals. On one hand, we aim at integrating heterogeneous geographic databases using schema matching techniques. On the other hand, we aim at querying a large collection of textual documents which are more various and for a larger readership than databases just mentioned before. This project is a collaboration between COGIT-IGN (Sébastien Mustière), the IC3 group at IRIT - Université Paul Sabatier (Nathalie Aussenac) and the DESI group at LIUPPA - Université de Pau et des Pays de l'Adour (Mauro Gaio). The home page of the project could be found at: http://geonto.lri.fr.

### 8.1.5. ANR JCJC WebStand

The objective of this ANR is to analyze the problems surrounding the use of semi-structured databases in social sciences. This ANR regroups both computer science and sociology laboratories. The contract has ended in July 2009. The results of our work led to the publication of [29].

### 8.1.6. ANR UNLOC project

L. Simon is the coordinator of the ANR BLANC project about "incomplete search for UNSAT", which is an important theoretical and practical question. This projects started in early 2009 and includes phycisists from Orsay and researchers from Univ. Artois (CRIL, Lens), Univ. Amiens and Univ. Marseille.

### 8.1.7. Digiteo EDIFlow project

This project, led by I. Manolescu, is a collaboration between data management and information visualisation researchers (with V. Benzaken from LRI and J.-D. Fekete from INRIA AVIZ). The purpose is to study models and build a corresponding platform for efficient data-intensive workflow systems. A first prototype, ReaViz, has been developed on top of an industrial-strength database and demonstrated this year [37]. It enables the declarative specification of a workflow via an XML file, and automatically compiles it into a database application. The integration with the InfoViz toolkit developed in AVIZ remains to be done.

### 8.1.8. Digiteo SHIRI 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.

### 8.1.9. FRAE SIRASAS project

The SIRASAS (“Stratégies Innovantes et Robustes pour l’Autonomie des Systèmes Aéronautiques et Spatiaux") project, funded by the FRAE (“Fondation de Recherche pour l’Aéronautique et l’Espace"), has for partners: IMS Bordeaux (coordinator), SATIE-ENS Cachan, LAAS-CNRS Toulouse, CRAN Nancy, LRI Paris-Sud, ONERA Toulouse, CNES Toulouse, Airbus Toulouse and Thales Alenia Space Cannes. It started on October 2007 and will end on September 2010, but the effective participation of Gemo is concentrated during the second half part. We work on space applications: on one side on the case study which is a rendezvous mission between two spacecrafts in Mars orbit with the task of detecting and isolating nozzles failures; on the other side on a more prospective topic which is studying reconfiguration task and its coupling with diagnosis in an autonomy context.

### 8.1.10. Participation to evaluation committees

P. Dague has participated to the evaluation committees of the thematic ANR program ARPEGE (“Systèmes Embarqués et Grandes Infrastructures”) and of the non thematic ANR programs “Blanc” and “Jeunes Chercheuses et Jeunes Chercheurs”. C. Reynaud has participated to the evaluation committee of the ANR program CONTINT and of the CAPERS-COFECUB program.

## 8.2. European Commission Financed Actions

### 8.2.1. Webdam ERC Grant

The Webdam grant (S. Abiteboul) started in December 2008. The goal to develop a formal model for Web data management. This model will open new horizons for the development of the Web in a well-principled way, enhancing its functionality, performance, and reliability. Specifically, the goal is to develop a universally accepted formal framework for describing complex and flexible interacting Web applications featuring notably data exchange, sharing, integration, querying and updating. We also propose to develop formal foundations that will enable peers to concurrently reason about global data management activities, cooperate in solving specific tasks and support services with desired quality of service.

The Webdam project is shared between the Dahu and Gemo project-teams, both from INRIA Saclay.

## 8.3. Bilateral International Relations

### 8.3.1. Cooperation within Europe

I. Manolescu is a member of the Advisory Board of the SeCo (Search Computing) ERC project, headed by S. Ceri from Politecnico di Milano.

Close links exist with University of Madrid (A. Gomez-Perez), University of Manchester (I. Horrocks), University of Rome (M. Lenzerini).

### 8.3.2. Cooperation with Senegal

Gemo started a cooperation with the Gaston Berger University in december 2006 which leads to a PhD in co-tutelle with Paris-Sud university. 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

Gemo had for many years an Associated Team with the data management group at the University of California at San Diego (V. Vianu, A. Deutch, Y. Papakonstantinou). After this association sponsored by INRIA International and the National Science Foundation, completed in 2008, the two teams continued to closely cooperate.

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

### 8.3.5. Cooperation with Australia

The planning and diagnosis group of NICTA's Canberra Research Laboratory visited Gemo in July, as well as INRIA Rennes Dream group. A synthesis workshop with the three partners was organized at Gemo in July in order to prepare the draft of an INRIA associated team's proposal. The proposal of the associated team Smarties ("Self-healing highly dynamical networks: Facing the Smart Grid challenge") was submitted by the three partners on October. The scientific challenges are: self-helability analysis, interleaving diagnosis and repair, use of symbolic techniques and application to smart grid.

### 8.3.6. Cooperation with Japan

A two days workshop was organized at Gemo (with colleagues from AMIB team) in September with the research group of Katsumi Inoue, from NII. We plan to collaborate on: consequence finding and hypothesis finding, SAT, distributed reasoning handling knowledge sharing, extension to formalisms beyond propositional logic and privacy issues. Both Philippe Dague and Katsumi Inoue were invited at the JST-ANR French Japanese Workshop in the field of Information and Communication Science and Technologies organized by ANR at Paris in November. They intend to submit a project at this JST-ANR call in January 2010.

## 8.4. Visiting Professors and Students

This year the following professors visited Gemo:

- Sihem Amer-Yahia, Yahoo Research (1 month)
- Angela Bonifati, researcher, CNR Italy (1.5 months)
- Amélie Marian, assistant professor at Rutgers University (2 months)
- Neoklis Polyzotis, professor at UC Santa Cruz (1 month)
- Philippe Rigaux, professor at Paris 9 (12 months)
- Marie-Christine Rousset, professor at the University of Grenoble (12 months)
- Dimitri Theodoratos, professor at New Jersey Institute of Science and Technology (1 month)
- Yuhong Yan, associate professor at Concordia University, Montréal (1.5 months)

# 9. Dissemination

## 9.1. Thesis

The following HDR (*Habilitation à Diriger les Recherches*) was defended in 2009:

- Ioana Manolescu [1], *Optimization techniques for XML query processing*

The following PhD thesis were defended in 2009:

- Bogdan Marinoiu [2], *Analysis and verification of distributed systems.*
- Cédric Pruski [3], *Une approche adaptative pour la recherche d'information sur le Web.*
- Spyros Zoupanos [4], *Efficient peer-to-peer data management*

The following members of the group participated to HDR and PhD committees:

- Serge Abiteboul: reviewer of the Habilitation thesis of Véronique Cortier (Univ. Nancy)
- Philippe Dague: member of the HDR committees of Annelyse Thévenin (Univ. Paris-Sud 11), Pauline Ribot (Univ. Toulouse III) and Arpad Rimmel (Univ. Paris-Sud 11)
- François Goasdoué: reviewer of the thesis of Antoine Zimmerman (Univ. Grenoble 1) and member of the PhD committee of Hanen Belhaj Frej (Univ. Paris Sud-11).
- Ioana Manolescu: reviewer of the PhD thesis of Yann Gripay (INSA Lyon) and Loredana Afanasiev (Univ. Amsterdam).
- Chantal Reynaud: member of the PhD committees of Bastien Rance (Univ. Paris Sud-11) and Rim Jedidi (Univ. Paris Sud-11), reviewer of the PhD thesis of Abdeltif Elbyed (Telecom & Management SudParis). Member of the HDR committee of Ladjel Bellatrèche (Univ. Poitiers)

## 9.2. Participation in Conferences

S. Abiteboul has been the general program chair of the Very Large Database Conference, Lyon, 2009.

I. Manolescu has been a ACM SIGMOD 2009 Repeatability and Workability co-chair, and a co-chair of the Web Engineering track of the WWW 2009 conference.

L. Romary has been the poster and demonstration co-chair of ECDL 2009 conference.

L. Simon has been the co-chair of the SAT 2009 competition, as a technical advisor.

Members of the project have participated in program committees:
P. Chatalic

- Journées Francophones de Programmation par Contraintes (JFPC 2009)
- Modèles Formels de l'Interaction (MFI 2009)

P. Dague

- 20th International Workshop on Principles of Diagnosis (DX) 2009
- 23th International Workshop on Qualitative Reasoning (QR) 2009
- IJCAI Workshop on Self and Autonomous Systems: reasoning and integration challenges (SAS) 2009
- Journées Nationales de l'IA Fondamentale (IAF) 2009

F. Goasdoué

- International Joint Conference on Artificial Intelligence (IJCAI) 2009

I. Manolescu

- ACM SIGMOD International Conference on the Management of Data 2009
- International Conference on Web Engineering 2009
- Journées en Bases de Données Avancées (BDA) 2009

N. Pernelle

- Atelier EvalECD Evaluation des méthodes d'Extraction des Connaissances dans les Donnés (EGC) 2009
- Atelier QDC Qualité des Données et des Connaissances(EGC) 2009

C. Reynaud

- International Conference on Web and Information Technologies (ICWIT) 2009
- Congrès francophone Reconnaissance des Formes et Intelligence Artificielle (RFIA)
- Conférence Extraction et Gestion des Connaissances (EGC) 2009
- 20èmes Journés Francophones d'Ingénierie des Connaissances (IC) 2009
- 3èmes Journés Francophones sur les Ontologies (JFO) 2009
- 2nd International Workshop on REsource Discovery (in conjunction with VLDB) 2009
- Atelier EvalECD Evaluation des méthodes d'Extraction des Connaissances dans les Donnés (EGC) 2009
- Atelier QDC Qualité des Données et des Connaissances (EGC) 2009
- Atelier Construction d'Ontologies : vers un guide de bonnes pratiques (Plate-forme AFIA) 2009

L. Romary

- Workshop on Linguistic Processing Pipelines 2009
- First International Workshop On Spoken Dialogue Systems Technology (IWSDS) 2009
- 13th workshop in the SemDial series (DiaHolmia) 2009
- 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL) 2009
- 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL) 2009

B. Safar

- 3èmes Journées Francophones sur les Ontologies (JFO) 2009

F. Saïs

- (co-chair) of Atelier EvalECD Evaluation des méthodes d'Extraction des Connaissances dans les Donnés (EGC) 2009
- Student workshop of the ACM international conference on Management of Emergent Digital Ecosystems (MEDES-SW) 2009
- Colloque sur L'Optimisation et les Systèmes d'Informations (COSI) 2009

## 9.3. Invited Presentations

Serge Abiteboul had an invited presentation at Time'09 [11].

Philippe Chatalic had a seminar on "Peer to Peer Infrerence Systems" at IRIT, Univ. Paul Sabatier, Toulouse.

I. Manolescu had an invited presentation at a joint session among the XSym and DBLP 2009 workshops, in conjunction with VLDB [25].

L. Romary had the following invited talks: "The role of standards in the LT community - an ISO perspective", at the Kyoto workshop; "Services for the eHumanities", at the Bielefeld conference; "Stabilizing knowledge through standards - a perspective for the humanities, Going Digital, Evolutionary and Revolutionary Aspects of Digitization", at the Nobel Symposium; "Standardization initiatives for language resources - A modeling perspective" at the Workshop on standards for phonological corpora; "Normes: convergences!", at the LexiPraxi.

## 9.4. Scientific Animations

**Editors**

S· Abiteboul is a member of the steering committee of Proceedings of the VLDB Endowment (PVLDB) Journal, a journal that started in 2008.
F. Goasdoué

- Guest editor of a special issue of Technique et Science Informatiques (TSI) on the Semantic Web, Hermès-Lavoisier, February 2009.

I. Manolescu

- Member of the reading committee of the book Search Computing Challenges and Directions, Springer.

C. Reynaud

- Member of the reading committee of the book Ontology Theory, Management and Design: Advanced Tools and Models, IGI Publisher.
- Member of the reading committee for the DKE Special Issue on Ontologies in Designing Advanced Information Systems, Elsevier.
- Revue Information - Interaction - Intelligence (RI3).

L. Romary

- Editorial Review Board of the International Journal of Digital Library Systems (IJDLS)
- ISO committee TC 37/SC 4 chairman
- TEI council chairman
- Member of DHO (Digital Humanities Observatory, Dublin (IR)) International Advisory Board
- Member of FIZ Karlsruhe (DE Scientific committee)
- Member of Textgrid project (BMBF, DE) advisory committee, 2007
- Member of Nestor project (DFG, DE) advisory committee
- Member of High Level Expert Group on Digital Libraries (European Commission, DG SINFO)
- High-Level Experts Group on Scientific Data (European Commission, DG SINFO)
- Member of the Working Group on Research Infrastructures in Social and Human Sciences of the German Council of Science and Humanities (Wissenschaftsrat)

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.

### 9.4.1. Presentations to a larger public

S. Abiteboul participated to the Téléphone Sonne on France Inter. He published interviews or articles in Science et Vie, Le Nouvel Economiste and L'informaticien. He participated as panelist to the colloquium "Une histoire de DIM, Domaines d'Intérêt Majeur" organized par by the Île-de-France région. He had presentations at the colloquium INRIA "Web et Industrie" in Lille, at "Demain, la République du Web : une utopie ?" at La Cantine, and at "Perspectives IT pour le Codir", of the OLG Clubs, at La Vilette.

Ioana Manolescu presents Gemo work on ViP2P on december 3, 2009, in an XQuery Meetup at La Cantine.

# 10. Bibliography

## Year Publications

### Doctoral Dissertations and Habilitation Theses

[1] I. MANOLESCU. *Optimization techniques for efficient XML data management*, Université Paris Sud-11, october 2009, HDR thesis.

[2] B. MARINOIU. *Analysis and verification of distributed systems*, Université de Paris Sud, june 2009, Ph. D. Thesis.

[3] C. PRUSKI. *Une approche adaptative pour la Recherche d'Information sur le Web*, Université de Paris Sud, april 2009, Ph. D. Thesis.

[4] S. ZOUPANOS. *Efficient peer-to-peer data management*, Université Paris-Sud 11, december 2009, Ph. D. Thesis.

### Articles in International Peer-Reviewed Journal

[5] N. ABDALLAH, F. GOASDOUÉ. *Non-conservative Extension of a Peer in a P2P Inference System*, in "AI Communications: The European Journal on Artificial Intelligence (AICOM)", 2009.

[6] S. ABITEBOUL, N. POLYZOTIS. *Searching Shared Content in Communities with the Data Ring*, in "IEEE Data Eng. Bull.", vol. 32, n$^{o}$ 2, 2009, p. 44-51 US .

[7] C. REYNAUD, B. SAFAR. *Construction automatique d'adaptateurs guidée par une ontologie pour l'intégration de sources et de données XML*, in "Technique et Science Informatiques (TSI)", vol. 28, 2009, p. 199 - 228, http://hal.inria.fr/inria-00432426/en/.

[8] B. SAFAR, C. REYNAUD. *Alignement d'ontologies basé sur des ressources complémentaires Illustration sur le système TaxoMap*, in "Technique et Science Informatiques (TSI)", vol. 28, 2009, p. 1213 - 1234.

### International Peer-Reviewed Conference/Proceedings

[9] N. ABDALLAH, F. GOASDOUÉ, M.-C. ROUSSET. *DL-LITER in the Light of Propositional Logic for Decentralized Data Management*, in "Internal Joint Conference on Artificial Intelligence (IJCAI)", 2009, p. 2010-2015.

[10] N. ABDALLAH, F. GOASDOUÉ, M.-C. ROUSSET. *Gestion décentralisée de données en DL-lite*, in "Congrès francophone de Reconnaissance des Formes et Intelligence Artificielle (RFIA)", 2009, p. 2010-2015.

[11] S. ABITEBOUL, P. BOURHIS, A. GALLAND, B. MARINOIU. *The AXML Artifact Model*, in "International Conference on Temporal Representation and Reasoning (TIME)", 2009.

[12] S. ABITEBOUL, P. BOURHIS, B. MARINOIU. *Efficient maintenance techniques for views over active documents*, in "EDBT", 2009, p. 1076-1087.

[13] S. ABITEBOUL, P. BOURHIS, B. MARINOIU. *Satisfiability and relevance for queries over active documents*, in "PODS", 2009, p. 87-96.

[14] S. ABITEBOUL, O. GREENSHPAN, T. MILO, N. POLYZOTIS. *MatchUp: Autocompletion for Mashups (demo)*, in "ICDE", 2009, p. 1479-1482 US IL .

[15] G. AUDEMARD, M. S. MODELIAR, L. SIMON. *Pourquoi les solveurs SAT modernes se piquent-ils contre des cactus ?*, in "Cinquièmes Journées Francophones de Programmation par Contraintes, Orléans, juin 2009, France", 06 2009, p. 245-255, http://hal.archives-ouvertes.fr/hal-00390919/en/.

[16] G. AUDEMARD, L. SIMON. *Predicting Learnt Clauses Quality in Modern SAT Solver*, in "Twenty-first International Joint Conference on Artificial Intelligence (IJCAI'09), Pasadena États-Unis d'Amérique", 07 2009, http://hal.inria.fr/inria-00433805/en/.

[17] G. BOURGNE, P. DAGUE, F. NOUIOUA, N. RAPIN. *Diagnosability of Input Output Symbolic Transition Systems*, in "1st International Conference on Advances in System Testing and Validation Lifecycle (VALID)", 2009 JP .

[18] F. HAMDI, B. SAFAR, N. NIRAULA, C. REYNAUD. *TaxoMap in the OAEI 2009 alignment contest*, in "The Fourth International Workshop on Ontology Matching, Chantilly, Washington DC. États-Unis d'Amérique", 2009, http://hal.inria.fr/inria-00432622/en/.

[19] F. HAMDI, B. SAFAR, H. ZARGAYOUNA, C. REYNAUD. *Partitionnement d'ontologies pour le passage à l'échelle des techniques d'alignement*, in "9eme Journées Francophones "Extraction et Gestion des Connaissances", Strasbourg France", Cepadues, 2009, http://hal.inria.fr/inria-00432551/en/.

[20] Y. LI, T. MELLITI, P. DAGUE. *A colored Petri nets model for diagnosing data faults of BPEL services*, in "20th International Workshop on Principles of Diagnosis (DX)", 2009.

[21] Y. LI, T. MELLITI, P. DAGUE. *A colored Petri nets model for the diagnosis of semantic faults of BPEL services*, in "International Workshop on Petri Nets and Software Engineering (PNSE)", 2009.

[22] Y. LI, T. MELLITI, L. YE, P. DAGUE. *A decentralized model-based diagnosis for BPEL services*, in "21st International Conference on Tools with Artificial Intelligence (ICTAI)", 2009.

[23] P. LOPEZ, L. ROMARY. *Multiple Retrieval Models and Regression Models for Prior Art Search*, in "CLEF 2009 Workshop, Corfu Grèce", 2009, 18p., http://hal.archives-ouvertes.fr/hal-00411835/en/DE.

[24] A. MADALINSKI, F. NOUIOUA, P. DAGUE. *Diagnosability verification with Petri net unfoldings*, in "13th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES)", 2009.

[25] I. MANOLESCU. *XML processing in DHT networks*, in "XSym/DBPL workshops", 2009, Joint invited talk of the two workshops.

[26] I. MANOLESCU, S. MANEGOLD. *Performance Evaluation in Database Research: Principles and Experience (tutorial)*, in "International Conference on Extending Database Technologies (EDBT)", 2009 NL .

[27] Y. MRABET, N. PERNELLE, N. BENNACER, M. THIAM. *Aggregative and Neighboring Approximations to Query Semi-Structured Documents*, in "Extraction et gestion des connaissances, Strasbourg France", RNTI E-15 (editor), Cépaduès, 2009, p. pp. 469-470, ISBN 978-2-85428-878-0.

[28] S. MUSTIÈRE, N. ABADIE, N. AUSSENAC- GILLES, M.-N. BESSAGNET, M. KAMEL, E. KERGOSIEN, C. REYNAUD, B. SAFAR. *GéOnto : Enrichissement d'une taxonomie de concepts topographiques*, in "Spatial Analysis and GEOmatics Sageo 2009, Paris France", 2009, http://hal.inria.fr/inria-00432628/en/.

[29] B. NGUYEN, A. VION, I. MANOLESCU, D. COLAZZO, F.-X. DUDOUET, P. SENELLART. *The WebStand Projet*, in "WebSci'09: Society On-Line", 2009.

[30] L. ROMARY. *ODD as a generic specification platform*, in "TEI 2009 Conference", 2009.

[31] D. SEBASTIEN, F. SAÏS, R. THOMOPOULOS. *Fusion évidentielle de références et interrogation flexible*, in "Rencontres francophones sur la Logique Floue et ses Applications, Annecy France", 11 2009, p. 15-22, http://hal.inria.fr/inria-00433011/en/.

[32] M. THIAM, N. BENNACER, N. PERNELLE, M. LÔ. *Incremental Ontology-Based Extraction and Alignment in Semi-Structured Documents*, in "20th International Conference DEXA (Database and Expert Systems Applications) 2009, Linz Autriche", Springer, 2009, p. pp. 211-218, http://hal-supelec.archives-ouvertes.fr/hal-00423575/en/.

[33] L. YE, P. DAGUE, Y. YAN. *A distributed approach for pattern diagnosability*, in "20th International Workshop on Principles of Diagnosis (DX)", 2009.

[34] L. YE, P. DAGUE, Y. YAN. *An incremental approach for pattern diagnosability in distributed discrete event systems*, in "21st International Conference on Tools with Artificial Intelligence (ICTAI)", 2009 CA .

[35] L. YE, P. DAGUE. *Diagnosability of patterns in distributed discrete event systems*, in "7th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes(SAFEPROCESS)", 2009.

[36] N. ZEMIRLINE, Y. BOURDA, C. REYNAUD, F. POPINEAU. *MESAM: A Protégé Plug-in for the Specialization of Models*, in "International Protege Conference, Amsterdam Pays-Bas", 2009, http://hal.inria.fr/inria-00432999/en/.

### Workshops without Proceedings

[37] I. MANOLESCU, W. KHEMIRI, V. BENZAKEN, J.-D. FEKETE. *Reactive workflows for visual analytics*, in "Journées Bases de Données Avancées, Belgique Naumur", B. AMANN (editor), 2009, http://hal.inria.fr/inria-00425666/en/.

[38] I. MANOLESCU, S. ZOUPANOS. *Vues matérialisées XML pour des entrepôts de données pair-à-pair*, in "Journées Bases de Données Avancées, Belgique Naumur", B. AMANN (editor), 2009, http://hal.inria.fr/inria-00425642/en/.

[39] I. MANOLESCU, S. ZOUPANOS. *XML materialized views in P2P networks*, in "Fourth International Workshop on Database Technologies for Handling XML Information on the Web, Russie Saint Petersburg", 2009, http://hal.inria.fr/inria-00425627/en/.

### Scientific Books (or Scientific Book chapters)

[40] S. ABITEBOUL, O. BENJELLOUN, T. MILO. *Active XML*, in "Encyclopedia of Database Systems", Springer US, 2009, p. 38-41 US IL .

[41] D. Barbosa, P. Bohannon, J. Freire, C.-C. Kanne, I. Manolescu, V. Vassalos, M. Yoshikawa. *XML Storage*, in "Encyclopedia of Database Systems", L. Liu, T. Ozsu (editors), Springer, 08 2009, http://hal.inria.fr/inria-00433434/en/CAUSGRJP.

[42] A. Bozzon, M. Brambilla, S. Ceri, P. Fraternali, I. Manolescu. *Liquid Queries and Liquid Results in Search Computing*, in "Search Computing Challenges and Directions", M. Brambilla, S. Ceri (editors), Springer, 2009, to appear IT .

[43] F. Hamdi, B. Safar, C. Reynaud, H. Zargayouna. *Alignment-based Partitioning of Large-scale Ontologies*, in "Advances in Knowledge Discovery And Management", H. Briand, F. Guillet, G. Ritschard, D. Zighed (editors), Springer, 2009, http://hal.inria.fr/inria-00432606/en/.

[44] I. Manolescu, Y. Papakonstantinou, V. Vassalos. *XML Tuple Algebra*, in "Encyclopedia of Database Systems", L. Liu, T. Ozsu (editors), Springer, 2009, p. 3640-3646, http://hal.inria.fr/inria-00431395/en/GRUS.

[45] C. Reynaud, N. Pernelle, M.-C. Rousset, B. Safar, F. Saïs. *Data Extraction, Transformation and Integration Guided by an Ontology*, in "Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction, Advances in Data Warehousing and Mining Book Series", L. Bellatreche (editor), IGI Global, 2009, http://hal.inria.fr/inria-00432585/en/.

[46] L. Romary. *Questions & Answers for TEI Newcomers*, in "Jahrbuch für Computerphilologie 10", Jahrbuch für Computerphilologie, Mentis Verlag, 2009, -, http://hal.archives-ouvertes.fr/hal-00348372/en/.

[47] F. Saïs, N. Pernelle, M.-C. Rousset. *Combining a Logical and a Numerical Method for Data Reconciliation*, in "Journal on Data Semantics", LNCS, Springer Berlin / Heidelberg, 06 2009, p. 66-94, http://hal.inria.fr/inria-00433007/en/.

### Research Reports

[48] S. Mustière, C. Reynaud, B. Safar, N. Abadie. *Same words? Same worlds? Comparing ontologies underlying geographic data*, Laboratoire de Recherche en Informatique - LRI - CNRS : UMR8623 - Université Paris Sud - Paris XI, 2009, http://hal.inria.fr/inria-00432827/en/, Interne.

### Other Publications

[49] K. Karanasos, I. Manolescu. *P2P Views Over Annotated Documents*, 2009, http://hal.inria.fr/inria-00433474/en/, Technical report.

[50] A. Katsifodimos, I. Manolescu, A. Tilea, S. Zoupanos. *Adaptive distributed XML views*, 2009, http://hal.inria.fr/inria-00433481/en/, Technical report.

[51] S. Manegold, I. Manolescu, L. Afanasiev, J. Feng, G. Gou, M. Hadjieleftheriou, S. Harizopoulos, P. Kalnis, K. Karanasos, D. Laurent, M. Lupu, N. Onose, C. Ré, V. Sans, P. Senellart, T. Wu, D. Shasha. *Repeatability & Workability Evaluation of SIGMOD 2009*, 2009, http://hal.inria.fr/inria-00433458/en/, To appear in SIGMOD RecordNLUSCNSAAT.