Activity Report 2012

Team OAK

Optimizations and Architectures for Complex large data

RESEARCH CENTER
Saclay - Île-de-France

THEME
Knowledge and Data Representation and Management

Table of contents

Team OAK

Keywords: Data Management, Reasoning, Semantics, Web, Cloud Computing, Distributed System

Creation of the Team: April 01, 2012 .

1. Members

Research Scientists

Ioana Manolescu [Team Leader, Senior Researcher, Inria, HdR]
Serge Abiteboul [Senior Researcher, Inria, HdR]

Faculty Members

Nicole Bidoit [Professor, Univ. Paris-Sud, HdR]
Dario Colazzo [Associate Professor, Univ. Paris-Sud, HdR]
François Goasdoué [Associate Professor, Univ. Paris-Sud, HdR]
Melanie Herschel [Associate Professor, Univ. Paris-Sud]

External Collaborators

Philippe Rigaux [Professor, CNAM, HdR]
Marie-Christine Rousset [Professor, Univ. Grenoble, HdR]
Virginie Thion-Goasdoué [Associate Professor, CNAM]

Engineers

Andrés Aranda_Andujar
André De_Amorim_Fonseca [until October]
Tushar Ghosh

PhD Students

Mohamed-Amine Baazizi [Contrat Doctoral Univ. Paris-Sud]
Jesùs Camacho_Rodriguez [Contrat Doctoral Univ. Paris-Sud]
Konstantinos Karanasos [ANR grant, Univ. Paris-Sud, till September]
Asterios Katsifodimos [Contrat Doctoral Univ. Paris-Sud]
Julien Leblay [Contrat Doctoral Univ. Paris-Sud]
Noor Malla [Grant of Syrian government, Univ. Paris-Sud]
Alexandra Roatis [Digiteo grant, Univ. Paris-Sud]
Aikaterini Tzompanaki [Contrat Doctoral Univ. Paris-Sud, since October]
Federico Ulliana [Contrat Doctoral Univ. Paris-Sud]
Stamatios Zampetakis [Allocataire CORDI, Inria, since October]

Post-Doctoral Fellows

Zoi Kaoudi
Marina Sahakyan [Grant ANR Codex]

Administrative Assistants

Céline Halter [ITA, until June]
Alexandra Merlin [ITA, interim since July]

2. Overall Objectives

2.1. Highlights of the Year

Our best results of the year appeared in extremely visible and selective venues: automated recommendation of materialized XML views in ACM SIGMOD conference [18], XML query-update independence [6] and RDF materialized view selection in the VLDB 2012 conference, and scalable duplicate detection in IEEE TKDE [8].

On the national scientific stage, our team has invested significant effort in the recently accepted LabEx DigiCosme proposal, where I. Manolescu is coordinating the “Scalable and secure data management” task, and in the national database conference where I. Manolescu has been the Program Committee chair, while Nicole Bidoit and François Goasdoué were part of the Program Committee.

Significant prototype development effort was invested in particular leading to the ACM CIKM Amada [10] and Nautilus [15] software demonstrations.

3. Scientific Foundations

3.1. Efficient XML and RDF data management

The development of Web technologies has led to a strong increase in the number and complexity of the applications which represent their data in Web formats, among which XML (for structured documents) and RDF (for Semantic Web data) are the most prominent. Oak has carried on research on algorithms and systems for efficiently processing expressive queries on such Web data formats. We have considered the efficient management of XML and RDF data, both for query evaluation and for efficiently applying updates, possibly in concurrence with queries. We have also started investigating multidimensional data analysis within RDF data warehouses.

For applications that integrate such Web data from various sources, we developed efficient and effective techniques to automatically recognize multiple representations of the same real-world object. That is, we devised main-memory resident algorithms that apply on hierarchical data [8] as well as algorithms that manipulate graph data leveraging off-the-shelf database management systems and parallelization to address both efficiency and scalability beyond main-memory [7].

3.2. Cloud-based Data Management

We have recently started to work on the efficient management of complex Web data, in particular structured XML documents and Semantic Web data under the form of RDF, in a cloud-based data management platform. We have investigated architectures and algorithms for storing Web data in elastic cloud-based stores and building an index for such data within the efficient key-value stores provided by off-the-shelf cloud data management platform. We have devised and prototyped such platforms for both XML and RDF data, and started experimenting with them in the Amazon Web Services (AWS) platform [13], [12], [10].

3.3. Data Transformation Management

With the increasing complexity of data processing queries, for instance in applications such as relational data analysis or integration of Web data (e.g., XML or RDF) comes the need to better manage complex data transformations. This includes systematically verifying, maintaining, and testing the transformations an application relies on. In this context, Oak has focused on verifying the semantic correctness of a declarative program that specifies a data transformation query, e.g., an SQL query. To this end, we have investigated how to leverage data provenance (the information of the origin of data and the query operators) for query debugging. More specifically, we developed and implemented algorithms to explain unexpected results produced by a query (why-provenance) as well as expected results that are however missing from the query result (why-not provenance). Results have been presented in form of a software prototype [15].

4. Application Domains

4.1. Online content management

This concerns archiving filtered content from online information sources (journals, blogs, ...) with the purpose of recording their perspective on facts involving specific countries, regions or enterprises, key actors etc. We have recently explored such an application in a demonstration proposal, built using our XR model for XML documents with semantic annotations (currently under evaluation). The benefit of XR was to simplify and streamline the specification of document- and semantics-rich data management applications. Support for documents allows to keep an evidence (source) for facts and statements that may be extracted from them. At the same time, support for semantics allows to connect documents to people, ideas, trends etc. and to reason across data sources [9].

4.2. Open data intelligence

Open data intelligence: the goal is to build and efficiently exploit warehouses of Open Data, integrated from several data sources on the Web, in order to produce consolidated rich information to be given to decision makers and to the citizens. Such projects have been started in France notably by the city areas of Rennes (pioneer in Open Data usage), Paris, and more recently by Grenoble, in a project to which we participate; the ICT Labs DataBridges also focused on such topics. Oak competencies required for such projects are related to large-scale RDF data management as well as to the design of innovative data models for semantic-rich content.

4.3. Efficient complex data management in the cloud

In a cloud environment, data catalogs and indices need to be efficiently built and maintained, typically in parallel; queries need to be routed to only those data subsets which are likely to lead to results, and efficiently executed, using parallelism and the available indexes. We work on such topics within the Europa ICT Labs activity. Algorithms for efficiently handling indexes and views in the cloud for heterogeneous, complex-structure data are also at the core of our work proposal in the Datalyse project (see below). Our 2012 work in this context has lead to [10], [12], [13].

4.4. 360 degree customer view

Companies seek to gather as much information as possible about their customers, for instance for more targeted publicity campaigns, market analysis, offer personalization, etc. That is, they want to consider data beyond the data they collected about their customers in their proprietary databases. Complementary data include for instance social data extracted from customers' activities in social networks, public data related to their place of residence (e.g., crime rate or housing price evolution). To achieve this goal, data integration on highly heterogeneous and massive data is necessary. Furthermore, as one means to assess both the correctness of the integration result and the quality (in this application most notably trustworthiness), we can resort to data provenance. These topics are explored in the project Datalyse which we submitted in 2012 to the French national "AAP Cloud 3: Big Data" call, a project headed by the "Business & Decision" company and whose ongoing evaluation will continue in early 2013.

5. Software

5.1. Amada

Name: Amada (https://team.inria.fr/oak/amada/)

Contact: Jesús Camacho-Rodríguez (jesus.camacho-rodriguez@inria.fr))

Other contacts: Zoi Kaoudi (zoi.kaoudi@inria.fr), Ioana Manolescu (ioana.manolescu@inria.fr), Dario Colazzo (dario.colazzo@lri.fr), François Goasdoué (fg@lri.fr)

Presentation: A platform for Web data management in the Amazon cloud

5.2. Nautilus Analyzer

Name: Nautilus Analyzer (http://nautilus.saclay.inria.fr/)

Contact: Melanie Herschel (melanie.herschel@lri.fr)

Other contacts: n.a.

Presentation: A tool for analyzing and debugging SQL queries using why-provenance and why-not provenance.

5.3. RDFViewS

Name: RDFViewS (http://tripleo.saclay.inria.fr/rdfvs/)

Contact: Konstantinos Karanasos (konstantinos.karanasos@inria.fr)

Other contacts: François Goasdoué (fg@lri.fr), Julien Leblay (julien.leblay@inria.fr), and Ioana Manolescu (ioana.manolescu@inria.fr)

Presentation: a storage tuning wizard for RDF applications

5.4. ViP2P

Name: ViP2P (views in peer-to-peer, http://vip2p.saclay.inria.fr)

Contact: Ioana Manolescu (ioana.manolescu@inria.fr)

Other contacts: Jesús Camacho_Rodriguez (jesus.camacho-rodriguez@inria.fr), Asterios Katsifodimos (asterios.katsifodimos@inria.fr), Konstantinos Karanasos (konstantinos.karanasos@inria.fr)

Presentation: a P2P platform for disseminating and querying XML and RDF data in large-scale distributed networks.

5.5. XUpOp

Name: XUpOp (XML Update Optimization)

Contact: Dario Colazzo (colazzo@lri.fr)

Other contacts: Nicole Bidoit (bidoit@lri.fr), Marina Sahakian (Marina.Sahakyan@lri.fr), and Mohamed Amine Baazizi (baazizi@lri.fr)

Presentation: a general purpose type based optimizer for XML updates

5.6. XUpIn

Name: XUpIn (XML Update Independence)

Contact: Federico Ulliana (Federico.Ulliana@lri.fr)

Other contacts: Dario Colazzo (colazzo@lri.fr), Nicole Bidoit (bidoit@lri.fr)

Presentation: an XML query-update independence tester

5.7. XUpTe

Name: XUpTe (XML Update for Tempora documents)

Contact: Dario Colazzo (colazzo@lri.fr)

Other contacts: Nicole Bidoit (bidoit@lri.fr), Mohamed-Amine Baazizi (amine.baazizi@gmail.com)

Presentation: a type-based optimizer for representing and updated XML temporal sata

5.8. XPUQ

Name: XPUQ (XML Partitioning for Updates and Queries)

Contact: Dario Colazzo (colazzo@lri.fr)

Other contacts: Nicole Bidoit (bidoit@lri.fr), Noor Malla (noorwm@hotmail.com)

Presentation: a static analyzer and partitioner for XML queries and updates

6. New Results

6.1. Efficient XML and RDF data management

6.1.1. Efficient and safe management of XML and JSON data

We addressed the problem of detecting independence between XML queries and updates. Since the problem is undecidable for XQuery queries and updates, and is intractable even for restricted fragments, we adopted an approximating technique based on a schema-based static analysis. Our analysis turned to be precise and, at the same time, fast to run. Main result about this research line have been published in [6], while the complete study is reported in Federico Ulliana's PhD Thesis (defended in December 12) [5].

To address the problem of manipulating large XML documents via main-memory XQuery engines, largely used for their efficiency and easiness of integration in a programming environment, we developed partitioning techniques for both XQuery queries and updates. Our technique is based on a static analysis over queries and updates (no schema is used) able to infer information that is used to partition the input document, in a streaming fashion. Besides allowing existing main-memory system to scale up in terms of query/update input size, our technique also admits a MapReduce implementation. Main results have been published in [11], while the complete study is reported in Noor Malla's PhD Thesis (defended on September 21) [3].

We also tackled the problem of safe manipulation of JSON data. Some typed and MapReduce-based programming languages for manipulating JSON data have been recently proposed. However, the problem of inferring a schema for untyped JSON data was still open, and having a schema for manipulated data is fundamental for the afore mentioned programming languages. We started investigating technique able to deal with massive JSON data sets. To ensure efficiency, our technique is based on Map-Reduce, while to ensure precision and conciseness it adopts type rewriting rules able to: i) compact as much as possible intermediate inferred types, and ii) to avoid gross approximation when compacting types. Some preliminary results are quite encouraging, and appeared in [21].

6.1.2. Hybrid models for XML and RDF

Considerable energy is spent towards enriching XML data on the web with semantics through annotations. These annotations can range from simple metadata to complex semantic relationships between data items. Although the vision of supporting such annotations is spreading, it still lacks the infrastructure that will enable it. To this end we have proposed a framework enabling the storage and querying of annotated documents. We have introduced (i) the XR data model, in which annotated documents are XML documents described by RDF triples and (ii) the query language XRQ to interrogate annotated documents through their structure and their semantics. A prototype platform XRP for the management of annotated documents has also been developed, to show the relevance of our approach through experiments [9].

6.1.3. RDF query answering

A promising method for efficiently querying RDF data consists of translating SPARQL queries into efficient RDBMS-style operations. However, answering SPARQL queries requires handling *RDF reasoning*, which must be implemented outside the relational engines that do not support it. We have introduced the *database (DB) fragment of RDF*, going beyond the expressive power of previously studied RDF fragments. Within this fragment, we have devised novel *sound and complete techniques* for answering *Basic Graph Pattern (BGP)* queries, exploring the two established approaches for handling RDF semantics, namely reformulation and saturation. In particular, we have focused on handling database *updates* within each approach and proposed a method for incrementally maintaining the saturation; updates raise specific difficulties due to the rich RDF semantics. Our techniques have been designed to be deployed on top of any RDBMS(-style) engine, and we have experimentally studied their performance trade-offs [20], [14], [25].

6.1.4. Efficient and scalable Web Data Entity Resolution

We addressed the problem of detecting multiple heterogeneous representations of a real-world object (often referred to as record linkage, duplicate detection, or entity resolution) in two contexts, i.e., for hierarchical data and for data where relationships between entities form a graph.

Concerning XML entity resolution, we contributed to a novel algorithm that uses a Bayesian network to determine the probability of two XML elements being duplicates. The probability is based both on content and on structure information given by the hierarchical XML model. To efficiently evaluate the Bayesian network to find duplicates, we devised two pruning techniques. Whereas the first is lossless in terms of not loosing any true duplicates, the second pruning heuristic trades off runtime for a somewhat lower accuracy of the duplicate detection result. An experimental evaluation shows that the proposed solutions are capable of outperforming other state-of-the art XML duplicate detection methods [8].

As for duplicate detection in entity graphs, we defined a general framework for algorithms tackling this problem. The general process consists of three steps, namely retrieval, classification, and update. We further proposed an algorithm complying to the framework that leverages an off-the-shelf relational database to store and to efficiently query information (both data and relationships) relevant for duplicate classification. We further extended our framework and algorithm to allow for parallel and batched processing. Our experimental validation on data of up to two orders of magnitude larger than data considered by other state-of-the-art algorithms showed that the proposed methods allow to scale duplicate detection in entity graphs to large volumes of data [7].

6.1.5. Warehousing RDF data

Data warehousing (DW) research has lead to a set of tools and techniques for efficiently analyzing large amounts of multi-dimensional data. As more data gets produced and shared in RDF, analytic concepts and tools for analyzing such irregular, graph-shaped, semantic-rich data are needed. We have introduced *the first all-RDF model for warehousing RDF graphs*. Notably, we have defined *RDF analytical schemas*, themselves full RDF graphs, and *RDF analytical queries*, corresponding to the relational DW star/snowflake schemas and cubes. We have shown how *RDF OLAP operations* can be performed on our RDF cubes. We have also performed experiments validating the practical interest of our approach.

6.2. Cloud-based Data Management

We investigate architectures for storing Web data (in particular, XML documents and RDF graphs) based on commercial cloud platforms. In particular, we have developed the AMADA platform, which operates in a Software as a Service (SaaS) approach, allowing users to upload, index, store, and query large volumes of Web data. Since cloud users support monetary costs directly connected to their consumption of cloud resources, we focus on indexing content in the cloud. We study the applicability of several indexing strategies, and show that they lead not only to reducing query evaluation time, but also, importantly, to reducing the monetary costs associated with the exploitation of the cloud-based warehouse [10], [12], [13].

6.3. Data Transformation Management

When developing data transformations – a task omnipresent in applications like data integration, data migration, data cleaning, or scientific data processing – developers quickly face the need to verify the semantic correctness of the transformation. Declarative specifications of data transformations, e.g. SQL or ETL tools, increase developer productivity but usually provide limited or no means for inspection or debugging. In this situation, developers today have no choice but to manually analyze the transformation and, in case of an error, to (repeatedly) fix and test the transformation.

The above observations call for a more systematic management of a data transformation. Within Oak, we have so far focused on the first phase of the process described above, namely the analysis phase. Leveraging results obtained in previous years (by us and others), we solidified the theory of why-not provenance. Analogously to a distinction between different types of why-provenance, we defined three types of why-not provenance. For each of the three types, we surveyed the semantics employed by different approaches, e.g., set vs. bag semantics or existential vs. universal quantification. We also identified cases of implication and equivalence between why-not provenance of different types. We have leveraged this theoretical work during the design of a novel algorithm that has the potential to overcome usability and efficiency limitations of previous algorithms after further optimization, implementation, and validation in the future. Furthermore, we implemented different approaches for why-provenance and why-not provenance and included them in the Nautilus Analyzer, a system prototype for declarative query debugging. We demonstrated this prototype at CIKM 2012 [15].

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

A collaboration grant is ongoing with DataPublica, which started based on our common work on Linked Data for Digital Cities.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. DW4RDF

This Digiteo DIM LSC (*Logiciels et Systèmes Complexes*) project has started in October 2011. The aim is to design and implement data warehouse-style models and technologies for RDF data. This project supports the PhD scholarship of A. Roatis.

8.2. National Initiatives

8.2.1. ANR

The ANR Codex project (Coordination, dynamicity and efficiency for XML, 2009-2012) has ended; the final review has taken place in Lyon in January 2012. The project was coordinated by Ioana Manolescu; Nicole Bidoit, Dario Colazzo and François Goasdoué also participated.

The ANR DataBridges project (Data integration for digital cities, 2011-2012) has ended; the final review has taken place in Paris in September 2012. The project was coordinated by Ioana Manolescu; François Goasdoué also participated.

The ANR ConnectedCities project (Clouds for digital cities, 2011-2012) has ended; the final review has taken place in Paris in September 2012. Dario Colazzo, François Goasdoué and Ioana Manolescu have participated to the project.

The ANR DataRing project (Massive data management in peer-to-peer, 2009-2012) has ended; the final review has taken place in Lyon in January 2012. Ioana Manolescu has participated to the project.

8.3. European Initiatives

8.3.1. Collaborations in European Programs, except FP7

Program: KIC EIT ICT Labs

Project acronym: DataBridges

Project title: Data Integration for Digital Cities

Duration: January 2012 - December 2012

Coordinator: Ioana Manolescu

Other partners: Université Paris Sud (France), Technical University of Delft (The Netherlands), DFKI (Germany), Aalto University (Finland), KTH (Sweden), Alcatel-Lucent Bell Labs (France), DataPublica (France)

Abstract: DataBridges work focuses on two main topics: (i) the interoperability, enrichment and personalization of data, e.g. data on the cultural activities within a city, based on user profiles; (ii) efficient techniques for large-scale RDF data management, to be applied (among others) on digital city data.

Program: KIC EIT ICT Labs

Project acronym: Europa

Project title: Efficient cloud-based data management

Duration: January 2012 - December 2012

Coordinator: Volker Markl (Technical Univ. Berlin)

Other partners: Université Paris Sud (France), Technical University of Delft (The Netherlands), DFKI (Germany), Aalto University (Finland), SICS (Sweden)

Abstract: Europa aims at developing techniques for large-scale efficient data management based on a cloud (massively parallel) processing paradigm. Within Europa, we have finalized the Amada platform, and our ongoing work focuses on an algebraic translation framework from XQuery into PACT programs. PACT is the parallel data processing language proposed by the Berlin partner.

8.3.2. Collaborations with Major European Organizations

Partner 1: organisme 1, labo 1 (pays 1)

Sujet 1 (max. 2 lignes)

Partner 2: organisme 2, labo 2 (pays 2)

Sujet 2 (max. 2 lignes)

8.4. International Research Visitors

8.4.1. Visits of International Scientists

We have been visited by:

- Prof. Paolo Atzeni (Università Roma Tré), in June
- Prof. Alin Deutsch (UCSD, USA) in June-July (Digiteo invited scientist)
- Prof. Evi Pitoura (University of Ioannina, Greece), in October
- Prof. Vassilis Christophides (FORTH, Greece) in December
- Prof. Themis Palpanas (University of Trento, Italy) in December
- Prof. Yanlei Diao (U. Massachussets at Amherst, USA) in December

8.4.1.1. Internships

Three students visited the team within the Inria Internship program: Karan Aggaral, Abishek Choudhary and Kuldeep Reddy.

9. Dissemination

9.1. Scientific Animation

9.1.1. Participation in Editorial Boards

M. Herschel is the guest editor of the special issue on Data Integration of the journal it-Information Technology, Volume 54, Issue 3, 2012.

I. Manolescu is the editor in chief of the ACM SIGMOD Record, an associate editor of the ACM Transactions on Internet Technologies, and a member of the "Experiments and Analysis" track of PVLDB.

9.1.2. Participation in Conference Organization

Members of the project have been chairs of scientific events:
Ioana Manolescu

- Bases de Données Avancées (BDA) 2012
- IEEE International Conference on Data Engineering (ICDE) 2012, "Semistructured Data, XML and RDF" track
- Extending Database Technologies (EDBT) 2012 conference, Tutorial track
- Co-chair of the VLDB 2012 PhD workshop

Members of the project have participated in program committees:
Nicole Bidoit

- Bases de données Avancées (BDA) 2012

Dario Colazzo

- IEEE International Conference on Data Engineering (ICDE) 2012

François Goasdoué

- Bases de Données Avancées (BDA) 2012
- IEEE International Conference on Tools with Artificial Intelligence (ICTAI) 2012

Melanie Herschel

- IEEE International Conference on Data Engineering (ICDE) 2012
- Conference on Very Large Databases (VLDB) 2012
- VLDB PhD Workshop 2012
- Bases de Données Avancées (BDA) 2012

Ioana Manolescu

- ACM SIGMOD Conference 2012
- IEEE International Conference on Data Engineering (ICDE), "Cloud, Data Warehousing, and Large Data" track
- Data Analytics in the Cloud Workshop, in collaboration with EDBT/ICDT 2012
- Non-conventional Data Access Workshop, in collaboration with the ER conference 2012
- Cloud Intelligence Workshop, in collaboration with VLDB 2012

Members of the project participate in steering committees:
Nicole Bidoit

- Ecole thématique Masse de données Distribuées, Aussois, 27 mai - 1er juin 2012

Ioana Manolescu

- Workshop on Open Data (WOD) 2012

9.1.3. Invited Presentations

Ioana Manolescu gave a keynote presentation titled “Triples with a purpose” at the Semantic Search Workshop (SSW) next to the VLDB Conference 2012.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence, Nicole Bidoit, Bases de données, 25.5h éq. TD, L3, Université Paris-Sud, France

Master, Nicole Bidoit, SGBD relationnels: implémentation, 18 éq. TD, M2, Université Paris-Sud, France

Master, Nicole Bidoit, Bases de données et Systèmes d’Information, 27 éq. TD, M2, Université Paris-Sud, France

Master, Nicole Bidoit, Mise à Niveau en Informatique - Bases de Données, 40 éq. TD, M1, Université Paris-Sud, France

Master, Nicole Bidoit, Base de données Avancées, 30 éq. TD, M1, Université Paris-Sud, France

Master, Nicole Bidoit, Données et connaissances pour le WEB, 32 éq. TD, M2, Université Paris-Sud, France

Master : Dario Colazzo, SGBD relationnels: tuning d’applications , 21h éq. TD, M2, Université Paris-Sud, France

Master : Dario Colazzo, Bases de données , 21h éq. TD, M2, Université Paris-Sud, France

Master : Dario Colazzo, Systèmes de Gestion de Bases de Données, 54h éq. TD, Polytech, Université Paris-Sud, France

Master : Dario Colazzo, Base de Données, 21h éq. TD, Polytech, Université Paris-Sud, France

Master : Dario Colazzo, Base de Données Avancées, 31h éq. TD, Polytech, Université Paris-Sud, France

Master : Dario Colazzo, Mise à niveau bases de données, 17h éq. TD, M1, Université Paris-Sud, France

Master : Dario Colazzo, Gestion des données sur Internet, 18h éq. TD, M1, Université Paris-Sud, France

Master : Dario Colazzo, Base de Données , 20h éq. TD, Polytech, Université Paris-Sud, France

Master : Dario Colazzo, Tutorat d’apprentis, 20h eq. TD, Polytech, Université Paris-Sud, France

Licence : François Goasdoué, Bases de données, 62,5h éq. TD, L3, Université Paris-Sud, France

Master : François Goasdoué, Web Sémantique, 74h éq. TD, M2, Université Paris-Sud, France

Master : François Goasdoué, Données et connaissances pour le Web, 7,5h éq. TD, M2, Université Paris-Sud, France

Master : François Goasdoué, Modèles de raisonnement distribué, 4,5h éq. TD, M2, Université Paris-Sud, France

Master : François Goasdoué, Décision et raisonnement, 9h éq. TD, M2, Université Paris-Dauphine, France

Master : François Goasdoué, Ontologies et Web Sémantique, 13,5h éq. TD, M2, Université Paris-Dauphine, France

Master: Melanie Herschel, Entrepôts de données et requêtes OLAP, 99.5h éq. TD, M2, Université Paris-Sud, France

Master: Melanie Herschel, Intégration de données et Web sémantique, 4,5h éq. TD, M2, Université Paris-Sud, France

Master : Ioana Manolescu, Données Sémi-Structurées et XML, 18h éq. TD M2, Université Paris-Sud, France

Master : Ioana Manolescu, Services Web, 27 éq. TD, M2, Université Paris-Dauphine, France

9.2.2. Supervision

PhD & HdR :

HdR : François Goasdoué, Knowledge Representation meets DataBases for the sake of ontology-based data management [1], Univ. Paris-Sud, 11/07/2012

PhD : Mohamed-Amine Baazizi, Satic Analysis for the Optimization of Temporal XML document Updates, Univ. Paris-Sud, 07/09/2012, Nicole Bidoit et Dario Colazzo

PhD : Noor Malla, Partitioning XML data, towards distributed and parallel management, Univ. Paris-sud, 21/09/2012, Nicole Bidoit et Dario Colazzo

PhD : Federico Ulliana, Types for Detecting XML Query-Update Independence, Univ. Paris-Sud, 12/12/2012, Nicole Bidoit et Dario Colazzo

PhD : Konstantinos Karanasos, View-Based Techniques for the Efficient Management of Web Data [2], Univ. Paris-Sud, 29/06/2012, François Goasdoué and Ioana Manolescu

PhD in progress : Asterios Katsifodimos, Efficient Distributed Views for XML Data Management, Univ. Paris-Sud, 29/06/2012, Ioana Manolescu

PhD in progress : Julien Leblay, Efficient management of annotated documents, 01/10/2010, François Goasdoué and Ioana Manolescu

PhD in progress : Alexandra Roatis, Traitement efficace de requêtes SPARQL avec extensions OLAP pour entrepôts RDF, 01/09/2011, Dario Colazzo, François Goasdoué and Ioana Manolescu

PhD in progress: Aikaterini Tzompanaki, Foundations and Algorithms to Compute the Provenance of Missing Data, 1/11/2012, Melanie Herschel and Nicole Bidoit

PhD in progress : Stamatis Zampetakis, Scalable algorithms for cloud-based semantic web data management, 15/10/2012, François Goasdoué and Ioana Manolescu

9.2.3. Juries

- Nicole Bidoit, PhD defense committee of François Hantry, “Problèmes basés sur les Causes dans le cadre de la Logique Temporelle Linéaire : Théorie et Applications”, Univ. Lyon 1, 17/09/2012.
- Dario Colazzo, PhD defense committee of Bogdan Butnaru, “Optimizations of XQuery in peer-to-peer distributed XML databases”, Univ. Versaille, 12/04/2012.
- François Goasdoué, PhD defense committee of Jitao Yang, “Un modèle de données pour bibliothèques numériques”, Univ. Paris-Sud, 30/05/2012.
- Ioana Manolescu, PhD defense committee of Silviu Maniu, “Data Management in Social Networks”, Telecom ParisTech, 28/09/2012.
- Ioana Manolescu, PhD defense committee of Jordi Creus, “ROSES : Un moteur de requêtes continues pour l’aggrégation de flux RSS à large échelle”, UPMC - Sorbonne Universités, 7/12/2012.
- Ioana Manolescu, HDR defense committee of Cédric du Mouza, “Indexing Very Large Datasets”, Université Paris-Dauphine, 30/11/2012.

9.3. Popularization

Ioana Manolescu presented the issues involved in cloud-based management of Web data, to a panel on the technology aspects of Big Data, at the “Big Data” industrial conference at Paris on March 20, 2012.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] F. GOASDOUÉ. *Knowledge Representation meets DataBases for the sake of ontology-based data management*, Université Paris Sud - Paris XI, July 2012, Habilitation à Diriger des Recherches, http://hal.inria.fr/tel-00759274.

[2] K. KARANASOS. *Techniques fondées sur des vues matérialisées pour la gestion efficace des données du web*, Université Paris Sud - Paris XI, June 2012, http://hal.inria.fr/tel-00755328.

[3] N. MALLA. *Méthode de Partitionnement pour le traitement distribué et parallèle de données XML.*, Université Paris Sud - Paris XI, September 2012, http://hal.inria.fr/tel-00759173.

[4] B. MOHAMED-AMINE. *Analyse statique pour l'optimisation des mises à jour de documents XML temporels*, Université Paris Sud - Paris XI, September 2012, http://hal.inria.fr/tel-00765066.

[5] F. ULLIANA. *Types for Detecting XML Query-Update Independence*, Université Paris Sud - Paris XI, December 2012, http://hal.inria.fr/tel-00757597.

Articles in International Peer-Reviewed Journals

[6] N. BIDOIT, D. COLAZZO, F. ULLIANA. *Type-Based Detection of XML Query-Update Independence*, in "Proceedings of the VLDB Endowment", May 2012, http://hal.inria.fr/hal-00757544.

[7] M. HERSCHEL, F. NAUMANN, S. SZOTT, M. TAUBERT. *Scalable Iterative Graph Duplicate Detection*, in "IEEE Transactions on Knowledge and Data Engineering", November 2012, vol. 24, n^{o} 11, p. 2094-2108, http://hal.inria.fr/hal-00757604.

[8] L. LEITÃO, P. CALADO, M. HERSCHEL. *Efficient and Effective Duplicate Detection in Hierarchical Data*, in "IEEE Transactions on Knowledge and Data Engineering", 2012, vol. 99, n^{o} PrePrints [*DOI :* 10.1109/TKDE.2012.60], http://hal.inria.fr/hal-00722505.

Articles in National Peer-Reviewed Journals

[9] F. GOASDOUÉ, K. KARANASOS, Y. KATSIS, J. LEBLAY, I. MANOLESCU, S. ZAMPETAKIS. *Des triplets sur des arbres: un modèle hybride XML-RDF pour documents annotés*, in "Ingénierie des Systèmes d'Information (ISI)", 2012, vol. 17, n^{o} 5, p. 87-111 [*DOI :* 10.3166/ISI.17.5.87-111], http://hal.inria.fr/hal-00765302.

International Conferences with Proceedings

[10] A. ARANDA-ANDÚJAR, F. BUGIOTTI, J. CAMACHO-RODRÍGUEZ, D. COLAZZO, F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU. *AMADA: Web Data Repositories in the Amazon Cloud*, in "ACM International Conference on Information and Knowledge Management", Maui, États-Unis, 2012, http://hal.inria.fr/hal-00730687.

[11] N. BIDOIT, D. COLAZZO, N. MALLA, C. SARTIANI. *Partitioning XML Data for Iterative Queries.*, in "International Database Engineering & Applications Symposium (IDEAS)", Prague, Tchèque, République, August 2012, http://hal.inria.fr/hal-00758699.

[12] F. BUGIOTTI, F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU. *RDF Data Management in the Amazon Cloud*, in "Workshop on Data analytics in the Cloud (DanaC 2012)", Berlin, Allemagne, February 2012, http://hal.inria.fr/hal-00670492.

[13] J. CAMACHO-RODRÍGUEZ, D. COLAZZO, I. MANOLESCU. *Building Large XML Stores in the Amazon Cloud*, in "DMC - Data Management in the Cloud Workshop - 2012", Washington, D.C., États-Unis, 2012, http://hal.inria.fr/hal-00669951.

[14] F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *Getting More RDF Support from Relational Databases*, in "WWW - 21st World Wide Web Conference - 2012", Lyon, France, ACM New York, NY, USA, April 2012, http://hal.inria.fr/hal-00697062.

[15] M. HERSCHEL, H. EICHELBERGER. *The Nautilus Analyzer: Understanding and Debugging Data Transformations*, in "ACM International Conference on Information and Knowledge Management", Maui, HI, États-Unis, October 2012, http://hal.inria.fr/hal-00757591.

[16] M. HERSCHEL, I. MANOLESCU. *Data Bridges: Data Integration for Digital Cities*, in "CDMW - International Workshop on City Data Management, November 2012", Maui, HI, États-Unis, October 2012, http://hal.inria.fr/hal-00757569.

[17] K. KARANASOS, A. KATSIFODIMOS, I. MANOLESCU, S. ZOUPANOS. *ViP2P: Efficient XML Management in DHT Networks*, in "ICWE - 12th International Conference on Web Engineering", Berlin, Allemagne, Springer, July 2012, http://hal.inria.fr/hal-00692827.

[18] A. KATSIFODIMOS, I. MANOLESCU, V. VASSALOS. *Materialized View Selection for XQuery Workloads*, in "SIGMOD - ACM SIGMOD International Conference on Management of Data 2012", Scottsdale, Arizona, États-Unis, May 2012, http://hal.inria.fr/hal-00680365.

[19] J. LEBLAY. *SPARQL query answering with bitmap indexes*, in "SWIM - 4th International Workshop on Semantic Web Information Management - 2012", Scottsdale, AZ, États-Unis, ASSOCIATION FOR COMPUTING MACHINERY (editor), May 2012, http://hal.inria.fr/hal-00691235.

National Conferences with Proceeding

[20] F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *Répondre aux requêtes par reformulation dans les bases de données RDF*, in "RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle)", Lyon, France, January 2012, p. 978-2-9539515-2-3, Session "Posters", http://hal.inria.fr/hal-00656566.

Conferences without Proceedings

[21] D. COLAZZO, G. GHELLI, C. SARTIANI. *Typing Massive JSON Datasets*, in "International Workshop on Cross-model Language Design and Implementation (XLDI)", Copenhagen, Danemark, September 2012, http://hal.inria.fr/hal-00758716.

Scientific Books (or Scientific Book chapters)

[22] S. ABITEBOUL, I. MANOLESCU, P. RIGAUX, M.-C. ROUSSET, P. SENELLART. *Web Data Management*, Cambridge University Press, 2012, 456, http://hal.inria.fr/hal-00677720.

[23] M. HERSCHEL, L. BERTI-EQUILLE. *Application de mesures de distance pour la détection de problèmes de qualité de données*, in "La qualité et la gouvernance de données au service de la performance des entreprises", L. BERTI-EQUILLE (editor), Hermes Science Publications, 2012, p. 145-175, http://hal.inria.fr/hal-00757559.

Books or Proceedings Editing

[24] M. HERSCHEL, P. MOLITOR, K. ROTHERMEL (editors). *it - Information technology Special issue on data integration*, Oldenbourg Wissenschaftsverlag, June 2012, 52, http://hal.inria.fr/hal-00722507.

Research Reports

[25] F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *BGP Query Answering against Dynamic RDF Databases*, Inria, July 2012, n^{o} RR-8018, 42, http://hal.inria.fr/hal-00719641.