IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2014

Project-Team OAK

Database optimizations and architectures for complex large data

RESEARCH CENTER
Saclay - Île-de-France

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team OAK

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

Creation of the Team: 2012 April 01*, updated into Project-Team:* 2013 January 01.

1. Members

Research Scientists
Ioana Manolescu Goujot [Team leader, Inria, Senior Researcher, HdR]
Nicole Bidoit [UNIV. PARIS-SUD, Researcher, HdR]

Faculty Members
Bogdan Cautis [UNIV. PARIS-SUD, Professor, HdR]
Benoît Groz [UNIV. PARIS-SUD, Associate Professor]
Melanie Herschel [UNIV. PARIS-SUD, Associate Professor, until August 2014]

Engineers
Elham Akbari Azirani [Inria, from Apr 2014]
Benjamin Djahandideh [Junior engineer, Inria]
Juan Alvaro Munoz Naranjo [Junior engineer, Inria, from Mar 2014]

PhD Students
Ioana-Alexandra Roatis [UNIV. PARIS-SUD, until October 2014]
Raphael Bonaque [Inria]
Damian Bursztyn [Inria]
Jesús Camacho Rodríguez [UNIV. PARIS-SUD, until October 2014]
Paul Lagree [UNIV. PARIS-SUD, from October 2014]
Aikaterini Tzompanaki [UNIV. PARIS-SUD]
Stamatios Zampetakis [Inria]

Post-Doctoral Fellows
Francesca Bugiotti [Inria]
Ioana Ileana [Inria, since December 2014]
Soudip Roy Chowdhury [Inria]

Administrative Assistant
Maëva Jeannot [Inria]

Others
Tianqi Lei [Inria intern (Ecole polytechnique), from April 2014 until July 2014]
Sejla Cebiric [Inria, from Feb 2014]
Dario Colazzo [Professor, Univ. Paris IX]
Sofoklis Floratos [Ecole Polytechnique, until June 2014]
François Goasdoué [Professor, Univ. Rennes I]
Yifan Li [PhD student (Chinese Scholarship Council), until Aug 2014]

2. Overall Objectives

2.1. Overall Objectives

Data is being created at unprecedented scale and speed, and processed in increasingly varied and complex fashion. OAK research aims at devising expressive models for flexible processing of complex data, in particular Web and social data; we also devise and develop strong software tools efficiently implementing such rich models.

The team has developed pointed expertise related to the processing of Web data (in particular XML, RDF, or social graph data), and in models and architecture for the massively parallel management of Web data.

3. Research Program

3.1. Scalable and Expressive Techniques for the Semantic Web

The Semantic Web vision of a world-wide interconnected database of *facts*, describing *resources* by means of *semantics*, is coming within reach as the W3C's RDF (Resource Description Format) data model is gaining traction. The W3C Linking Open Data initiative has boosted the publication and interlinkage of a large number of datasets on the semantic web resulting to the Linked Open Data Cloud. These datasets of billions of RDF triples have been created and published online. Moreover, numerous datasets and vocabularies from different application domains are published nowadays as RDF graphs in order to facilitate community annotation and interlinkage of both scientific and scholarly data of interest. RDF storage, querying, and reasoning is now supported by a host of tools whose scalability and expressive power vary widely. Unsurprisingly, some of the most scalable tools draw upon the existing models and architecture for managing structured data. However, such tools often ignore the semantic aspects that make RDF interesting. For what concerns the semantics, a delicate balance must be found between expressive power and the efficiency of the resulting data management algorithms.

- The team works on identifying tractable dialects of RDF, amenable to highly efficient query answering algorithms, taking into account both data and semantics.
- Another line of research investigates the usage of RDF data and semantics to help structure, organize, and enrich structured documents from social media. Based on such a rich model, we devised novel query answering algorithms which attempt to explore efficiently the rich social dataset in order to return the most pertinent answers to the users, from a social, structured and semantic perspective. This research is related to the DIGICOSME LabEx grant "Structured, Social and Semantic Search".
- Last but not least, we investigate novel models and algorithms for efficient Semantic Web data management, going beyond the existing standard languages. We have finalized our proposal of an all-RDF data analytics framework, combining the rich structure and semantics of RDF with the power of analysis tools previously developed for relational data, such as analytical schemas and queries. Recent and ongoing work focuses on the automated selection of RDF analytical schemas as well as on efficient view-based analytical query answering strategies. The research is related to the "Investissement d'Avenir" project DATALYSE.

3.2. Massively Distributed Data Management Systems

Large and increasing data volumes have raised the need for distributed storage architectures. Among such architectures, computing in the cloud is an emerging paradigm massively adopted in many applications for the scalability, fault-tolerance and elasticity features it offers, which also allows for effortless deployment of distributed and parallel architectures. At the same time, interest in massively parallel processing has been renewed by the MapReduce model and many follow-up works, which aim at simplifying the deployment of massively parallel data management tasks in a cloud environment. For these reasons, cloud-based stores are an interesting avenue to explore for handling very large volumes of RDF data.

Our research aims at taking advantage of such widely available, large-scale distributed architectures to build scalable platforms for massively distributed management of complex data. We consider many different wide-scale distributed back-ends in this context, ranging from those provided by commercial cloud platforms to simple MapReduce and to more complex extensions thereof. In particular, we have considered the Stratosphere platform developed at TU Berlin, currently distributed by Apache under the name Flink.

This research is part of our participation to the Datalyse project previously mentioned, as well as the KIC EIT ICT Labs Europa activity, part of the "Computing in the Cloud" action line. We have completed our objectives within Europa, and our participation ended in 2014.

A recent development in this area is the start of our collaboration with social scientists from UNIV. PARIS-SUD, working on the management of innovation; we have started a collaborative research projects (ANR "Cloud-Based Organizational Design") where we perform an interdisciplinary analysis (both from a computing and from a business management perspective) on the adoption of cloud technologies within an enterprise.

3.3. Advanced Algorithms for Data Querying and Transformation

The *efficient* evaluation of queries over large databases remains a challenging task, to which certain optimization approaches based on static analysis of queries, data properties (such as integrity constraints), and indexing capabilities (such as materialized views) can still provide practically-relevant solutions. In this area, mainly for relational stores, we focus on query reformulation under constraints and views, as a uniform solution to problems such as view-based rewriting under constraints, semantic query optimization, and physical access path selection in query optimization.

With the increasing amount of available data, as well as the increasing complexity of data processing and transformations 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. In particular, it has become essential to analyze and debug data transformations. 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 . In particular, we study one important sub-problem of data transformation analysis, namely the one of Why-Not questions. Such questions can explain to developers of complex data transformations or manipulations why their data transformation did not produce some specific results, although they expected them to do so.

3.4. Social Data Management and Crowdsourcing

The social Web blurs today the distinction between search, recommendation, and advertising (three paradigms for information access that have been so far considered mostly in separation). Our research in this area strives to find better adapted and scalable ways to answer information needs in the social Web, often by techniques at the intersection of databases, information retrieval, and data mining.

In particular, we study models and algorithms for personalized, or social-aware search in social applications. While progress has been made in this area, more remains to be done in order to address users' needs in practice, especially towards richer data models, and improving applicability and result relevance. For instance, when searching for tweets, their geographical location and recency may be as important for relevance as the textual and social aspects.

Furthermore, regarding quality of answers in response to searches, for various reasons (e.g., sparsity or tagging quality), meaningful results may often not be available. One response to this observation could be to turn to the crowd, the very users/publishers of the social media platform, and to turn this crowd into on-demand and query-driven sources of data. We study principled approaches for crowd selection (expert sourcing) and task assignment (data sourcing), in order to better answer ongoing social queries.

Beyond social links that represent just ties, a promising direction we also focus on in user-centric applications is to uncover implicit, potentially richer relationships from user interactions and to exploit them to improve core functionality such as search.

Moreover, we plan to investigate how crowdsourcing can be exploited to extract informations on user preferences, using techniques about noisy data management and provenance analysis.

4. Application Domains

4.1. Social Networks

We develop models and algorithms for efficiently exploiting, enhancing, and querying social network data, in particular based on structured content, semantic annotations, and user interaction networks. We pursue this research with many industrial partners within the ALICIA project (Section 7.2.1) as well as in the Structured, Social, and Semantic Search project (Section 7.2.2).

4.2. Computational Journalism

Modern journalism increasingly relies on content management technologies in order to represent, store, and query source data and media objects themselves. Writing news articles increasingly requires consulting several sources, interpreting their findings in context, and crossing links between related sources of information. OAKresearch results directly applicable to this area provide techniques and tools for rich Web content warehouse management. We have launched a collaboration with Le Monde's "Les Décodeurs" team to investigate these topics.

4.3. Open Data Intelligence

The Web is a vast source of information, to which more is added every day either in unstructured form (Web pages) or, increasingly, as partially structured sources of information, in particular as Open Data sets, which can be seen as connected graphs of data, most frequently described in the RDF data format recommended by the W3C. Further, RDF data is also the most appropriate format for representing structured information extracted automatically from Web pages, such as the DBPedia database extracted from Wikipedia or Google's InfoBoxes. To intelligently exploit such Open Data collections, OAKhas developed a complete framework for RDF data analytics within the recently completed DW4RDF project and continues work on this topic within the ODIN project started this year.

4.4. Hybrid Data Warehousing

Increasingly many modern applications need to exploit data from a variety of formats, including relations, text, trees, graphs etc. The recent development of data management systems aimed at "Big Data", including NoSQL platforms, large-scale distributed systems etc. provides enteprise architects with many systems to chose from. This makes it hard to decide which part of the application data to handle in which system, especially given that each system is best at handling a specific kind of data and a certain class of operations. OAKinvestigates principled techniques for distributing an application's data sources across a variety of systems and data models, based on materialized views. We test our ideas in this area within the Datalyse project.

5. New Software and Platforms

5.1. Amada

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

Contact: Jesús Camacho-Rodríguez (jcamachor[at]gmail.com))

Other contacts: Ioana Manolescu (ioana.manolescu[at]inria.fr), Dario Colazzo (dario.colazzo[at]dauphine.fr), François Goasdoué (fg[at]irisa.fr)

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

5.2. PAXQuery

Name: PAXQuery (https://team.inria.fr/oak/projects/paxquery/)

Contact: Jesús Camacho-Rodríguez (jcamachor[at]gmail.com))

Other contacts: Ioana Manolescu (ioana.manolescu[at]inria.fr), Dario Colazzo (dario.colazzo[at]dauphine.fr), Juan Alvaro M. Naranjo (juan-alvaro.munoz-naranjo[at]inria.fr)

Presentation: A system for the massively parallel processing of XQuery queries, developed as an extension of the Apache Flink system (http://flink.apache.org/)

5.3. CliqueSquare

Name: CliqueSquare (https://team.inria.fr/oak/projects/cliquesquare/)

Contact: Stamatis Zampetakis (stamatis.zampetakis[at]inria.fr))

Other contacts: Ioana Manolescu (ioana.manolescu[at]inria.fr), François Goasdoué (fg[at]irisa.fr), Benjamin Djahandideh (benjamin.djahandideh[at]inria.fr)

Presentation: A system for the massively parallel evaluation of conjunctive SPARQL queries, built on top of Hadoop. The system has been released in open-source: https://sourceforge.net/projects/cliquesquare/.

5.4. FactMinder

Name: FactMinder (http://tripleo.saclay.inria.fr/xr/demo/)

Contact: Ioana Manolescu (ioana.manolescu[at]inria.fr)

Other contacts: Stamatis Zampetakis (stamatis.zampetakis[at]inria.fr), François Goasdoué (fg[at]irisa.fr).

Presentation: A system for archiving, annotating, and querying semantic-rich Web content.

5.5. Nautilus Analyzer

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

Contact: Melanie Herschel (melanie.herschel[at]lri.fr)

Other contacts: n.a.

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

5.6. PigReuse

Name: PigReuse

Contact: Jesús Camacho-Rodríguez (jcamachor[at]gmail.com)

Other contacts: Ioana Manolescu (ioana.manolescu[at]inria.fr), Dario Colazzo (dario.colazzo[at]dauphine.fr)

Presentation: A PigLatin optimization tool based on identifying and sharing repeated subexpressions.

5.7. WARG

Name: WARG (https://team.inria.fr/oak/warg/)

Contact: Alexandra Roatiş (alexandra.roatis[at]gmail.com)

Other contacts: Ioana Manolescu (ioana.manolescu[at]inria.fr), Sejla Cebiric (sejla.cebiric[at]inria.fr), François Goasdoué (fg[at]irisa.fr)

Presentation: A platform for specifying and exploiting warehouses of RDF data.

6. New Results

6.1. Highlights of the Year

The year has allowed reaching important results in four research areas of the group: query-based why-not provenance with explanations , minimal query reformulations under constraints [15], Linked Open Data analytics , and RDF data management in the cloud .

BEST PAPERS AWARDS :

[7] **Query-Based Why-Not Provenance with NedExplain in Extending Database Technology (EDBT)**. N. BIDOIT, M. HERSCHEL, K. TZOMPANAKI.

[11] **RDF Analytics: Lenses over Semantic Graphs in 23rd International World Wide Web Conference**. D. COLAZZO, F. GOASDOUÉ, I. MANOLESCU, A. ROATIS.
[4], [24] **RDF in the Clouds: A Survey in The International Journal on Very Large Databases**. Z. KAOUDI, I. MANOLESCU.

6.2. Scalable and Expressive Techniques for the Semantic Web

A main scientific topic of the team is the design of expressive and efficient tools for analyzing and manipulating Semantic Web data, in particular RDF. Our 2014 results in this area follow three complementary directions.

First, we have finalized our model for RDF analytics and proposed a full framework in which we fully redesign, from the bottom up, core data analytics concepts and tools in the context of RDF data, leading to the first complete formal framework for warehouse-style RDF analytics. Notably, we defined (i) *analytical schemas* tailored to heterogeneous, semantics-rich RDF graph, (ii) *analytical queries* which (beyond relational cubes) allow flexible querying of the data and the schema as well as powerful aggregation and (iii) *OLAP-style operations*. We implemented our RDF analytics platform on top of the KDB system and ported it on Postgres as well [11], [29]; work is ongoing to adapt it on a massively parallel RDF query evaluation platform, namely CliqueSquare (see below). In [25], we describe novel techniques for optimizing the evaluation of RDF analytical queries based on previously computed analytical query results.

Second, we continued our work on efficient evaluation of queries on RDF data, in the presence of constraints. *Reformulation-based query answering* is a query processing technique aiming at answering queries against data, under constraints. It consists of reformulating the query based on the constraints, so that evaluating the reformulated query directly against the data (i.e. without considering any more the constraints) produces the correct answer set. We have show how to optimize reformulation-based query answering in the setting of *ontology-based data access*, where SPARQL conjunctive queries are posed against RDF facts on which constraints expressed by an RDF Schema hold. The literature provides solutions for various fragments of RDF, aiming at computing the equivalent union of maximally-contained conjunctive queries w.r.t. the constraints. However, in general, such a union is large, thus it cannot be efficiently processed by a query engine. In this context, we have shown that generalizing the query reformulation language allows considering a space of reformulated queries (instead of a single possible choice), and selecting tthe reformulated query with lower estimated evaluation cost. We have shown experimentally that our technique enables reformulation-based query answering where the state-of-the-art approaches are simply unfeasible, while it may decrease their costs by orders of magnitude in other cases [21], [27].

Third, we have continued our work on cloud-based RDF data management. In [23], we have demonstrated CliqueSquare, a platform we developed in the team for the massively parallel processing of RDF queries. CliqueSquare enjoys the benefits of a query optimization algorithm which creates query plans as flat as possible, which in turn translates into massive opportunities for parallel processing. In [24], we have finalized our work on managing RDF data within the Amazon Web Services cloud. Finally, we have conducted a study of the existing models and algorithms published so far for the massively parrallel processing of RDF queries, which appeared as a survey in the VLDB Journal [4] and was also the basis of a tutorial at the ACM SIGMOD conference.

6.3. Massively Distributed Data Management Systems

Work in this area concerning the massively parallel processing of Semantic Web data was covered within the respective module.

We have finalized our work on massively parallel processing of XML queries based on the Apache Flink framework, formerly known as Stratosphere from the Technical University of Berlin, which implements the PACT model (an expressive extension of MapReduce). In [22], we have addressed the problem of efficiently parallelizing the execution of complex nested data processing, expressed in XQuery. We provided novel algorithms showing how to translate such queries into PACT, a recent framework generalizing MapReduce in particular by supporting many-input tasks. We presented the first formal translation of complex XQuery

algebraic expressions into PACT plans, and demonstrated experimentally the efficiency and scalability of our approach. The work has recently been accepted for publication to IEEE TKDE (to appear in 2015),

Finally, we have considered improving the performance of massively parallel data processing programs expressed using the PigLatin language. PigLatin is a popular language within the data management community interested in the efficient parallel processing of large data volumes. The dataflow-style primitives of PigLatin provide an intuitive way for users to write complex analytical queries, which are in turn compiled into MapReduce jobs. Currently, subexpressions occurring repeatedly in PigLatin scripts are executed as many times as they occur, leading to avoidable MapReduce jobs. The current PigLatin optimizer is not capable of recognizing, and thus optimizing, such repeated subexpressions. In [19], we have presented We present a novel approach for identifying and reusing common subexpressions occurring in PigLatin scripts. In particular, we lay the foundation of our reuse-based algorithms by formalizing the semantics of the PigLatin query language with extended nested relational algebra for bags. Our algorithm, named PigReuse, operates on the algebraic representations of PigLatin scripts, identifies subexpression merging opportunities, selects the best ones to execute based on a cost function, and merges other equivalent expressions to share its result; our experiments have confirmed the efficiency and effectiveness of our reuse-based algorithms and optimization strategies.

6.4. Advanced Algorithms for Data Querying and Transformation

We revisit in [15] the Chase&Backchase (C&B) algorithm for query reformulation under constraints. For an important class of queries and constraints, C&B has been shown to be complete, i.e. guaranteed to find all (join-)minimal reformulations under constraints. C&B is based on constructing a canonical rewriting candidate called a universal plan, then inspecting its exponentially many sub-queries in search for minimal reformulations, essentially removing redundant joins in all possible ways. This inspection involves chasing the subquery. Because of the resulting exponentially many chases, the conventional wisdom has held that completeness is a concept of mainly theoretical interest. We show that completeness can be preserved at practically relevant cost by introducing a novel reformulation algorithm that instruments the chase to maintain provenance information connecting the joins added during the chase to the universal plan subqueries responsible for adding these joins. This allows it to directly "read off" the minimal reformulations from the result of a single chase of the universal plan, saving exponentially many chases of its subqueries. We exhibit natural scenarios yielding speedups of over two orders of magnitude between the execution of the best view-based rewriting found by a commercial query optimizer and that of the best rewriting found by our algorithm.

Different types of explanations that serve as Why-Not answers have been proposed in the past and are either based on the available data, the query tree, or both. A first approach to this so called why-not provenance has been recently proposed. In [7], we show that this first approach has some shortcomings. To overcome these shortcomings, we propose Ned, an algorithm to explain data missing from a query result. NedExplain computes the why-not provenance for monotone relational queries with aggregation. This work contributes to providing necessary formalization in which the new algorithm is build. It also develops a comparative evaluation showing that it is both more efficient and effective than the state-of-the-art approach.

Solutions to answering Why-Not questions are generally more efficient and easier to interpret by developers than solutions solely based on data. However, algorithms producing such query-based explanations including ours ([7]) so far may return different results for reordered conjunctive query trees, and even worse, these results may be incomplete. Clearly, this represents a significant usability problem, as the explanations developers get may be partial and developers have to worry about the query tree representation of their query, losing the advantage of using a declarative query language. As remedy to this problem, in [6][18], we propose to capture query based answers of Why-Not questions through operator polynomial and we devised an algorithm called Ted that produces the same complete query-based explanations for reordered conjunctive query trees.

6.5. Social Data Management and Crowdsourcing

In [13], we focused on the issue of defining models and metrics for reciprocity in signed graphs. In unsigned directed networks, reciprocity quantifies the predisposition of network members in creating mutual

connections. On the other hand, this concept has not yet been investigated in the case of signed graphs. We capitalize on the graph degeneracy concept to identify subgraphs of the signed network in which reciprocity is more likely to occur. This enables us to assess reciprocity at a global level, rather than at an exclusively local one as in existing approaches. The large scale experiments we perform on real world data sets of trust networks lead to both interesting and intuitive results. We believe these reciprocity measures can be used in various social applications such as trust management, community detection and evaluation of individual nodes. The global reciprocity we define in this paper is closely correlated to the clustering structure of the graph, more than the local reciprocity as it is indicated by the experimental evaluation we conducted.

As initial step towards better answering information needs in applications managing social content that is structured and possibly enriched with semantic annotations, in [20], we present a preliminary data model and an approach for answering queries over structured, social and semantic-rich content, taking into account all dimensions of the data in order to return the most meaningful results.

7. Partnerships and Cooperations

7.1. Regional Initiatives

DW4RDF is a Digiteo project joint between Inria and U. Paris Sud, focused on analytic platforms for RDF data. The project has ended in October 2014, it has lasted three years, and it was coordinated by François Goasdoué. The project has provided the framework for the PhD of Alexandra Roatis [11], [29], [5], [29], [2].
S4 (Social, Structured and Semantic Search) is a Digicosme project joint between Inria and U. Paris Sud, focused on developing novel models and algorithms for user-centric search in a social context where complex documents are authored and endowed with rich semantics. The project provides the framework for the PhD of Raphael Bonaque [20].

7.2. National Initiatives

7.2.1. ANR

Apprentissage Adaptatif pour le Crowdsourcing Intelligent et l'Accès à l'Information (ALICIA) is a 4-year project, started in February 2014, supported by the ANR CONTINT call. The project is coordinated by Bogdan Cautis, with Nicole Bidoit, and Ioana Manolescu. Its goal is to study models, techniques, and the practical deployment of adaptive learning techniques in user-centric applications, such as social networks and crowdsourcing.
Cloud-Based Organizational Design (CBOD) is a 4-year ANR started in 2014, coordinated by prof. Ahmed Bounfour from UNIV. PARIS-SUD. Its goal is to study and model the ways in which cloud computing impacts the behavior and operation of companies and organizations, with a particular focus on the cloud-based management of data, a crucial asset in many companies.
Datalyse is funded for 3.5 years as part of the *Investissement d'Avenir - Cloud & Big Data* national program. The project is led by the Grenoble company Eolas, a subsidiary of Business & Decision. It is a collaboration with LIG Grenoble, U. Lille 1, U. Montpellier, and Inria Rhône-Alpes aiming at building scalable and expressive tools for Big Data analytics.

7.2.2. LabEx, IdEx

Structured, Social and Semantic Search is a 3-year project started in October 2013, financed by the *LabEx (Laboratoire d'Excellence)* DIGICOSME. The project aims at developing a data model for rich structured content enriched with semantic annotations and authored in a distributed setting, as well as efficient algorithms for top-k search on such content.
BizModel4Cloud is a one-year (2014) interdisciplinary research project funded under a *Projet Exploratoire Premier Soutien (PEPS)* call joint between the CNRS and the IdEx Paris Saclay. It reunites the same partners as the ANR CBOD project of which it is an initial, short version.

7.2.3. Others

ODIN is a four-year project started in 2014, funded by the Direction Générale de l'Armement, between the SemSoft company, IRISA Rennes and Inria Saclay (OAK). The project aims to develop a complete framework for analytics on Web data, in particular taking into account uncertainty, based on Semantic Web technologies such as RDF.

7.3. European Initiatives

7.3.1. Collaborations in European Programs, except FP7 & H2020

Program: COST

Project acronym: Keystone

Project title: Semantic keyword-based search on structured data sources

Duration: Oct 2013 – Oct 2018

Coordinator: Francesco Guerra (U. Modena, Italy)

Other partners: The project involves 24 countries, see http://www.cost.eu/domains_actions/ict/Actions/IC1302?parties

Abstract: To build efficient and expressive keyword search tools, the action "semantic KEYword-based Search on sTructured data sOurcEs" (KEYSTONE) proposes to draw upon competencies from several disciplines, such as semantic data management, the semantic web, information retrieval, artificial intelligence, machine learning, user interaction, service science, service design, and natural language processing.

7.4. International Initiatives

7.4.1. Inria Associate Teams

7.4.1.1. OAKSAD

Title: Languages and techniques for efficient large-scale Web data management

International Partner (Institution - Laboratory - Researcher):

University of California, San Diego (ÉTATS-UNIS)

Duration: 2013 - 2015

See also: https://team.inria.fr/oak/oaksad/

Data on the Web is increasingly large and complex. The ways to process and share it have also evolved, from the classical scenario where users connect to a database, to today's complex processes whereas data is jointly produced on the Web, disseminated through streams, corroborated and enriched through annotations, and exploited through complex business processes, or workflows. The OAK and San Diego teams work together to devise expressive languages, efficient techniques and scalable platforms for such applications. The main areas on which our interest is shared are: semantic Web annotations; large-scale distributed data sharing; monitoring and verification of automated data processing workflows in the cloud.

7.4.2. Inria International Partners

7.4.2.1. Informal International Partners

We have started discussions with the University of Tsukuba (Japan) and prepare a future submission of an associate team with them, on topics related to efficient techniques for querying distributed heterogeneous data sources.

7.5. International Research Visitors

7.5.1. Visits of International Scientists

- Yannis Velegrakis (U. Trento) visited the team in December 2014 and gave a seminar on recommender systems.
- Konstantinos Karanasos (Microsoft Research) visited the team in November 2014 and gave a seminar on dynamic query optimization in large-scale data processing platforms.
- Tamer Ozsu (U. Waterloo) visited the team in October 2014 and gave a seminar on distributed RDF data management.
- Alin Deutsch (UCSD) visited the team in October 2014 as part of our OAKSADjoint work.
- Dan Olteanu (Oxford U.) visited the team in October 2014 and gave a seminar on modern Datalog evaluation engines.
- Julien Leblay (Oxford U.) visited the team in May 2014 and gave a seminar on querying the deep web.
- Laurent Daynès (Oracle) visited the team in February 2014 and gave a seminar on optimization techniques for evaluating arithmetic expressions in Oracle.

7.5.1.1. Internships

- Sejla Cebiric (M2 intern), from University of Sarajevo, Bosnia (March - August 2014)
- Elham Akbari Azirani (M2 intern), from University of Teheran, Iran (April - September 2014)

7.5.2. Visits to International Teams

7.5.2.1. Research stays abroad

Bogdan Cautis visited Yahoo Labs Barcelona, in July, on the account of ongoing collaborations in as-you-type search and query recommendation in social media. He also visited the University of Singapore for one week in April (Stephane Bressan's team).

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific events organisation

8.1.1.1. General chair, Scientific chair

- I. Manolescu has been a co-chair of the 2nd International Workshop on Benchmarking RDF Systems (BeRSyS), in conjunction with VLDB 2014.

8.1.1.2. Member of the organizing committee

- N. Bidoit has been co-organiser of the summer school Ecole Thèmatique Bases de Données, June 2014, Oléron, whose focus was on Massive and Distributed Data Management.

8.1.2. Scientific events selection

8.1.2.1. Responsable of the conference program committee

- I. Manolescu has been a group leader (track chair) of the "Text Databases, XML and Keyword Search" track of the ACM SIGMOD Conference 2014.

8.1.2.2. Member of the conference program committee

- I. Manolescu: Extending Database Technologies Conference (EDBT), IEEE International Conference on Data Engineering (ICDE), SIGMOD/PODS PhD Symposium, Data4U workshop (in conjunction with VLDB 2014). She has also participated in the program committee of Bases de Données Avancées (BDA) 2014.

- Bogdan Cautis served as program committee member for the international conferences EDBT 2014, SIGMOD 2014, and CIKM 2014, as well as the workshops UnCrowd 2014 and BUDA 2014 (with ACM SIGMOD). He also participated in the program committee of Bases de Données Avancées (BDA) 2014.

8.1.2.3. Reviewer

- Bogdan Cautis served as reviewer for IEEE Transactions on Knowledge and Data Engineering.

8.1.3. Journal

8.1.3.1. Member of the editorial board

- I. Manolescu has been the Editor in Chief of the ACM SIGMOD Record until June 2014; she is an associate editor of the ACM Transaction on Internet Technology, and a member of the review board of PVLDB 2014.

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Master: Bogdan Cautis, Masses de Données, M2R IAC, 18 ETD, M2, UNIV. PARIS-SUD

Licence: Bogdan Cautis, Introduction au Bases de Données, 44 ETD, L1, IUT Orsay

Licence: Bogdan Cautis, Programmation et Administration des Bases de Données, 94 ETD, L1, IUT Orsay

Master: Benoît Groz, Bases de données dimensionnelles et OLAP, M2 Miage et M2 CA, 33 ETD, M2, UNIV. PARIS-SUD

Master: Benoît Groz, Principe d'utilisation des SGBD, M1 Miage et M1, 56 ETD, M1, UNIV. PARIS-SUD

Master: Ioana Manolescu, Masses de Données, M2R IAC, 26 ETD, M2, UNIV. PARIS-SUD

8.2.2. Supervision

Raphael Bonaque: "Structured, Social and Semantic Search", since October 2013, Bogdan Cautis, François Goasdoué, and Ioana Manolescu.

Damien Bursztyn: "Scalable Techniques for Web Data Management", since January 2014, François Goasdoué and Ioana Manolescu

Jesús Camacho-Rodríguez : "Cloud-based Web Data Management", defended in September 2014, Dario Colazzo and Ioana Manolescu.

Alexandra Roatiş: "Scalable Database Techniques for the Semantic Web", defended in September 2014, Dario Colazzo, François Goasdoué, and Ioana Manolescu.

Aikaterini Tzompanaki: Foundations and Algorithms to Compute the Provenance of Missing Data, since November 2012, Melanie Herschel and Nicole Bidoit.

Stamatis Zampetakis: "Massively Parallel Algorithms for Semantic Web Data", since October 2012, François Goasdoué and Ioana Manolescu.

Ioana Ileana: "Extracting and Archiving Web Data: a holistic approach", defended in October 2014, Bogdan Cautis.

Paul Lagrée: "Adaptive Learning for Intelligent Crowdsourcing and Information Access", since October 2014, Bogdan Cautis.

8.2.3. Juries

Bogdan Cautis has been a member of the PhD committee of Ryadh Dahimene (December 2014), CNAM, France.

Ioana Manolescu has reported on the PhD thesis of Laurent Pellegrino (Avril 2014), Université de Nice, France.
She has been a member of the PhD committees of Alexandra Roatis (September 2014, UNIV. PARIS-SUD), Jesús Camacho Rodríguez (September 2014, UNIV. PARIS-SUD) and Ioana Ileana (October 2014, Telecom ParisTech).
She has also been a member of the HDR committee of Fabian Suchanek (October 2014, Telecom ParisTech).

8.3. Popularization

At the International Conference on Database Systems for Advanced Applications (DASFAA 2014), B. Cautis gave a lecture for the local Indonesian students, on “Searching the Social Web”.

9. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] J. CAMACHO-RODRÍGUEZ. *Efficient techniques for large-scale Web data management*, Université Paris Sud - Paris XI, September 2014, https://tel.archives-ouvertes.fr/tel-01089388

[2] R. ROATIS. *Efficient Querying and Analytics of Semantic Web Data*, Université Paris Sud - Paris XI, September 2014, https://tel.archives-ouvertes.fr/tel-01082065

Articles in International Peer-Reviewed Journals

[3] B. CAUTIS, T. RISSE, E. DEMIDOVA, S. DIETZE, W. PETERS, N. PAPAILIOU, K. DOKA, Y. STAVRAKAS, V. PLACHOURAS, P. SENELLART, F. CARPENTIER, A. MANTRIC, B. CAUTIS, P. SIEHNDEL, D. SPILIOTOPOULOS. *The ARCOMEM Architecture for Social- and Semantic-Driven Web Archiving*, in "Future Internet", November 2014, 29 p. [*DOI :* 10.3390/FI6040688], https://hal.inria.fr/hal-01095075

[4] *Best Paper*
Z. KAOUDI, I. MANOLESCU. *RDF in the Clouds: A Survey*, in "The International Journal on Very Large Databases", June 2014, https://hal.inria.fr/hal-01020977.

Articles in National Peer-Reviewed Journals

[5] D. COLAZZO, F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *RDF analytics. Lenses over semantic graphs*, in "Ingénierie des Systèmes d’Information", 2014, vol. 19, n^{o} 4, pp. 87-117 [*DOI :* 10.3166/ISI.19.4.87-117], https://hal.inria.fr/hal-01080824

International Conferences with Proceedings

[6] N. BIDOIT, M. HERSCHEL, K. TZOMPANAKI. *Immutably Answering Why-Not Questions for Equivalent Conjunctive Queries*, in "TaPP 2014 - 6th USENIX Workshop on the Theory and Practice of Provenance", Cologne, Germany, June 2014 [*DOI :* 10.1145/NNNNNNN.NNNNNNN], https://hal.archives-ouvertes.fr/hal-01095479

[7] *Best Paper*
N. BIDOIT, M. HERSCHEL, K. TZOMPANAKI. *Query-Based Why-Not Provenance with NedExplain*, in "Extending Database Technology (EDBT)", Athens, Greece, March 2014, https://hal.inria.fr/hal-00962157.

[8] F. BUGIOTTI, D. BURSZTYN, A. DEUTSCH, I. ILEANA, I. MANOLESCU. *Invisible Glue: Scalable Self-Tuning Multi-Stores*, in "Conference on Innovative Data Systems Research (CIDR)", Asilomar, United States, January 2015, https://hal.inria.fr/hal-01087624

[9] F. BUGIOTTI, L. CABIBBO, P. ATZENI, R. TORLONE. *Database Design for NoSQL Systems*, in "International Conference on Conceptual Modeling", Atlanta, United States, October 2014, pp. 223 - 231 [*DOI :* 10.1007/978-3-319-12206-9_18], https://hal.inria.fr/hal-01092440

[10] J. CAMACHO-RODRÍGUEZ, D. COLAZZO, I. MANOLESCU. *PAXQuery: A Massively Parallel XQuery Processor*, in "DanaC'14", Snowbird, UT, United States, June 2014 [*DOI :* 10.1145/2627770.2627772], https://hal.archives-ouvertes.fr/hal-01086808

[11] *Best Paper*
D. COLAZZO, F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *RDF Analytics: Lenses over Semantic Graphs*, in "23rd International World Wide Web Conference", Seoul, South Korea, April 2014 [*DOI :* 10.1145/2566486.2567982], https://hal.inria.fr/hal-00960609.

[12] B. DJAHANDIDEH, F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU, J.-A. QUIANÉ-RUIZ, S. ZAMPETAKIS. *CliqueSquare in Action: Flat Plans for Massively Parallel RDF Queries*, in "International Conference on Data Engineering", Seoul, South Korea, April 2015, https://hal.inria.fr/hal-01108710

[13] C. GIATSIDIS, B. CAUTIS, S. MANIU, M. VAZIRGIANNIS, D. M. THILIKOS. *Quantifying trust dynamics in signed graphs, the S-Cores approach*, in "SDM'2014: SIAM International Conference on Data Mining", Philadelphia, United States, Siam, August 2014, pp. 668-676 [*DOI :* 10.1137/1.9781611973440.77], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01083529

[14] F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU, J.-A. QUIANÉ-RUIZ, S. ZAMPETAKIS. *CliqueSquare: Flat Plans for Massively Parallel RDF Queries*, in "International Conference on Data Engineering", Seoul, South Korea, April 2015, https://hal.inria.fr/hal-01108705

[15] I. ILEANA, B. CAUTIS, A. DEUTSCH, Y. KATSIS. *Complete Yet Practical Search for Minimal Query Reformulations Under Constraints*, in "SIGMOD Conference 2014", France, June 2014 [*DOI :* 10.1145/2588555.2593683], https://hal.inria.fr/hal-01086494

[16] K. STEFANIDIS, V. EFTHYMIOU, M. HERSCHEL, V. CHRISTOPHIDES. *Entity Resolution in the Web of Data*, in "International World Wide Web Conference (WWW)", Seoul, South Korea, April 2014, https://hal.inria.fr/hal-00962165

[17] K. TZOMPANAKI. *Semi-automatic SQL Debugging and Fixing to solve the Missing-Answers Problem*, in "Very Large Databases (VLDB'14) PhD Workshop", Hangzhou, China, September 2014, https://hal.inria.fr/hal-01095488

National Conferences with Proceedings

[18] N. BIDOIT, M. HERSCHEL, K. TZOMPANAKI. *Répondre à des requêtes Why-Not indépendamment de la représentation des requêtes*, in "Bases de données avancées (BDA14)", Autrans-Grenboble, France, October 2014, https://hal.inria.fr/hal-01095491

[19] J. CAMACHO-RODRÍGUEZ, D. COLAZZO, M. HERSCHEL, I. MANOLESCU, S. ROY CHOWDHURY. *Reuse-based Optimization for Pig Latin*, in "BDA'2014: 30e journées Bases de Données Avancées", Grenoble-Autrans, France, October 2014, https://hal.inria.fr/hal-01086497

Conferences without Proceedings

[20] R. BONAQUE, B. CAUTIS, F. GOASDOUÉ, I. MANOLESCU. *Toward Social, Structured and Semantic Search*, in "Surfacing the Deep and the Social Web (SDSW)", Riva del Garda, Trentino, Italy, Proceedings of the Workshop on Surfacing the Deep and the Social Web co-located with the 13th International Semantic Web Conference (ISWC 2014), COST Action KEYSTONE, October 2014, vol. 1310, https://hal.inria.fr/hal-01109123

[21] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *Optimizing Reformulation-based Query Answering in RDF*, in "BDA'14 - Gestion de Données – Principes, Technologies et Applications", Grenoble-Autrans, France, October 2014, https://hal.inria.fr/hal-01091205

[22] J. CAMACHO-RODRÍGUEZ, D. COLAZZO, I. MANOLESCU. *PAXQuery: Efficient Parallel Processing of Complex XQuery*, in "BDA'2014: 30e journées Bases de Données Avancées", Grenoble-Autrans, France, October 2014, https://hal.archives-ouvertes.fr/hal-01086809

[23] B. DJAHANDIDEH, F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU, J. QUIANÉ-RUIZ, S. ZAMPETAKIS. *How to Deal with Cliques at Work*, in "BDA'2014: 30e journées Bases de Données Avancées", Grenoble-Autrans, France, October 2014, https://hal.inria.fr/hal-01072060

Scientific Books (or Scientific Book chapters)

[24] F. BUGIOTTI, J. CAMACHO-RODRÍGUEZ, F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU, S. ZAMPETAKIS. *SPARQL Query Processing in the Cloud*, in "Linked Data Management", A. HARTH, K. HOSE, R. SCHENKEL (editors), Emerging Directions in Database Systems and Applications, Chapman and Hall/CRC, April 2014, https://hal.inria.fr/hal-00909121

Research Reports

[25] E. AKBARI AZIRANI, F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *Efficient OLAP Operations for RDF Analytics*, OAK team, Inria Saclay, January 2015, n° RR-8668, https://hal.inria.fr/hal-01101843

[26] F. BUGIOTTI, L. CABIBBO, P. ATZENI, R. TORLONE. *Database Design for NoSQL Systems*, Università degli studi Roma Tre, 2014 [*DOI :* 10.1007/978-3-319-12206-9_18], https://hal.inria.fr/hal-01092445

[27] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU. *Optimizing Reformulation-based Query Answering in RDF*, Inria Saclay, December 2014, n° RR-8646, https://hal.inria.fr/hal-01091214

[28] F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU, J. QUIANÉ-RUIZ, S. ZAMPETAKIS. *CliqueSquare: Flat Plans for Massively Parallel RDF Queries*, October 2014, n° RR-8612, https://hal.inria.fr/hal-01071984

Scientific Popularization

[29] A. ROATIS. *Analysing RDF Data: A Realm of New Possibilities*, in "ERCIM News", January 2014, n^{o} 96, https://hal.inria.fr/hal-00960743