Activity Report 2016

Team CEDAR

Rich Data Exploration at Cloud Scale

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

RESEARCH CENTER
Saclay - Île-de-France

THEME
Data and Knowledge Representation and Processing

Table of contents

Team CEDAR

Creation of the Team: 2016 January 01

Keywords:

Computer Science and Digital Science:

3.1.1. - Modeling, representation
3.1.2. - Data management, quering and storage
3.1.3. - Distributed data
3.1.6. - Query optimization
3.1.7. - Open data
3.1.8. - Big data (production, storage, transfer)
3.1.9. - Database
3.2.1. - Knowledge bases
3.2.3. - Inference
3.2.4. - Semantic Web
3.2.5. - Ontologies
3.3. - Data and knowledge analysis
3.3.1. - On-line analytical processing
3.3.2. - Data mining
3.3.3. - Big data analysis
8.1. - Knowledge

Other Research Topics and Application Domains:

1.1.6. - Genomics
8.5.1. - Participative democracy
9.4.5. - Data science
9.7.2. - Open data

1. Members

Research Scientists

Ioana Manolescu [Team leader, Inria, Senior Researcher, HDR]
Michael Thomazo [Inria, Researcher]

Faculty Member

Yanlei Diao [Ecole Polytechnique, Professor, HDR]

Engineers

Oscar Santiago Mendoza Rivera [Inria]
Swen Ribeiro [Inria]

PhD Students

Damian Bursztyn [Inria]
Tien Duc Cao [Inria, from Apr 2016]
Sejla Cebiric [Inria]
Raphael Bonaque [Inria, until Sep 2016]
Enhui Huang [École Polytechnique, from Sep 2016]

Post-Doctoral Fellows

Alessandro Solimando [Inria, until Sep 2016]
Fei Song [Inria, from Dec 2016]

Visiting Scientists

Davide Lanti [FUB, from Mar 2016]
Rana Alotaibi [UCSD, August to October 2016]
Xavier Tannier [UPS, Associate Professor, from May 2016, HDR]
François Goasdoué [U. de Rennes,Professor, HDR]

Administrative Assistants

Tiffany Caristan [Inria, from Jun 2016]
Helena Kutniak [Inria]
Maeva Jeannot [Inria]

2. Overall Objectives

2.1. Overall Objectives

Today data is being generated at an unprecedented rate, so much that 90created in the past two years. Such significant increase of data volume is due to the new ways that we gather data: from software tools that record system and user activities; from sensors and scientific instruments that monitor our built and natural environment; from medical instruments that enable genomic diagnosis of patients; and from user-initiated sources on the Web or social networks. Data often comes with semantics, enriching its interpretation and enhancing its value. Importantly, we observe that in today's data-intensive application, variety is the norm, and is likely to re- main so for a while. This is because different applications are best served by different kinds of data: traditional commerce-oriented applications use relational databases, Web content management systems handle semistruc- tured documents, sensors provide numerical streams, science applications manipulate arrays, highly heterogeneous data sets is often exported in RDF graphs, software system logs consist of structured text etc. At the scale and speed of consumption of today's Big Data, unifying data across such formats into a single architecture (approach formerly known as extract-transform-load in a data warehouse context) is no longer feasible. Instead, Cedar aims at inventing expressive models and highly efficient data management tools, focused from the start on Big Data variety. Our tools will be designed for deployment in the cloud, and validated at large scale.

3. Research Program

3.1. Scalable Heterogeneous Stores

Big Data applications increasingly involve *diverse* data sources, such as: structured or unstructured documents, data graphs, relational databases etc. and it is often impractical to load (consolidate) diverse data sources in a single repository. Instead, interesting data sources need to be exploited "as they are", with the added value of the data being realized especially through the ability to combine (join) together data from several sources. Systems capable of exploiting diverse Big Data in this fashion are usually termed *polystores*. A current limitation of polystores is that data stays captive of its original storage system, which may limit the data exploitation performance. We work to devise highly efficient storage systems for heterogeneous data across a variety of data stores.

3.2. Semantic Query Answering

In the presence of data semantics, query evaluation techniques are insufficient as they only take into account the database, but do not provide the reasoning capabilities required in order to reflect the semantic knowledge. In contrast, (ontology-based) query answering takes into account both the data and the semantic knowledge in order to compute the full query answers, blending query evaluation and semantic reasoning.

We aim at designing efficient semantic query answering algorithms, both building on cost-based reformulation algorithms developed in the team and exploring new approaches mixing materialization and reformulation.

3.3. Multi-Model Querying

As the world's affairs get increasingly more digital, a large and varied set of data sources becomes available: they are either structured databases, such as government-gathered data (demographics, economics, taxes, elections, ...), legal records, stock quotes for specific companies, un-structured or semi-structured, including in particular graph data, sometimes endowed with semantics (see e.g. the Linked Open Data cloud). Modern data management applications, such as data journalism, are eager to combine in innovative ways both static and dynamic information coming from structured, semi-structured, and un-structured databases and social feeds. However, current content management tools for this task are not suited for the task, in particular when they require a lenghy rigid cycle of data integration and consolidation in a warehouse. Thus, we see a need for flexible tools allowing to interconnect various kinds of data sources and to query them together.

3.4. Interactive Data Exploration at Scale

In the Big Data era we are faced with an increasing gap between the fast growth of data and the limited human ability to comprehend data. Consequently, there has been a growing demand of data management tools that can bridge this gap and help users retrieve high-value content from data more effectively. To respond to such user information needs, we aim to build interactive data exploration as a new database service, using an approach called "explore-by-example".

3.5. Exploratory Querying of Semantic Graphs

Semantic graphs including data and knowledge are hard to apprehend for users, due to the complexity of their structure and oftentimes to their large volumes. To help tame this complexity, in prior research (2014), we have presented a full framework for RDF data warehousing, specifically designed for heterogeneous and semantic-rich graphs. However, this framework still leaves to the users the burden of chosing the most interesting warehousing queries to ask. More user-friendly data management tools are needed, which help the user discover the interesting structure and information hidden within RDF graphs.

3.6. Representative Semantic Query Answering

Top-k search is a classical topic, studied in relational databases, semantic web, recommandation systems,... It is extremely useful, among other, when a human user face a large number of query results, allowing the user to reformulate the query if necessary. However, we argue that top-k search incurs a bias on the perception of the set of results which is out of the control of the user. Our goal is to provide the user with k answers as well which are chosen so as to represent the diversity of the answer set. We will first consider this problem in the setting of relational or RDF databases. We will then extend to more heterogeneous sources, including in particular plain text.

4. Application Domains

4.1. 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. CEDARresearch results directly applicable to this area provide techniques and tools for rich Web content warehouse management. This work will be funded by the ANR ContentCheck project, and a Google Award on Even Thread Extraction. We work in collaboration with Le Monde's "Les Décodeurs" team to investigate these topics.

4.2. 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. We work on this topic within the 4-year project ODIN started in 2014.

4.3. 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. CEDARinvestigates 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. Highlights of the Year

5.1. Highlights of the Year

ERC Proposal Accepted

Y. Diao's ERC Consolidator proposal "Charting a New Horizon of Big and Fast Data Analysis through Integrated Algorithm Design" has been accepted by the EU.

Awards

- A team of five including the team's PhD student Tien Duc Cao has won the first place at the Start-up Week-End in Artificial Intelligence (SWAI) in November 2016 (https://twitter.com/i/moments/796004617410711552, http://swai.fr/).
- Šejla Čebirić has been awarded the Google Anita Borg Scholarship.
- The paper "On the Complexity of Evaluating Regular Path Queries over Linear Existential Rules." by M. Bienvenu and M. Thomazo received the best paper award at the RR'16 conference .

BEST PAPER AWARD:

[3]
M. BIENVENU, M. THOMAZO. *On the Complexity of Evaluating Regular Path Queries over Linear Existential Rules*, in "10th International Conference on Web Reasoning and Rule Systems", Aberdeen, United Kingdom, 10th International Conference on Web Reasoning and Rule Systems, September 2016, https://hal.inria.fr/hal-01341787

6. New Software and Platforms

6.1. New Software

6.1.1. CliqueSquare

CliqueSquare allows storing and querying very large volumes of RDF data in a massively parralel fashion in a Hadoop cluster. The system uses its own partitioning and storage model for the RDF triples in the cluster.

CliqueSquare evaluates queries expressed in a dialect of the SPARQL query language. It is particularly efficient when processing complex queries, because it is capable of translating them into MapReduce programs guaranteed to have the minimum number of successive jobs. Given the high overhead of a MapReduce job, this advantage is considerable.

6.1.2. Compact

Compact reformulates conjunctive queries in the setting of ontology-based query anwering. It takes as input a conjunctive query and an ontology, and outputs a first-order rewriting of that query whenever it exists (without termination guarantee when it does not exists). To ease its use and dissemination, a novel version has been implemented by M. Thomazo based on the framework GRAAL, developed within the Inria Sophia-Antipolis team GraphIK by C. Sipieter, an engineer funded by an ADT. It will in particulary ease the integration with Semantic Web standards, as well as the use of query optimization techniques developed within Cedar for RDFS and DL-Lite$_{\mathcal{R}}$ to more general ontology languages.

6.1.3. RDF-Commons

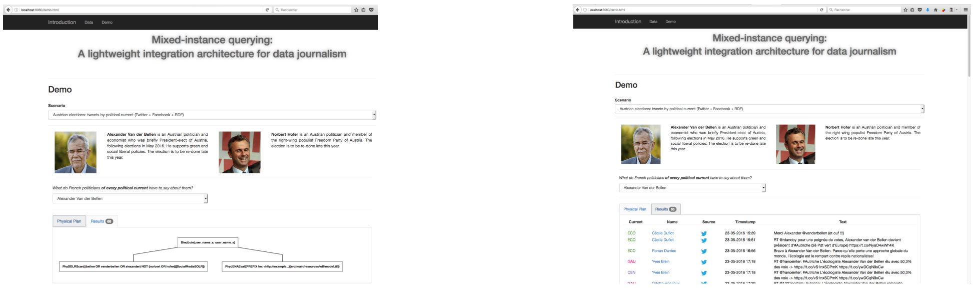
RDF-Commons is a set of modules providing the abilities to *i)* load and store RDF data in a DBMS *ii)* parse RDF conjunctive queries *iii)* encode URIs and literals into integers *iv)* encode RDF conjunctive queries *v)* build statistics on RDF data *vi)* estimate the cost of the evaluation of a conjunctive query *vii)* saturate the RDF data, with respect to an RDF Schema *viii)*reformulate a conjunctive query with respect to an RDF Schema (ix) propose algebraic plans.

The algebraic plan part has been developed by A. Solimando and D. Bursztyn. An ADT funding for two years has been granted to consolidate and extend the development of RDF-Commons. The hiring process is ongoing.

6.1.4. RDFSummary

RDF Summary is a standalone Java software capable of building summaries of RDF graphs. Summaries are compact graphs (typically several orders of magnitude smaller than the original graph), which can be used to get acquainted quickly with a given graph, they can also be used to perform static query analysis, infer certain things about the answer of a query on a graph, just by considering the query and the summary.

6.1.5. Tatooine



Mixed-instance querying:
A lightweight integration architecture for data journalism

Demo

Scenario: Austrian elections: tweets by political current (Twitter + Facebook + RDF)

Alexander Van der Bellen is an Austrian politician and economist who was briefly President-elect of Austria, following elections in May 2016. He supports green and social liberal policies. The election is to be re-done late this year.

Norbert Hofer is an Austrian politician and member of the right-wing populist Freedom Party of Austria. The election is to be re-done late this year.

*What do French politicians **of every political current** have to say about them?*

Alexander Van der Bellen

BindJoin(user_name_s, user_name_s)

PhySOLRScan(|(bellen OR vanderbellen OR alexander) NOT (norbert OR hofer)||SocialMediaSOLR|)

PhyJENAEval(|PREFIX fm: <http://example...|(src/main/resources/rdf/model.ttl|)

Current	Name	Source	Timestamp	Text
ECO	Cécile Duflot	🐦	23-05-2016 15:39	Merci Alexander @vanderbellen (et ouf !!!)
ECO	Cécile Duflot	🐦	23-05-2016 15:51	RT @rdandoy pour une poignée de votes, Alexander van der Bellen devient président d'#Autriche (2è Pdt vert d'Europe) https://t.co/NyaO4wWh4K
ECO	Ronan Dantec	🐦	23-05-2016 16:56	Bravo à Alexander Van der Bellen. Parce qu'elle porte une approche globale du monde, l'écologie est le rempart contre replis nationalistes!
GAU	Yves Blein	🐦	23-05-2016 17:18	RT @franceinter: #Autriche L'écologiste Alexander Van der Bellen élu avec 50,3% des voix -> https://t.co/vS1nxSCPmK https://t.co/ywDCqN9xCw
CEN	Yves Blein	🐦	23-05-2016 17:18	RT @franceinter: #Autriche L'écologiste Alexander Van der Bellen élu avec 50,3% des voix -> https://t.co/vS1nxSCPmK https://t.co/ywDCqN9xCw

Figure 1. Tweet enrichment in Tatooine: evaluation plan (left) and results (right).

We developed lightweight data integration system called Tatooine, based on our discussions with our journalist partners in the ANR ContentCheck project from the team "Les Décodeurs". Tatooine allows to exploit heterogeneous data sources of different data models, which we view as a mixed data instance, by querying them together; Tatooine combines data from various sources within an integrated engine complemented by information extraction and data visualization modules. Figure 1 illustrates the functioning of Tatooine through screen captures: a set of tweets (JSON documents stored in SOLR) obtained through a full-text search are combined with information about their authors (RDF metadata stored in Jena TDB) and the results are presented to the users highlighting the political affiliation of the tweet authors.

7. New Results

7.1. Scalable Heterogeneous Stores

To improve data querying performance within polystores (Section 3.1), we developed Estocada, a novel system capable of exploiting side-by-side a practically unbound variety of data management system, all the while guaranteeing the soundness and completeness of the store, and striving to extract the best performance out of the various DMSs. Estocada leverages recent advances in the area of query rewriting under constraints, which we use to capture the various data models and describe the fragments stored within each data management system. Estocada was demonstrated at the IEEE ICDE conference [12]; recent experimental results demonstrated performance improvements by many orders of magnitude brought by the fragments Estocada supports, with respect to the setting where data is stored only in the system it originates from. This work continues, in collaboration with Alin Deutsch and Rana Alotaibi from UCSD.

7.2. Semantic Query Answering

This is a core topic for the team, in which the year has been particularly fruitful.

First, we investigated efficient query answering techniques in knowledge bases. A large and useful set of ontologies enjoys FOL (first-order logic) reducibility of query answering, that is: answering a query q can be reduced to evaluating a certain first-order logic (FOL) formula (obtained from the query and ontology) against only the explicit facts. We devised a novel query optimization framework for ontology-based data access settings enjoying FOL reducibility. Our framework is based on searching within a set of alternative equivalent FOL queries, that is, FOL reformulations, one with minimal evaluation cost when evaluated through a relational database system. We applied this framework to the DL-Lite$_{\mathcal{R}}$ Description Logic underpinning the W3C's OWL2 QL ontology language, and demonstrated through experiments its performance benefits when two leading SQL systems, one open-source and one commercial, are used for evaluating the FOL query reformulations. This work has lead to a major publication in the PVLDB journal [13], and a demonstration at the Semantic Web conference [4], while the complete details appear in [16] and the PhD thesis of the student author. [2].

Second, we initiated a study of extensions of conjunctive queries to conjunctive regular path queries. The first step has been to study regular path queries under linear existential rules, generalizing previous work on DL-Lite$_{\mathcal{R}}$, which is at the core of the Semantic Web OWL 2 QL profile. Regular path queries are queries that check for a path between two individuals, which is labeled by a word belonging to a given regular language. Such navigational languages are very popular for graph-based data representation, such as RDF. We have studied the complexity for this query language, and shown that it is NL-complete in data complexity, and EXPTIME-complete in combined complexity (and PTIME complete with bounded arity). This work has received the best paper award at RR'16 , and is currently being extented to conjunctive regular path queries.

Last, we studied the expressivity of several variants of Datalog, the classical language for deductive databases. In particular, we have studied its expressivity when given access (or not) to input negation (the ability to check if an extensional atoms hold or not) and to a linear order. We provided a complete Venn diagram regarding the expressivity of all the variants when considering homomorphism-closed query. The trickiest (and most surprising) points is the existence of polynomial-time computable homomorphism closed queries that are not expressible within Datalog with linear order but without input negation. These results have been published at IJCAI'16 [7].

7.3. Multi-model Querying

We have proposed a lightweight data integration architecture implemented within Tatooine (see Section 6.1.5); the system was demonstrated on a data journalism use case at the prestigious VLDB conference [9].

A separate effort in the area of multi-model querying considered querying databases of interconnected documents, users and concepts, by means of keywords. In this context, it is important that query results reflect not only the keywords present in documents but also the links between users and documents (so as to return to one user first the results authored in his social neighborhood), links between documents (for instance when a tweet answers another or an article has a link to another), and last but not least semantic information which allows interconnecting and interpreting terms mentioned in text. This research was finalized as part of the PhD of Raphaël Bonaque [1] and appeared at the EDBT conference 2016 [11].

7.4. Interactive Data Exploration at Scale

In the work with Enhui Huang (PhD student at Ecole Polytechnique), we seek to minimize the number of samples presented to the user for reviewing in order to build an accurate model of the user interest. In particular, as the dimensionality of the data space increases, the number of samples needed to build an accurate user interest model increases fast. We examine a range of popular feature selection techniques for data exploration, and for the best-performing feature selection technique, Gradient boosting regression trees (GBRT), we propose optimizations to overcome the issue of unbalanced training data and to dynamically determine the number of relevant features to select. Experimental results show that our optimized GBRT improves F-measure from nearly 0 without feature selection, to high F-measure (>0.8), by adaptively choosing the number of relevant features.

This work is currently under submission to a database conference.

7.5. Exploratory Querying of Semantic Graphs

We have started work with an intern (Zheng Zhang) toward automatically exploring the structure of an RDF graph and visualizing it with the help of a D3.js (https://d3js.org/) visualization library. These initial steps should serve to guide the beginning of an interactive exploration of the RDF graph in order to identify interesting analytical queries to be asked and evaluated. This work continues.

Separately, with a different intern (Javier Letelier), we have investigated efficient algorithms for keyword search in an RDF graph, exploiting structural and semantic knowledge about the graph; such knowledge is organized as an RDF summary which is an RDF graph itself. The algorithm was implemented and integrated as a text search tool within the Tatooine prototype; the work is ongoing.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

- AIDE (“A New Database Service for Interactive Exploration on Big Data”) is an ANR “Young Researcher” project led by Y. Diao, to start at the end of 2016.
- CBOD (“Cloud-Based Organizational Design”) is a 4-year ANR started in 2014, coordinated by prof. Ahmed Bounfour from UPS. 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.
- ContentCheck (2015-2018) is an ANR project in collaboration with U. Rennes 1 (F. Goasdoué), INSA Lyon (P. Lamarre), the LIMSI lab from U. Paris Sud, and the Le Monde newspaper, in particular their fact-checking team Les Décodeurs. Its aim is to investigate content management models and tools for journalistic fact-checking.

- 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. The project has ended in November 2016.

8.1.2. LabEx, IdEx

- Structured, Social and Semantic Search (S4) 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. The project has ended in September 2016.
- CloudSelect is a three-years project started in October 2015. It is financed by the *Institut de la Société Numérique* (ISN) of the IDEX Paris-Saclay; it funds the PhD scholarship of S. Cebiric. The project is a collaboration with A. Bounfour from the economics department of Université Paris Sud. The project aims at exploring technical and business-oriented aspects of data mobility across clo ud services, and from the cloud to outside the cloud. Research contributing to this project is carried in collaboration with U. California in San Diego (UCSD) (see Section 3.1).

8.1.3. 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 Cedar. 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.
- Google Award I. Manolescu has received a Google Award in collaboration with X. Tannier from LIMSI/CNRS and Université de Paris-Sud. The award is given within a call specifically dedicated to computing tools for computational journalism. The project given the award focuses on "Event Thread Extraction for Viewpoint Analysis"; the project has finished at the end of 2016.

8.2. European Initiatives

8.2.1. FP7 & H2020 Projects

The permanent members of the team participate to build a proposal called GDMA (Graph Data Management and Analytics, for an European Joint Doctorate within the Initial Training Network (ITN) chapter of Europe's H2020 program, with the University of Aalborg (Denmark), Université Libre de Bruxelles, Universitat Politecnica de Catalunya, and University of Ioannina (Greece). If successful the project would involve six PhD thesis co-supervised in Cedar and starting in 2018, three students mostly residing with us, and three abroad working with our partners from Aalborg and Brussels.

I. Manolescu has submitted a Marie-Curie proposal titled IDEAA (An interactive toolbox to help citizens understand and build a viewpoint on specific issues by monitoring, analysing, and interlinking public data from EU institutions) to host a junior researcher (Mirjana Mazuran from Politecnico di Milano) for two years.

8.3. International Initiatives

8.3.1. Inria International Partners

8.3.1.1. Informal International Partners

We continue collaborating with U. California in San Diego (UCSD) following the OAKSAD associated team (2013-2015), in particular in the Estocada project (Section 7.1).

8.4. International Research Visitors

8.4.1. Visits of International Scientists

Several international guests gave seminars in our group:

- L. Ach and M. Rezk (Rakuten)
- D. Calvanese (University of Bolzano)
- R. Cheng (Hong Kong University)
- M. Franklin (University of Berkeley)
- R. Kontchakov, S. Kikot, M. Zakharyaschev (Birbeck University College)
- Y. Papakonstantinou (University of California in San Diego)
- V. Vianu (University of California in San Diego)

8.4.1.1. Internships

- R. Alotaibi visited the team for two months working on scalable heterogeneous stores with D. Bursztyn and I. Manolescu.
- D. Lanti visited the team for five months, working on efficient semantic query answering with D. Bursztyn.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Selection

9.1.1.1. Chair of Conference Program Committees

Y. Diao has been the Program Committee Chair of the ACM Symposium on Cloud Computing (SoCC) 2016, Chair of the ACM Symposium on Cloud Computing, October 2016 and Chair of the Inaugural Paris Big Data Summit, March 2016.

I. Manolescu has been the Program Committee Chair of the International Conference on Scientific and Statistical Database Management (SSDBM) 2016, and the Vision Papers track chair in the Extending Database Technologies (EDBT) conference 2016.

9.1.1.2. Member of the Conference Program Committees

- I. Manolescu has been a member of program committee of the IEEE Conference on Data Engineering (ICDE) 2016, of the Workshop on Data Engineering for the Semantic Web (DESWeb, in conjunction with ICDE) 2016 and of the Workshop on Semantic Big Data (SBD) 2016 in conjunction with the ACM SIGMOD conference.
- M. Thomazo has been member of the PC of ICCS'16, IJCAI'16 and PRIMA'16.

9.1.1.3. Reviewer

- M. Thomazo has reviewed for RR'16, VLDB'16 and ICDT'17.

9.1.2. Journal

9.1.2.1. Member of the Editorial Boards

- Y. Diao is Editor-in-Chief of the ACM SIGMOD Record, Associate Editor of ACM Transactions on Databases (TODS) and Editor of the Proceedings of Very Large Databases (PVLDB), 2016
- I. Manolescu has been an Associate Editor for the ACM Transactions of the Web (TWeb) and a member of the editorial board of PVLDB 2016.

9.1.2.2. Reviewer - Reviewing Activities

- Y. Diao has reviewed for the EDBT vision track
- I. Manolescu has reviewed for the ACM Transactions on Database Systems (TODS).
- M. Thomazo has reviewed for the Journal of Web Semantics.

9.1.3. Invited Talks

- I. Manolescu has given an invited talk “Estocada: Flexible Hybrid Stores” at the INESC institute in Portugal.
- I. Manolescu has given an invited talk “CliqueSquare: Flat Plans for Massively Parallel RDF Queries” at the Aalborg University, in Denmark.
- M. Thomazo has presented FactMinder, a tool developed in the prequel of Cedar, to a panel of computer scientists and journalists at the Tech&Check event organized by Duke University

9.1.4. Leadership within the Scientific Community

I. Manolescu has been a coordinator of Task 1 (Scalable and Secure Data Processing) of the DataSense axis of the DigiCosme Labex (*Laboratoire d’Excellence*). I. Manolescu has been a member of the ACM SIGMOD Jim Gray PhD Award Committee, of the IEEE TCDE Awards Committee. She has also been a member of the EDBT (Extending Database Technology) Association and of the Steering Committee of BDA (*Bases de Données Avancées*). I. Manolescu has joined the Board of Trustees of the Proceedings of VLDB (PVLDB) Endowment in 2016.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

- Master: Y. Diao has taught the course “Systems for Big Data Analytics” in the Data Science Master Program of Université Paris Saclay
- Master: I. Manolescu, Architectures for Massive Data Management, 30h, M2, Université Paris-Saclay, France.
- Master: I. Manolescu, Database Management Systems, 52h, M1, École Polytechnique, France.
- Doctorat: I. Manolescu, Scalable Tools for Linked Data Analytics, 20h, Aalborg University, Denmark.
- Master: M. Thomazo, Database Management Systems, 20h, M1, École Polytechnique, France
- Master: M. Thomazo, Introduction to Big Data Systems, 40h, M2, Université Paris-Saclay, France

9.2.2. Supervision

Raphael Bonaque: “Structured, Social and Semantic Search”, Université Paris Saclay, September 30, 2016, Bogdan Cautiş, François Goasdoué, and Ioana Manolescu.

Damian Bursztyn: “Scalable Techniques for Web Data Management”, December 15, 2016, François Goasdoué and Ioana Manolescu.

Tien Duc Cao: “Extraction et interconnexion de connaissances appliquée aux données journalistiques”, since October 2016, Ioana Manolescu and Xavier Tannier (LIMSI/CNRS and Université de Paris Sud)

Sejla Čebirić: “CloudSelect: Data Mobility Within, Across and Outside Clouds”, since September 2015, F. Goasdoué and I. Manolescu.

Enhui Huang: Interactive Data Exploration at Scale, since October 2016, Y. Diao.

Y. Diao has as well supervised 5-10 students each semester for the 3A research project and been mentor and chair of the defense of 17 student summer internship projects.

9.2.3. Juries

I. Manolescu has been a member of the PhD committee of Raphaël Bonaque and Damian Burstzyn, and a referee and jury member for the PhD of Gonçalo Simões at INESC-ID, Portugal.

9.3. Popularization

- M. Thomazo has presented a game based on RDF graphs and social networks as part of Fête de la Science in Inria Saclay, in October.
- The ANR ContentCheck project on **content management techniques for journalistic fact-checking** has attracted news attention in a series of general-audience articles published by Ouest France, Le Devoir (Canada), Rue89, Le Monde, Inria and CNRS, among others:
  - http://www.ouest-france.fr/leditiondusoir/data/569/reader/reader.html#!preferred/1/package/569/pub/570/page/8
  - http://www.ledevoir.com/politique/canada/450937/sur-la-piste-du-mensonge
  - http://rue89.nouvelobs.com/2015/10/26/algorithmes-antimensonge-fin-bobards-politique-261827
  - http://data.blog.lemonde.fr/2015/10/23/le-fact-checking-peut-il-sautomatiser/
  - http://www.inria.fr/centre/saclay/actualites/un-logiciel-de-fact-checking-pour-comprendre-le-monde-qui-nous-entoure
  - https://lejournal.cnrs.fr/articles/un-logiciel-qui-decrypte-la-politique

  The project has also lead to an interview by Science Po students studying in the Communication Master of the school, around automated fact-checking topics: https://inciviveritas.wordpress.com/menu-page/lautomatisation-du-fact-checking/.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] R. BONAQUE. *Top-k search over rich web content*, Université Paris-Saclay, September 2016, https://tel.archives-ouvertes.fr/tel-01418124

[2] D. BURSZTYN. *Efficient Big Data query answering in the presence of constraints*, Université Paris-Saclay, December 2016, https://tel.archives-ouvertes.fr/tel-01449287

International Conferences with Proceedings

[3] *Best Paper*
M. BIENVENU, M. THOMAZO. *On the Complexity of Evaluating Regular Path Queries over Linear Existential Rules*, in "10th International Conference on Web Reasoning and Rule Systems", Aberdeen, United Kingdom, 10th International Conference on Web Reasoning and Rule Systems, September 2016, https://hal.inria.fr/hal-01341787.

[4] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU. *Optimizing FOL reducible query answering: understanding performance challenges*, in "ISWC 2016: The 15th International Semantic Web Conference", Kobe, Japan, October 2016, https://hal.inria.fr/hal-01400568

[5] J. CAMACHO-RODRÍGUEZ, D. COLAZZO, M. HERSCHEL, I. MANOLESCU, S. ROY CHOWDHURY. *Reuse-based Optimization for Pig Latin*, in "25th ACM International on Conference on Information and Knowledge Management", Indianapolis, France, Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, October 2016, pp. 2215 - 2220 [*DOI :* 10.1145/2983323.2983669], https://hal.inria.fr/hal-01425321

[6] M. KRÖTZSCH, T. MASOPUST, M. THOMAZO. *On the Complexity of Universality for Partially Ordered NFAs* , in "41st International Symposium on Mathematical Foundations of Computer Science", Krakow, Poland, August 2016 [*DOI :* 10.4230/LIPICS.MFCS.2016.62], https://hal.inria.fr/hal-01334958

[7] S. RUDOLPH, M. THOMAZO. *Expressivity of Datalog Variants – Completing the Picture*, in "25th International Joint Conference on Artificial Intelligence", New-York, United States, July 2016, https://hal.inria.fr/hal-01302832

Conferences without Proceedings

[8] R. B. AL-OTAIBI, F. BUGIOTTI, D. BURSZTYN, A. DEUTSCH, I. MANOLESCU, S. ZAMPETAKIS. *Estocada: Stockage Hybride et Ré-écriture sous Contraintes d'Intégrité*, in "BDA: Conférence sur la Gestion de Données", Poitiers, France, November 2016, https://hal.archives-ouvertes.fr/hal-01355933

[9] R. BONAQUE, T. D. CAO, B. CAUTIS, F. GOASDOUÉ, J. LETELIER, I. MANOLESCU, O. MENDOZA, S. RIBEIRO, X. TANNIER, M. THOMAZO. *Mixed-instance querying: a lightweight integration architecture for data journalism*, in "VLDB", New Delhi, India, September 2016, https://hal.inria.fr/hal-01321201

[10] R. BONAQUE, B. CAUTIS, F. GOASDOUÉ, I. MANOLESCU. *Recherche Sociale, Structurée et Sémantique*, in "32ème Conférence sur la Gestion de Données - Principes, Technologies et Applications", Poitiers, France, November 2016, https://hal.inria.fr/hal-01426532

[11] R. BONAQUE, B. CAUTIS, F. GOASDOUÉ, I. MANOLESCU. *Social, Structured and Semantic Search*, in "International Conference on Extending Database Technology", Bordeaux, France, March 2016, https://hal.inria.fr/hal-01277939

[12] F. BUGIOTTI, D. BURSZTYN, A. DEUTSCH, I. MANOLESCU, S. ZAMPETAKIS. *Flexible Hybrid Stores: Constraint-Based Rewriting to the Rescue*, in "32nd IEEE International Conference on Data Engineering", Helsinki, Finland, May 2016, https://hal.inria.fr/hal-01321138

[13] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU. *Teaching an RDBMS about ontological constraints*, in "Very Large Data Bases", New Delhi, India, September 2016, https://hal.inria.fr/hal-01354592

[14] Š. ČEBIRIĆ, F. GOASDOUÉ, I. MANOLESCU. *Query-Oriented Summarization of RDF Graphs*, in "BDA (Bases de Données Avancées)", Poitiers, France, November 2016, https://hal.inria.fr/hal-01363625

Books or Proceedings Editing

[15] E. PITOURA, S. MAABOUT, K. GEORGIA, A. MARIAN, T. LETIZIA, I. MANOLESCU, K. STEFANIDIS (editors). *Proceedings of the 19th International Conference on Extending Database Technology, EDBT* , OpenProceedings.org, Bordeaux, France, March 2016, https://hal.archives-ouvertes.fr/hal-01285191

Research Reports

[16] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU. *Efficient query answering in the presence of DL-LiteR constraints*, Inria Saclay ; Inria, August 2016, n^{o} RR-8714, https://hal.inria.fr/hal-01143498

[17] J. CAMACHO-RODRÍGUEZ, D. COLAZZO, M. HERSCHEL, I. MANOLESCU, S. ROY CHOWDHURY. *PigReuse: A Reuse-based Optimizer for Pig Latin*, Inria Saclay, August 2016, https://hal.inria.fr/hal-01353891

[18] Š. ČEBIRIĆ, F. GOASDOUÉ, I. MANOLESCU. *Query-Oriented Summarization of RDF Graphs*, Inria Saclay ; Université Rennes 1, June 2016, n^{o} RR-8920, https://hal.inria.fr/hal-01325900