IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Cachan

Activity Report 2013

Project-Team DAHU

Verification in databases

IN COLLABORATION WITH: Laboratoire specification et vérification (LSV)

RESEARCH CENTER
Saclay - Île-de-France

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team DAHU

Keywords: Data Management, Databases, Web, Verification, Distributed System

Creation of the Project-Team: 2009 January 01.

1. Members

Research Scientists

Luc Segoufin [Team leader, Inria, Senior Researcher, HdR]
Serge Abiteboul [Inria, Senior Researcher, HdR]

Faculty Members

Arnaud Durand [Université Paris 7, Professor, from Sep 2013, HdR]
Sylvain Schmitz [ENS-Cachan, Associate Professor, from Sep 2013]
Cristina Sirangelo [ENS Cachan, Associate Professor]

PhD Students

Nadime Francis [ENS Cachan]
Wojciech Kazana [Inria, FP7 ERC WEBDAM project, until Jun 2013]
Émilien Antoine [Inria, FP7 ERC WEBDAM project]

Post-Doctoral Fellow

Johann Brault-Baron [Inria, from Sep 2013]

Visiting Scientists

Sergio Abriola [University of Buenos Aires, from Jun 2013 until Aug 2013]
Benoît Larose [Concordia University, from Nov 2013 until Nov 2013]
Victor Vianu [UCSD, from Jun 2013]

Administrative Assistant

Thida Iem [Inria]

2. Overall Objectives

2.1. Introduction

For more information see http://www.lsv.ens-cachan.fr/axes/DAHU/dahu.php.

The need to access and exchange data on the Web has led to database management systems (DBMS) that are increasingly distributed and autonomous. Data extraction and querying on the Web is harder than in classical DBMS, because such data is heterogeneous, redundant, inconsistent and subject to frequent modifications. DBMS thus need to be able to detect errors, to analyze them and to correct them. Moreover, increasingly complex Web applications and services rely on DBMS, and their reliability is crucial. This creates a need for tools for specifying DBMS in a high-level manner that is easier to understand, while also facilitating verification of critical properties.

The study of such specification and verification techniques is the main goal of Dahu.

2.2. Highlights of the Year

Serge Abiteboul was awarded the 2013 Milner awards.

3. Research Program

3.1. Research Program

Dahu aims at developing mechanisms for high-level specifications of systems built around DBMS, that are easy to understand while also facilitating verification of critical properties. This requires developing tools that are suitable for reasoning about systems that manipulate data. Some tools for specifying and reasoning about data have already been studied independently by the database community and by the verification community, with various motivations. However, this work is still in its infancy and needs to be further developed and unified.

Most current proposals for reasoning about DBMS over XML documents are based on tree automata, taking advantage of the tree structure of XML documents. For this reason, the Dahu team is studying a variety of tree automata. This ranges from restrictions of "classical" tree automata in order to understand their expressive power, to extensions of tree automata in order to understand how to incorporate the manipulation of data.

Moreover, Dahu is also interested in logical frameworks that explicitly refer to data. Such logical frameworks can be used as high level declarative languages for specifying integrity constraints, format change during data exchange, web service functionalities and so on. Moreover, the same logical frameworks can be used to express the critical properties we wish to verify.

In order to achieve its goals, Dahu brings together world-class expertise in both databases and verification.

4. Application Domains

4.1. Application Domains

Databases are pervasive across many application fields. Indeed, most human activities today require some form of data management. In particular, all applications involving the processing of large amounts of data require the use of a database. Increasingly complex Web applications and services also rely on DBMS, and their correctness and robustness is crucial.

We believe that the automated solutions that Dahu aims to develop for verifying such systems will be useful in this context.

5. New Results

5.1. Specification and Verification of Database Driven Systems

Participants: Serge Abiteboul, Luc Segoufin, Victor Vianu.

We continued our investication on the verification of database driven systems using an automata model with registers. We have exhibited new classes of decidable scenarios using nominal set theory [25]. These new classes contain the previously known relational cases but also the some semistructered ones.

We introduce in [24] and study a model of collaborative data-driven workflows. In a local-as-view style, each peer has a partial view of a global instance that remains purely virtual. Local updates have side effects on other peers' data, defined via the global instance. We also assume that the peers provide (an abstraction of) their specifications, so that each peer can actually see and reason on the specification of the entire system. We study the ability of a peer to carry out runtime reasoning about the global run of the system, and in particular about actions of other peers, based on its own local observations. A main contribution is to show that, under a reasonable restriction (namely, key-visibility), one can construct a finite symbolic representation of the infinite set of global runs consistent with given local observations. Using the symbolic representation, we show that we can evaluate in pspace a large class of properties over global runs, expressed in an extension of first-order logic with past linear-time temporal operators, PLTL-FO. We also provide a variant of the algorithm allowing to incrementally monitor a statically defined property, and then develop an extension allowing to monitor an infinite class of properties sharing the same temporal structure, defined dynamically as the run unfolds. Finally, we consider an extension of the language, augmeting work-flow control with PLTL-FO formulas. We prove that this does not increase the power of the workflow specification language, thereby showing that the language is closed under such introspective reasoning.

5.2. Distributed data management

Participants: Serge Abiteboul, Émilien Antoine, Cristina Sirangelo.

We have studied the feasibility of query answering in the presence of incomplete information in data. In particular we have investigated when it is the case that classical query evaluation techniques, which are commonly used over complete data, suffice to answer queries also in the presence of incompleteness [26]. These results allowed to find syntactic classes of queries that can be answered efficiently under many well known semantics of incompleteness, using query answering techniques which are already implemented (and optimized) in classical database systems.

The management of Web users' personal information is increasingly distributed across a broad array of applications and systems, including online social networks and cloud-based services. While users wish to share and integrate data using these systems, it is increasingly difficult to avoid the risks of unintended disclosures or unauthorized access by applications.

In [21], [20], we propose a novel access control model that operates within a distributed data management framework based on datalog. Using this model, users can control access to data they own and control applications they run. They can conveniently specify access control policies providing flexible tuple-level control derived using provenance information. We present a formal specification of the model, a theoretical analysis, and an implementation. We show that the computational cost of access control is acceptable.

5.3. Query Processing for the Web

Participants: Johann Brault-Baron, Arnaud Durand, Nadime Francis, Wojciech Kazana, Luc Segoufin, Cristina Sirangelo.

In many applications the output of a query may have a huge size and enumerating all the answers may already consume too many of the allowed resources. In this case it may be appropriate to first output a small subset of the answers and then, on demand, output a subsequent small numbers of answers and so on until all possible answers have been exhausted. To make this even more attractive it is preferable to be able to minimize the time necessary to output the first answers and, from a given set of answers, also minimize the time necessary to output the next set of answers - this second time interval is known as the *delay*. We have shown that this was doable with a linear preprocessing time and constant enumeration delay for first-order queries over structures of bounded expansion [27] and for monadic second-order queries over structures of bounded tree-width [15]. We also presented a survey about this work at the Intl. Conf. on Database Theory (ICDT) [19].

Web data is often structured in the XML format. In [18] we have surveyed results about static analysis of pattern-based queries over XML documents. These queries are analogs of conjunctive queries, their unions and Boolean combinations, in which tree patterns play the role of atomic formulae. These can be viewed as both queries and incomplete documents, and thus static analysis problems can also be viewed as answering queries over such documents. We looked at satisfiability of patterns under schemas, containment of queries for various features of XML used in queries, query answering, and applications of pattern-based queries in reasoning about schema mappings for data exchange.

6. Partnerships and Cooperations

6.1. European Initiatives

6.1.1. FP7 Projects

6.1.1.1. Webdam

Title: WebDam

Type: IDEAS

Instrument: ERC Advanced Grant (Advanced)

Duration: December 2008 - November 2013

Coordinator: Serge Abiteboul, Inria (France)

Others partners: Pierre Senellart, Telecom Paristech.

See also: http://webdam.inria.fr

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

6.2. International Initiatives

6.2.1. Inria International Partners

6.2.1.1. Declared Inria International Partners

Victor Vianu, UC San Diego, USA.

6.3. International Research Visitors

6.3.1. Visits of International Scientists

- Benoît Larose

  Subject: Constraint Satisfaction Problems

  Institution: concordia Univeresity, Montreal, Canada.

7. Dissemination

7.1. Scientific Animation

Organization of workshops and conferences.

- Serge Abiteboul, Pierre Senellart and Victor Vianu organized the Final Workshop for Web data management (nicknamed Webdone) in Paris 2013.

Program Committees.

- Cristina Sirangelo: 16th Intl. Conf. on Database Theory (ICDT 2013). ICDT 2013 Test of Time Award.
- Luc Segoufin: Intl. Conf. on Logic in Computer Science (LICS'13).

Responsibilities.

- Luc Segoufin is since 2010 part of the "bureau du commité des projets" à l'Inria Saclay. Since 2011 he is part of the scientific board of Inria. Since 2010 he is responsible of the groupe de travail "Complexité et Modèles Finis" du GDR "Mathématique et Informatique" (http://www.gdr-im.fr/).
- S. Abiteboul is the principal investigator of the European Research Council Grant Webdam on Web Data Management.

  As a member of the Sciences Academy, S. Abiteboul wrote a report on "L'enseignement de l'informatique en France - Il est urgent de ne plus attendre".

  S. Abiteboul is since 2013 a member of the Conseil national de la recherche. As a member, he participated in 2013 in reports on Net neutrality, Computer science education, and digital inclusion.

  S. Abiteboul is chairman of the Inria Awards committee.

  S. Abiteboul is chairman of the Scientific Board of Société d'Informatique de France.

  S. Abiteboul is a member of the Academic Senat of the University Paris-Saclay.

  S. Abiteboul is a member of the Academia Europea.

Larger audience. In 2013, S. Abiteboul gave talks to Entretiens de la Cité in Lyon, Conférence cultures numériques, éducation aux médias et à l'information in Lyon, Conférences Science et société in Nancy, Congrès de la Société informatique de France in Nice, Forum régionaux des Savoirs in Rouen. Journéee Economie de la connaissance et économie numérique at IHEST, Saclay.

In particular, S. Abiteboul gave talks on Big data in insurances at Journée SCOR sur les Big data et les assurances, and Journéee Ifpas de l'assurance in Paris; on Big data and health at Open data et santé, Congrés Health IT. He also discussed the place of fiction in the Web at Séminaire Vérifiction, CNAM, Paris, 2013.

S. Abiteboul a été également auditionné à l'Assemblée nationale par la commission des affaires économiques, Mission d'information sur l'économie numérique; et par l'Office parlementaire d'évaluation des choix scientifiques et technologiques, sur le Risque numérique. 2013.

S. Abiteboul gave interviews to Le Monde, Famille Chrétienne, and 01Net.

7.2. Teaching - Supervision - Juries

7.2.1. Teaching

Master : Cristina Sirangelo, Complexité avancée, 18 hours ETD, M1, MPRI, France

Master : Cristina Sirangelo, Algorithms, 15 hours ETD, Préparation à l'agrégation, École Normale Supérieure de Cachan, France

Licence : Cristina Sirangelo, Bases de données, 30 hours ETD, L3, École Normale Supérieure de Cachan, France

Doctorat : Cristina Sirangelo, Bases de données et sites Web dynamiques, 18 hours ETD, École Normale Supérieure de Cachan, France

Doctorat : Cristina Sirangelo, Création de sites Web, 18 hours ETD, École Normale Supérieure de Cachan, France

Licence : Serge Abiteboul, Base de données , ENS Cachan and ENS Paris

Master : Serge Abiteboul, Web data management, MPRI Paris

Master : Luc Segoufin, Finite Model Theory and Descriptive Complexity, MPRI.

Licence : Émilien Antoine, Algorithmique et complexité, 32h, L3, Université de Paris-Sud, France

7.2.2. Supervision

PhD: Wojciech Kazana, Query Evaluation with Constant Delay, 16/09/2013, Luc Segoufin

PhD in Progress: Nadime Francis, graph databases, 01/09/2011, Cristina Sirangelo and Luc Segoufin

PhD : Émilien Antoine, Data management in social network, 05/12/2013, Serge Abiteboul

7.2.3. Juries

- Luc Segoufin was reviewer for the PhD thesis of Stefan Mengel, Paderborn, Germany.
- Luc Segoufin was reviewer for the PhD thesis of Johann Brault-Baron, Université de Caen, France.

7.3. Popularization

Serge Abiteboul participated to a popularization book on mathematics, « Mathématiques, l'explosion continue », with an article, Chercher sur le Web : juste un point fixe et quelques algorithmes.

Serge Abiteboul wrote with Pierre Senellart an article on " Un déluge de données", in Pour la science sur le Big bang numérique, 2013.

8. Bibliography

Major publications by the team in recent years

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

[2] S. ABITEBOUL, L. SEGOUFIN, V. VIANU. *Static Analysis of Active XML Systems*, in "ACM Transactions on Database Systems", 2009, vol. 34, n^{o} 4

[3] P. BARCELÓ, L. LIBKIN, A. POGGI, C. SIRANGELO. *XML with incomplete information*, in "J. ACM", 2010, vol. 58, n^{o} 1

[4] M. BOJANCZYK, L. SEGOUFIN, H. STRAUBING. *Piecewise testable tree languages*, in "Logical Methods in Computer Science (LMCS)", 2012, vol. 8, n^{o} 3

[5] M. BOJAŃCZYK, C. DAVID, A. MUSCHOLL, T. SCHWENTICK, L. SEGOUFIN. *Two-variable logic on words with data*, in "ACM Trans. on Computational Logic (ToCL)", 2011, vol. 12, n^{o} 4

[6] M. BOJAŃCZYK, A. MUSCHOLL, TH. SCHWENTICK, L. SEGOUFIN. *Two-variable logic on data trees and applications to XML reasoning*, in "Journal of the ACM", 2009, vol. 56, n^{o} 3

[7] BALDER TEN. CATE, L. SEGOUFIN. *Transitive Closure Logic, Nested Tree Walking Automata, and XPath*, in "Journal of the ACM", 2010, vol. 57, n^{o} 3

[8] B. CAUTIS, S. ABITEBOUL, T. MILO. *Reasoning about XML update constraints*, in "Journal of Computer and System Sciences", 2009, vol. 75, n^{o} 6, pp. 336-358

[9] L. LIBKIN, C. SIRANGELO. *Reasoning about XML with temporal logics and automata*, in "Journal of Applied Logic", 2010, vol. 8, n^{o} 2, pp. 210-232, http://www.lsv.ens-cachan.fr/Publis/PAPERS/PDF/LS-jal10.pdf

[10] L. LIBKIN, C. SIRANGELO. *Data exchange and schema mappings in open and closed worlds*, in "Journal of Computer System Sciences (JCSS)", 2011

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] É. ANTOINE. , *Gestion des données distribuées avec le langage de règles: Webdamlog*, Université Paris Sud - Paris XI, December 2013, http://hal.inria.fr/tel-00908155

[12] W. KAZANA. , *l'évaluation de requêtes avec un délai constant*, École normale supérieure de Cachan - ENS Cachan, September 2013, http://hal.inria.fr/tel-00908434

Articles in International Peer-Reviewed Journals

[13] S. ABITEBOUL, Y. KATSIS, B. T. CATE. *On the equivalence of distributed systems with queries and communication*, in "Journal of Computer and System Sciences", 2013, http://hal.inria.fr/hal-00879029

[14] B. T. CATE, L. SEGOUFIN. *Unary negation*, in "Logical Methods in Computer Science", 2013, vol. 9, n^{o} 3, http://hal.inria.fr/hal-00904567

[15] W. KAZANA, L. SEGOUFIN. *Enumeration of monadic second-order queries on trees*, in "ACM Transactions on Computational Logic", 2013, vol. 14, n^{o} 4, http://hal.inria.fr/hal-00916400

Articles in National Peer-Reviewed Journals

[16] S. ABITEBOUL. *Vers une nouvelle science des risques ?*, in "Risques", September 2013, http://hal.inria.fr/hal-00908090

Invited Conferences

[17] S. ABITEBOUL. *Les connaissances de la toile*, in "Cultures numériques, éducation aux médias et à l'information", Lyon, France, Scéron Edition, May 2013, http://hal.inria.fr/hal-00915477

[18] A. GHEERBRANT, L. LIBKIN, C. SIRANGELO. *Reasoning About Pattern-Based XML Queries*, in "RR - 7th International Conference on Web Reasoning and Rule Systems, 2013", Mannheim, Germany, July 2013, http://hal.inria.fr/hal-00908414

[19] L. SEGOUFIN. *Enumerating with constant delay the answers to a query*, in "Intl. Conf. on Database Theory", Genes, Italy, March 2013, http://hal.inria.fr/hal-00907085

International Conferences with Proceedings

[20] S. Abiteboul, É. Antoine, G. Miklau, J. Stoyanovich, J. Testard. *Rule-Based Application Development using Webdamlog*, in "SIGMOD - Special Interest Group on Management Of Data", New York, United States, 2013, http://hal.inria.fr/hal-00817791

[21] S. Abiteboul, É. Antoine, G. Miklau, J. Stoyanovich, V. Zaychik Moffitt. *Introducing Access Control in Webdamlog*, in "DBPL - 14th International Symposium on Database Programming Languages - 2013", Riva del Garda, Trento, Italy, 2013, http://hal.inria.fr/hal-00850754

[22] S. Abiteboul, P. Bourhis, A. Muscholl, Z. Wu. *Recursive queries on trees and data trees*, in "Proceedings of the 16th International Conference on Database Theory", Genoa, Italy, 2013, http://hal.inria.fr/hal-00809297

[23] S. Abiteboul, D. Deutch, V. Vianu. *Deduction with Contradictions in Datalog*, in "International Conference on Database Theory", Athens, Greece, 2014, http://hal.inria.fr/hal-00923265

[24] S. Abiteboul, V. Vianu. *Collaborative data-driven workflows: think global, act local*, in "Proceedings of the 32nd symposium on Principles of database systems", New York, NY, USA, United States, ACM, 2013, pp. 91–102 [*DOI :* 10.1145/2463664.2463672], http://hal.inria.fr/hal-00840306

[25] M. Bojańczyk, L. Segoufin, S. Toruńczyk. *Verification of database-driven systems via amalgamation*, in "ACM conf. on Principle of Database Systems (PODS)", New-York, United States, June 2013, pp. 63-74, http://hal.inria.fr/hal-00908771

[26] A. Gheerbrant, L. Libkin, C. Sirangelo. *When is Naïve Evaluation Possible?*, in "PODS - 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2013", New York, United States, June 2013, http://hal.inria.fr/hal-00908404

[27] W. Kazana, L. Segoufin. *Enumeration of first-order queries on classes of structures with bounded expansion*, in "ACM conf. on Principle of Database Systems (PODS)", New-York, United States, 2013, pp. 297-308, http://hal.inria.fr/hal-00908779

Research Reports

[28] S. Abiteboul, É. Antoine, J. Stoyanovich. , *The Webdamlog System Managing Distributed Knowledge on the Web*, April 2013, http://hal.inria.fr/hal-00813300