



IN PARTNERSHIP WITH:
CNRS

**Ecole normale supérieure de
Cachan**

Activity Report 2014

Project-Team DAHU

Verification in databases

IN COLLABORATION WITH: Laboratoire spécification et vérification (LSV)

RESEARCH CENTER
Saclay - Île-de-France

THEME
**Data and Knowledge Representation
and Processing**

Table of contents

1. Members	1
2. Overall Objectives	1
3. Research Program	2
4. Application Domains	2
5. New Software and Platforms	2
6. New Results	2
6.1. Highlights of the Year	2
6.2. Distributed data management	3
6.3. Query Processing for the Web	3
6.4. Complexity in Counter Systems and Substructural Logics	3
6.5. Incomplete Databases	4
7. Bilateral Contracts and Grants with Industry	4
8. Partnerships and Cooperations	5
8.1. National Initiatives	5
8.2. International Initiatives	5
8.3. International Research Visitors	5
9. Dissemination	5
9.1. Promoting Scientific Activities	5
9.1.1. Scientific events organisation	5
9.1.1.1. general chair, scientific chair	5
9.1.1.2. member of the organizing committee	5
9.1.2. Scientific events selection	6
9.1.3. Journal	6
9.2. Teaching - Supervision - Juries	6
9.2.1. Teaching	6
9.2.2. Supervision	7
9.2.3. Juries	7
9.3. Popularization	7
10. Bibliography	7

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 [CNRS, Professor, until Aug 2014, HdR]

Sylvain Schmitz [ENS Cachan, Associate Professor]

Cristina Sirangelo [ENS Cachan, Associate Professor]

PhD Students

Émilien Antoine [Inria, until Jan 2014]

Nadime Francis [ENS Cachan]

Nathan Grosshans [CNRS, from Sep 2014]

David Montoya [Cifre, from Jan 2014]

Post-Doctoral Fellow

Johann Brault-Baron [Inria, until Aug 2014]

Administrative Assistant

Thida Iem [Inria]

Others

Victor Vianu [Chaire Inria, Professor]

Jakub Kalas [ENS Cachan, from Sep 2014]

2. Overall Objectives

2.1. Overall Objectives

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.

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 Software and Platforms

5.1. New Software

5.1.1. *The Webdamlog system*

The Webdamlog system is a distributed knowledge management system. A new version of the system has been developed in collaboration with Drexel University (Prof. Julia Stoyanovich). The new version includes access control.

6. New Results

6.1. Highlights of the Year

Victor Vianu was elected member of Academia Europaea.
BEST PAPER AWARD :

[21] **Implicational Relevance Logic is 2-ExpTime-Complete in Joint 25th International Conference on Rewriting Techniques and Applications and 12th International Conference on Typed Lambda Calculi and Applications.** S. SCHMITZ.

6.2. Distributed data management

Participants: Serge Abiteboul, Émilien Antoine, Victor Vianu.

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 [15], we study deduction in the presence of inconsistencies. Following previous works, we capture deduction via datalog programs and inconsistencies through violations of functional dependencies (FDs). We study and compare two semantics for datalog with FDs: the first, of a logical nature, is based on inferring facts one at a time, while never violating the FDs; the second, of an operational nature, consists in a fixpoint computation in which maximal sets of facts consistent with the FDs are inferred at each stage. Both semantics are nondeterministic, yielding sets of possible worlds. We introduce a PTIME (in the size of the extensional data) algorithm, that given a datalog program, a set of FDs and an input instance, produces a c-table representation of the set of possible worlds. Then, we propose to quantify nondeterminism with probabilities, by means of a probabilistic semantics. We consider the problem of capturing possible worlds along with their probabilities via probabilistic c-tables. We then study classical computational problems in this novel context. We consider the problems of computing the probabilities of answers, of identifying most likely supports for answers, and of determining the extensional facts that are most influential for deriving a particular fact. We show that the interplay of recursion and FDs leads to novel technical challenges in the context of these problems.

Jakub Kalas (ENS Cachan) spent 4 months in the team working on Personal Information Management Systems, using primarily positioning from data mobile phone and data from search engines.

6.3. Query Processing for the Web

Participants: Johann Brault-Baron, Arnaud Durand, Nadime Francis, 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 almost linear preprocessing time and constant enumeration delay for first-order queries over structures of low degree [18]. We also presented a survey about this work at the Intl. Symp. on Theoretical Aspects of Computer Science (STACS) [22].

We have also been interested in querying data structured as graphs, which is nowadays spreading on the Web. Examples are social networks, linked data and the semantic web, via the RDF format. We have tackled the problem of answering queries over graph databases which are available only through a given set of views. This is a common situation in many applications where access to data needs to be either controlled or optimized. In [19] we have studied when it is possible to rewrite over the views queries issued on the original data, and which query languages are needed for this purpose. We have considered views and queries expressed as Regular path queries, a very common graph query language.

6.4. Complexity in Counter Systems and Substructural Logics

Participant: Sylvain Schmitz.

The ties between propositional substructural logics (like linear logic, relevance logic, affine logic, etc.) on the one hand and extensions of vector addition systems on the other hand have long been known, as they lie for instance at the heart of undecidability proof of provability in linear logic. In a series of papers we recently revisited these connections with an eye on complexity issues. This allowed us to prove tight complexity bounds on provability in affine and contractive fragments of linear logic [20], in affine $(!, \oplus)$ -Horn linear logic [16], and in implicative relevance logic [21] (an open problem for more than 25 years, with consequences on type inhabitation in the λI -calculus). Our work also yields a new Tower lower bound on reachability in branching vector addition systems [20], which entails the same lower bound for logics on XML trees [4], for which decidability is still open.

Although the connection with data logics might not seem obvious at first, the models of counter systems considered in these papers are tightly connected with logics for XML processing [5], [4]. Further investigations in the relationships between data logics, substructural logics, and counter systems are the main thrust behind the just accepted ANR PRODAQ project (see Section 8.1.1).

6.5. Incomplete Databases

Participants: Nadime Francis, Cristina Sirangelo.

Incomplete databases appear in several different scenarios. Intuitively, pieces of information might not be available, or can get lost due to failures in storage or transmission. Alternatively, some complex data management tasks, such as data integration or data exchange, use incomplete databases as a model for databases with missing or unspecified information. In the context of the Web, these tasks have become even more crucial, which increased the need to handle incomplete databases. Given an incomplete database, one of the first questions to answer is that of consistency: can we make sure that the incomplete database can be completed as a real database conforming to some specified schema.

Together with Claire David and Filip Murlak, we have considered this problem when incomplete instances are represented as incomplete XML documents, where labels and nodes might be missing, and we additionally assume the DOM semantics, meaning that nodes never lose their identity (otherwise, they are considered completely lost). These are further modeled as injective tree patterns using child and descendant relations. In [17], we close the question of the complexity of checking the consistency of such patterns with regards to a fixed regular tree language: it is polynomial for patterns that do not use child edges, and for patterns that use both, it is already NP-complete for patterns using at most two descendant edges per branch, the case for at most one descendant edge being already known to be polynomial.

In [12] we have studied the feasibility of query answering in the presence of incomplete information in data. In particular we have generalized conditions allowing classical query evaluation techniques to be applicable also in the presence of incompleteness. Our results show that conditions found in some of our previous work can be significantly relaxed so as to account for more complex semantics of incompleteness, originating in the fields of logic programming, programming semantics and data exchange.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

The CIFRE scholarship of David Montoya started in 2014, with Sinovia, Cofely Ineo (group GDF Suez). The topic is on analysis of multimodal itineraries.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

Acronym: PRODAQ

Title: Proof systems for Data Queries

Coordinator: Sylvain Schmitz

Duration: January 2015 – September 2019

Abstract: The project aims at developing proof systems for data logics. It is at the interface between several research communities in database theory, infinite-state system verification and proof theory. The main thrust behind the project is the investigation of proof-theoretic tools for data logic, using in particular insights from substructural logics, and using counter systems as a means to obtain algorithms and complexity results.

8.2. International Initiatives

8.2.1. Inria International Partners

8.2.1.1. Declared Inria International Partners

Victor Vianu, UC San Diego, USA. Chaire Inria depuis 2013.

8.3. International Research Visitors

8.3.1. Visits of International Scientists

- Thomas Schwentick
Subject: Automata methods for data processing
Institution: Dortmund University, Dortmund, Germany.
- Sławomir Lasota
Subject: Reasoning with data using sets with atoms
Institution: Warsaw University, Warsaw, Poland.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. general chair, scientific chair

Luc Segoufin was the general chair of the conference Highlights of Logic, Games and Automata, held in Paris in September 2014.

Serge Abiteboul co-chaired with Patrick Flandrin the Conference: La Datamasse : directions et enjeux pour les données massives, Sciences Academy, Paris 2014

9.1.1.2. member of the organizing committee

Sylvain Schmitz and Luc Segoufin were part of the organizing committee for the conference Highlights of Logic, Games and Automata, held in Paris in september 2014.

9.1.2. Scientific events selection

9.1.2.1. member of the conference program committee

Luc Segoufin: Intl. Conf. on Foundations of Software Science and Computation Structures (FoSSaCS'14).

Victor Vianu: Joint meeting of Conf. on Computer Science Logic (CSL'14) and ACM/IEEE Symp. on Logic in Computer Science (LICS'14). Intl. Conf. on Data, Processes, and Software Systems, Hangzhou, China. Intl. Symp. on Foundations of Information and Knowledge Systems (FoIKS'14). SIGMOD PhD Symposium 2014.

Serge Abiteboul: Intl. Conf. on Digital Intelligence. Buda Intl. Workshop on Big Uncertain Data. Intl. Conf. on Extending database Technology. Intl. Conf. on Principles of database systems (PODS'14), Award committee. Intl. Conf. on Data engineering (ICDE'2015). Intl. Conf. on Database Theory (ICDT'2015), Award committee. Intl. Conf. on the Management of data (2015), Demonstrations.

Cristina Sirangelo: Intl. Conf. on Database Theory (ICDT'15). Symp. on Principles of Database Systems (PODS'15).

9.1.3. Journal

9.1.3.1. member of the editorial board

Victor Vianu: Editor-in-Chief of the Journal of the ACM Associate Editor of ACM Transactions on Computational Logic (TOCL)

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

E-learning

Mooc: Serge Abiteboul (with Benjamin Nguyen), Bases de données relationnelles, 6 weeks, Association Sillages, june 2014, accredited by Ocean

Serge Abiteboul (with Benjamin Nguyen and Yannick Lebras) Introduction aux Bases de Données Relationnelles, course for the CPGE, pdf, <http://abiteboul.com/Lili/bdrelationnelles.pdf>

Master : Cristina Sirangelo, Complexité avancée, 36 hours ETD, M1, MPRI, France

Master : Cristina Sirangelo, Algorithmique, 25 hours ETD, Préparation à l'agrégation, École Normale Supérieure de Cachan, France

Licence : Cristina Sirangelo, Bases de données, 24 hours ETD, L3, École Normale Supérieure de Cachan, France

Licence : Cristina Sirangelo, Projet de bases de données, 22 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

Doctorat : Cristina Sirangelo, Recherche sur le Web, 18 hours ETD, École Normale Supérieure de Cachan, France

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

Master : Serge Abiteboul (with Pierre Senellart), Web data management, MPRI Paris

Licence : Serge Abiteboul, Bases de données pour les professeurs de CPGE, ENS Cachan, 3 heures.

Master : Sylvain Schmitz, Langages formels, 30 hours ETD, M2 Préparation à l'agrégation, École Normale Supérieure de Cachan, France

Master : Sylvain Schmitz, Logical and Computational Structures for Linguistic Modeling, 18 hours ETD, M2 MPRI, France

9.2.2. Supervision

HdR : Cristina Sirangelo, Representing and querying incomplete information: a data interoperability perspective, Ecole Normale Supérieure de Cachan, December 5th 2014 [11]

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

PhD in progress: Nathan Grosshans, branching program, 15/09/2014, Luc segoufin and Pierre McKenzie (Montreal)

PhD in progress: David Montoya, Personal information management systems, 01/02/2014, Serge Abiteboul

9.2.3. Juries

Luc Segoufin was a member of the HDR jury of Cristina sirangelo.

Cristina Sirangelo was member of a hiring committee for an assistant professor position at Polytech Paris-Sud.

Sylvain Schmitz was a member in two hiring committees for "agrégé préparateur" assistant professor positions at ENS Cachan.

9.3. Popularization

Serge Abiteboul participated in 2014 to the following reports of the Conseil national du numérique: Digital education (Jules Ferry 3.0, Bâtir une école créative et juste dans un monde numérique) and Platform neutrality (Neutralité des plateformes)

Serge Abiteboul started in 2014 a Blog invited by Le monde newspaper: <http://binaire.blog.lemonde.fr>

Serge Abiteboul coordinated the production of a report on Education to Big data [23].

Serge Abiteboul was audited on the Digital Risk by Office Parlementaire d'Évaluation des Choix Scientifiques et Technologiques.

Serge Abiteboul was interviewed on radio notably by: La Matinale d'Europe 1, BFM Business, Radio Campus, Autour de la question, RFI. He had articles notably in: Le monde, Fenêtres sur Cours, Sciences et Avenir, Le Parisien Magazine. He gave talks to Cap digital, Assises Médias Entreprises, Conférence Idées, Université d'été de l'Institut des Actuaire, 10 ans de la Fondation Steri, Orange Labs, Google.

10. Bibliography

Major publications by the team in recent years

- [1] S. ABITEBOUL, I. MANOLESCU, P. RIGAU, 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° 4
- [3] P. BARCELÓ, L. LIBKIN, A. POGGI, C. SIRANGELO. *XML with incomplete information*, in "J. ACM", 2010, vol. 58, n° 1
- [4] 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° 4

- [5] M. BOJAŃCZYK, A. MUSCHOLL, T. 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
- [6] M. BOJAŃCZYK, L. SEGOUFIN, H. STRAUBING. *Piecewise testable tree languages*, in "Logical Methods in Computer Science (LMCS)", 2012, vol. 8, 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] C. SIRANGELO. *Representing and Querying Incomplete Information: a Data Interoperability Perspective*, Ecole Normale Supérieure de Cachan, December 2014, Habilitation à diriger des recherches, <https://hal.inria.fr/tel-01092547>

Articles in International Peer-Reviewed Journals

- [12] A. GHEERBRANT, L. LIBKIN, C. SIRANGELO. *Naive Evaluation of Queries over Incomplete Databases*, in "ACM Transactions on Database Systems (TODS)", December 2014, 42 p. , <https://hal.inria.fr/hal-01101184>
- [13] C. HAASE, S. SCHMITZ, P. SCHNOEBELEN. *The Power of Priority Channel Systems*, in "Logical Methods in Computer Science", December 2014, vol. 10, n^o 4:4, pp. 1–39 [DOI : 10.2168/LMCS-10(4:4)2014], <https://hal.inria.fr/hal-01091086>

Invited Conferences

- [14] S. SCHMITZ. *Complexity Bounds for Ordinal-Based Termination*, in "8th International Workshop on Reachability Problems", Oxford, United Kingdom, J. OUAKNINE, I. POTAPOV, J. WORRELL (editors), Lecture Notes in Computer Science, Springer, September 2014, vol. 8762, pp. 1–19 [DOI : 10.1007/978-3-319-11439-2_1], <https://hal.inria.fr/hal-01076701>

International Conferences with Proceedings

- [15] S. ABITEBOUL, D. DEUTCH, V. VIANU. *Deduction with Contradictions in Datalog*, in "International Conference on Database Theory", Athens, Greece, 2014, <https://hal.inria.fr/hal-00923265>
- [16] J.-B. COURTOIS, S. SCHMITZ. *Alternating Vector Addition Systems with States*, in "39th International Symposium on Mathematical Foundations of Computer Science", Budapest, Bulgaria, E. CSUHAI-VARJÚ,

- M. DIETZFELBINGER, Z. ÉSIK (editors), Lecture Notes in Computer Science, Springer, August 2014, vol. 8634, pp. 220–231 [DOI : 10.1007/978-3-662-44522-8_19], <https://hal.inria.fr/hal-00980878>
- [17] C. DAVID, N. FRANCIS, F. MURLAK. *Consistency of injective tree patterns*, in "Foundations of Software Technology and Theoretical Computer Science", New Dehli, India, December 2014, <https://hal.inria.fr/hal-01094596>
- [18] A. DURAND, N. SCHWEIKARDT, L. SEGOUFIN. *Enumerating answers to first-order queries over databases of low degree*, in "Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS'14," Snowbird, United States, June 2014, pp. 121–131 [DOI : 10.1145/2594538.2594539], <https://hal.inria.fr/hal-01070898>
- [19] N. FRANCIS, L. SEGOUFIN, C. SIRANGELO. *Datalog Rewritings of Regular Path Queries using Views*, in "Proc. 17th International Conference on Database Theory (ICDT)", Athens, Greece, March 2014, pp. 107–118 [DOI : 10.5441/002/ICDT.2014.14], <https://hal.inria.fr/hal-01070903>
- [20] R. LAZIĆ, S. SCHMITZ. *Non-Elementary Complexities for Branching VASS, MELL, and Extensions*, in "Joint meeting of the Twenty-Third EACSL Annual Conference on Computer Science Logic and the Twenty-Ninth Annual ACM/IEEE Symposium on Logic in Computer Science", Vienna, Austria, ACM, July 2014, pp. 61:1–61:1 [DOI : 10.1145/2603088.2603129], <https://hal.inria.fr/hal-01076694>
- [21] *Best Paper*
S. SCHMITZ. *Implicational Relevance Logic is 2-ExpTime-Complete*, in "Joint 25th International Conference on Rewriting Techniques and Applications and 12th International Conference on Typed Lambda Calculi and Applications", Vienna, Austria, G. DOWEK (editor), Lecture Notes in Computer Science, Springer, July 2014, vol. 8560, pp. 395–409 [DOI : 10.1007/978-3-319-08918-8_27], <https://hal.inria.fr/hal-01076698>.
- [22] L. SEGOUFIN. *A glimpse on constant delay enumeration*, in "31st International Symposium on Theoretical Aspects of Computer Science (STACS 2014)", Lyon, France, March 2014, pp. 13–27 [DOI : 10.4230/LIPIcs.STACS.2014.13], <https://hal.inria.fr/hal-01070893>

Research Reports

- [23] S. ABITEBOUL, F. BANCILHON, F. BOURDONCLE, S. CLEMENCON, C. DE LA HIGUERA, G. SAPORTA, F. FOGELMAN SOULIÉ. *L'émergence d'une nouvelle filière de formation : « data scientists »*, Inria Saclay, April 2014, <https://hal.inria.fr/hal-01092062>
- [24] F. JACQUEMARD, L. SEGOUFIN, J. DIMINO. *FO2($<, +1, \cdot$) on data trees, data tree automata and an branching vector addition systems*, Inria Saclay, January 2015, 32 p. , <https://hal.inria.fr/hal-00769249>

Other Publications

- [25] S. ABITEBOUL, B. ANDRÉ, D. KAPLAN. *Managing your digital life with a Personal information management system*, September 2014, <https://hal.inria.fr/hal-01068006>