



IN PARTNERSHIP WITH:
CNRS

**Ecole normale supérieure de
Cachan**

Activity Report 2015

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. Highlights of the Year	2
6. New Results	3
6.1. Specification and verification of data-driven systems	3
6.2. Query processing for the Web	3
6.3. Distributed knowledge base.	3
7. Bilateral Contracts and Grants with Industry	3
8. Partnerships and Cooperations	4
8.1. National Initiatives	4
8.2. European Initiatives	4
8.3. International Research Visitors	5
8.3.1. Visits of International Scientists	5
8.3.2. Visits to International Teams	5
9. Dissemination	5
9.1. Promoting Scientific Activities	5
9.1.1. Scientific events organisation	5
9.1.2. Scientific events selection	5
9.1.2.1. Member of the conference program committees	5
9.1.2.2. Reviewer	5
9.1.3. Journal	5
9.1.4. Invited talks	5
9.1.5. Leadership within the scientific community	6
9.1.6. Scientific expertise	6
9.1.7. Research administration	6
9.2. Teaching - Supervision - Juries	6
9.2.1. Teaching	6
9.2.2. Supervision	6
9.2.3. Juries	6
9.3. Popularization	7
10. Bibliography	7

Project-Team DAHU

Creation of the Project-Team: 2009 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
- 3.1.4. - Uncertain data
- 3.1.5. - Control access, privacy
- 3.1.9. - Database
- 4.7. - Access control

Other Research Topics and Application Domains:

- 9.8. - Privacy

1. Members

Research Scientists

- Luc Segoufin [Team leader, Inria, Senior Researcher, HdR]
- Serge Abiteboul [Inria, Senior Researcher, HdR]
- Victor Vianu [Inria, Advanced Research position]

Faculty Members

- Sylvain Schmitz [ENS Cachan, Associate Professor]
- Cristina Sirangelo [ENS Cachan, Associate Professor, until Sep 2015, HdR]

PhD Students

- Nathan Grosshans [ENS Cachan, granted by Digicosme]
- David Montoya [ENS Cachan, granted by CIFRE]
- Karima Rafes [ENS Cachan]
- Alexandre Vigny [Univ. Paris VII, from Sep 2015]
- Nadime Francis [ENS Cachan, until Oct 2015]

Post-Doctoral Fellow

- Matthias Niewerth [Inria, from Oct 2015]

Administrative Assistant

- Thida Iem [Inria]

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. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

Best student paper award for Nadime Francis [22] at the conference ICDT'15.

Luc Segoufin and Victor Vianu obtained the ACM Alberto O. Mendelzon PODS Test of Time Award in 2015.

6. New Results

6.1. Specification and verification of data-driven systems

Process-centric views of data-driven workflows. Declarative, data-aware workflow models are becoming increasingly pervasive. While these have numerous benefits, views describing valid sequences of tasks are also useful to provide stake-holders with high-level descriptions of the workflow. In [23], we study the problem of recovering process-centric views from declarative, data-aware workflow specifications. The views are most naturally specified by finite-state transition systems describing regular languages. The results characterize when process-centric views of artifact systems are regular, with both linear and branching-time semantics.

Complexity in counter systems and in proof systems. The static analysis of queries on XML trees and data streams relies in a majority of cases on decision procedures expressed in terms of formal systems like counter systems or proof systems. For instance, two-variables first-order data queries on words can be related to reachability in vector addition systems (VAS), and the same queries on trees to reachability in a branching extension of VAS. We have fundamental results on the computational complexity of these problems, including the first explicit upper bound for reachability in VAS [24] and the best known lower bound for reachability in branching VAS [17] (where it is currently unknown whether reachability is decidable at all). We have furthermore defined a first sequent calculus for a modal data logic [29] as preliminary groundwork for the ANR PRODAQ project on proof systems for data queries.

6.2. Query processing for the Web

Query languages for graph databases. Graph-structured data on the Web can be found in emerging applications such as *RDF* and *linked data*, or *social networks*. Classical database languages are not suitable to query such data, essentially because they do not allow to (easily) express simple connectivity queries, which are the basic building block in graph navigation. We use Regular Path Queries, computing pairs of nodes reachable via a path satisfying a regular expression. We have tackled the problem of answering queries over graph databases which are available only through a given set of views. We have shown that in the “asymptotic case”, i.e. when the query is large enough relative to the view definition, it is decidable whether the view determines the query [22].

6.3. Distributed knowledge base.

Webdamlog The Webdamlog language is an extension of datalog to the distributed context, with *delegation* as the main novelty. A summary of the project was presented in [20].

We introduced an access control mechanism based on provenance in [26]. This access control is designed for a distributed setting. Peers getting data are also willing to enforce access control on that data, so that the owner of the data keeps some control over it when the data is passed around in the network. A second version of Webdamlog was developed in 2015 at Drexel, primarily by Vera Moffit also as part of her thesis (in collaboration with S. Abiteboul). It includes access control mechanism.

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 and the integration of itinerary data with other personal data.

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. European Initiatives

8.2.1. MEALS

Title: Mobility between Europe and Argentina applying Logics to Systems

Programm: FP7

Duration: October 2011 - September 2015

Coordinator: Université de la sarre

Partners:

Imperial College of Science, Technology and Medicine (United Kingdom)

Rheinisch-Westfaelische Technische Hochschule Aachen (Germany)

Technische Universiteit Eindhoven (Netherlands)

Technische Universitaet Dresden (Germany)

University of Leicester (United Kingdom)

Universitaet Desarlandes (Germany)

Inria contact: Castuscia Palamidessi

Computing systems are getting ever more ubiquitous, making us dependent on their proper functioning. Therefore we require that they are correct (i.e. they conform their intended behavior), safe (i.e. its operation does not have catastrophic consequences), reliable, available to provide the intended service, and secure (i.e., no user without appropriate clearance can access or modify protected data). Guarantees for such characteristics rely on rigid specification and analysis techniques for both the required system functionality as well as its behavior. Formal methods provide a mathematical approach to model, understand, and analyze systems, especially at early development stages. In this project we focus on three aspects of formal methods: specification, verification, and synthesis. We consider the study of both qualitative behavior and quantitative behavior (extended with probabilistic information). We aim to study formal methods in all their aspects: foundations (their mathematical and logical basis), algorithmic advances (the conceptual basis for software tool support) and practical considerations (tool construction and case studies). The MEALS project includes five tightly interconnected thematic work packages. They focus on quantitative analysis of concurrent program behaviour (WP1), reasoning tasks for specification and verification (WP2), security and information flow properties (WP3), synthesis in model-based systems engineering (WP4) and foundations for the elaboration and analysis of requirements specifications (WP5). The crosscutting concern of all these work packages is the development of formal techniques for the specification, verification and synthesis of dependable ubiquitous computing systems. Five carefully planned MEALS gatherings and workshops give the project an effective structure for knowledge transfer, community building, and result dissemination, aimed at a sustained transcontinental collaboration.

8.3. International Research Visitors

8.3.1. Visits of International Scientists

Victor Vianu, June 1st till December 31st, UC San diego

8.3.2. Visits to International Teams

8.3.2.1. Research stays abroad

Sylvain Schmitz visited the University of Warwick for six months thanks to a grant by the Leverhulme Trust, from February 1st to July 31st, 2015.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. General chair, scientific chair

Serge Abiteboul participated with Olivier Baudon to the organization in 2015 of the seminar of the Société Informatique de France “Enseignement de l’informatique pour les humanités et les sciences sociales”.

9.1.2. Scientific events selection

9.1.2.1. Member of the conference program committees

Serge Abiteboul: Data engineering (ICDE’2015); Management of data (Sigmod’2015); Very Large Database (VLDB’2015); Web Database (WebDB’2015). In 2015, Serge Abiteboul also participated in Awards committees for the French Academy of Sciences and the British Royal Society (Milner Award).

Sylvain Schmitz: Reachability Problems (RP’15).

Cristina Sirangelo: ICDT’15. PODS’15. PODS’16.

Victor Vianu: ICDT’15.

9.1.2.2. Reviewer

The members of the team reviewed numerous papers for numerous international conferences and journals.

9.1.3. Journal

9.1.3.1. Member of the editorial boards

Victor Vianu: Editor-in-Chief of JACM, Associate Editor of ACM TOCL, Editor of the Database Theory column in SIGACT News.

9.1.4. Invited talks

Victor Vianu gave two invited talks: One at the Alberto. O. Mendelzon Workshop, Lima, Peru, May 6-8 2015. A second one at the Martin Abadi Honoris Causa Symposium, ENS Cachan, June 2015.

Sylvain Schmitz gave two invited talks, at ACTS 2015 in Chennai, India, and at the DIMAP Logic Day 2015 in Warwick, UK.

Serge Abiteboul gave invited talks at the Advances in Database and Information Systems Conference (keynote - Poitiers), Profile, Predict, and Prevent Conference (Law, Paris), TTOW Conference (History, Paris), DSAAI Conference (Paris), Filocs Workshop (Tel Aviv), Data science symposium (Rennes), CNRS Mastodons Workshop (Paris), Inria EPFL Workshop (Lausanne), UPMC Seminar (Paris 6)...

He also gave invited talks in industry (Thalles, Accenture).

9.1.5. Leadership within the scientific community

Serge Abiteboul was nominated co-chair of the Parity and Equality committee of Inria.

9.1.6. Scientific expertise

Sylvain Schmitz reviewed grant proposals for the Agence Nationale de la Recherche (ANR) and the Austrian Science Fund (FWF).

9.1.7. Research administration

Since 2015 Luc Segoufin is an elected member of the CNRS of Inria.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence : Cristina Sirangelo, Databases, 24h L3, ENS Cachan
 Licence : Cristina Sirangelo, Database project, 22.5h, L3, ENS Cachan
 Master : Cristina Sirangelo, Algorithms, 37h, M2, ENS Cachan
 Master : Cristina Sirangelo, Advanced complexity theory, 30h M1, ENS Cachan
 Graduate : Cristina Sirangelo, Web sites, 18h, Ecole doctorale ENS Cachan
 Graduate : Cristina Sirangelo, Databases, 18h, Ecole doctorale ENS Cachan
 Graduate : Cristina Sirangelo, Web search, 18h, Ecole doctorale ENS Cachan
 Licence : Sylvain Schmitz, Complexity, 22.5h, L3, ENS Cachan, France
 Master : Sylvain Schmitz, Introduction to Verification, 19h, M1, ENS Cachan, France
 Master : Sylvain Schmitz, Formal Languages, 37.5h, M2, ENS Cachan, France
 Master : Sylvain Schmitz, Logical and Computational Structures for Linguistic Modeling, 18h, M2, ENS Cachan, France
 Master : Serge Abiteboul, Web data management, 15h, MPRI
 Licence : Serge Abiteboul, Relational databases, 30h, ENS Cachan

E-learning

Mooc Bases de Données Relationnelles - Bador (in French), Inria Fun, Serge Abiteboul, Benjamin Nguyen and Philippe Rigaux, Start January 2016; about 6000 students registered, 6+ weeks. Target audience: students in L3/Master, engineers and scientists using databases.

9.2.2. Supervision

PhD : Nadime Francis, View-based query determinacy and rewritings over graph databases, 08/12/2015 [11], Luc Segoufin and Cristina Sirangelo
 PhD in progress: Nathan Grosshans, branching program, 01/09/2014, Luc Segoufin and Pierre McKenzie (University of Montreal)
 PhD in progress: David Montoya, Personal information management systems, 01/02/2014, Serge Abiteboul
 PhD in progress: Alexandre Vigny, enumeration, 01/09/2015, Luc Segoufin and Arnaud Durand
 PhD in progress: Karima Rafes, Web semantic and Internet of Thing, 01/02/2015, Serge Abiteboul
 PhD in progress: Simon Halfon, Well quasi orders, 01/09/2015, Sylvain Schmitz and Philippe Schnoebelen

9.2.3. Juries

Cristina Sirangelo was in the jury for ingénieur d'étude of ENS-Cachan in July 2015.

Sylvain Schmitz was on the Ph.D. committee of Jérôme Kirman, Université de Bordeaux, in December 2015. Serge Abiteboul was on the Ph.D. committee of Stamatis Zampetakis, University of Paris-Saclay, in September 2015.

9.3. Popularization

For the last couple of years, Serge Abiteboul has initiated and is coordinating an invited blog for Le Monde newspaper, namely Binaire, <http://binaire.blog.lemonde.fr>. He is the initiator of “The Cabale Informatique de France”, a joint action between the Société Informatique de France and Wikipédia France to improve the quality of Wikipédia pages about Computer Science in France.

Serge Abiteboul participated in a documentary "Nos ordinateurs ont-ils la mémoire courte" (directed by Vincent Amouroux, Zed production) that was showed on Arte in September 2015. He wrote popularization articles, in particular, [13] (on Personal information systems), [28] (on the Responsibility of algorithms, and notably data analysis algorithms; short version in French in Le Monde) and [27] (on Digital humanities).

Serge Abiteboul gave talks to Congrès National des Profs de Physique Chimie (La Rochelle), Forum Big Data (La Vilette), Congrès APLCPCGE ("prépas", Paris), UPMC Colloquium (Paris)... He was audited for the French Parliament (OPECST, Big data and Agriculture).

In 2015, Serge Abiteboul was a member of the "Conseil national du numérique".

10. 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] V. BARANY, B. T. CATE, L. SEGOUFIN. *Guarded negation*, in "Journal of the ACM", 2015, vol. 62, n^o 3, 24 p. , <https://hal.inria.fr/hal-01184763>
- [4] P. BARCELÓ, L. LIBKIN, A. POGGI, C. SIRANGELO. *XML with incomplete information*, in "J. ACM", 2010, vol. 58, n^o 1
- [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] N. FRANCIS, L. SEGOUFIN, C. SIRANGELO. *Datalog Rewritings of Regular Path Queries using Views*, in "Logical Methods in Computer Science (LMCS)", December 2015, vol. 11, n^o 4, <https://hal.inria.fr/hal-01248391>

- [9] R. LAZIĆ, S. SCHMITZ. *Non-Elementary Complexities for Branching VASS, MELL, and Extensions*, in "ACM Transactions on Computational Logic", May 2015, vol. 16, n^o 3:20, pp. 1–30 [DOI : 10.1145/2733375], <https://hal.archives-ouvertes.fr/hal-01168290>
- [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] N. FRANCIS. *View-based query determinacy and rewritings over graph databases*, Université Paris-Saclay, December 2015, <https://hal.archives-ouvertes.fr/tel-01247115>

Articles in International Peer-Reviewed Journals

- [12] S. ABITEBOUL, Y. AMSTERDAMER, D. DEUTCH, T. MILO, P. SENELLART. *Optimal Probabilistic Generation of XML Documents*, in "Theory of Computing Systems", 2015 [DOI : 10.1007/s00224-014-9581-5], <https://hal.inria.fr/hal-01248048>
- [13] S. ABITEBOUL, B. ANDRÉ, D. KAPLAN. *Managing your digital life with a Personal information management system*, in "Communications of the ACM", May 2015, vol. 58, n^o 5, pp. 32-35, <https://hal.inria.fr/hal-01068006>
- [14] S. ABITEBOUL, P. BOURHIS, V. VIANU. *Highly Expressive Query Languages for Unordered Data Trees*, in "Theory of Computing Systems", 2015, 30 p. , <https://hal.inria.fr/hal-01167068>
- [15] V. BARANY, B. T. CATE, L. SEGOUFIN. *Guarded negation*, in "Journal of the ACM", 2015, vol. 62, n^o 3, 24 p. , <https://hal.inria.fr/hal-01184763>
- [16] N. FRANCIS, L. SEGOUFIN, C. SIRANGELO. *Datalog Rewritings of Regular Path Queries using Views*, in "Logical Methods in Computer Science (LMCS)", December 2015, vol. 11, n^o 4, <https://hal.inria.fr/hal-01248391>
- [17] R. LAZIĆ, S. SCHMITZ. *Non-Elementary Complexities for Branching VASS, MELL, and Extensions*, in "ACM Transactions on Computational Logic", May 2015, vol. 16, n^o 3:20, pp. 1–30 [DOI : 10.1145/2733375], <https://hal.archives-ouvertes.fr/hal-01168290>
- [18] T. PLACE, L. SEGOUFIN. *DECIDING DEFINABILITY IN FO 2 (<math>\langle v, h) ON TREES*, in "Logical Methods in Computer Science (LMCS)", 2015, vol. 11, n^o 3 [DOI : 10.2168/LMCS-11(3:5)2015], <https://hal.inria.fr/hal-01223373>
- [19] L. SEGOUFIN. *Constant Delay Enumeration for Conjunctive Queries*, in "SIGMOD record", 2015, vol. 44, n^o 1, 8 p. , <https://hal.inria.fr/hal-01184764>

Invited Conferences

- [20] S. ABITEBOUL. *The Story of Webdamlog*, in "19th East-European Conference on Advances in Databases and Information Systems", Futuroscope, Poitiers, France, 2015, <https://hal.inria.fr/hal-01195717>

International Conferences with Proceedings

- [21] A. DAWAR, L. SEGOUFIN. *Capturing MSO with One Quantifier*, in "Fields of Logic and Computation II - Essays Dedicated to Yuri Gurevich on the Occasion of His 75th Birthday", Berlin, Germany, LNCS, 2015, vol. 9300 [DOI : 10.1007/978-3-319-23534-9_8], <https://hal.inria.fr/hal-01223378>
- [22] N. FRANCIS. *Asymptotic Determinacy of Path Queries using Union-of-Paths Views*, in "18th International Conference on Database Theory (ICDT 2015)", Brussels, Belgium, March 2015 [DOI : 10.4230/LIPIcs.ICDT.2015.44], <https://hal.inria.fr/hal-01150780>
- [23] A. KOUTSOS, V. VIANU. *Process-Centric Views of Data-Driven Business Artifacts*, in "International Conference on Database Theory (ICDT)", Bruxelles, Belgium, 2015, 18 p. [DOI : 10.4230/LIPIcs.ICDT.2015.247], <https://hal.inria.fr/hal-01167033>
- [24] J. LEROUX, S. SCHMITZ. *Demystifying Reachability in Vector Addition Systems*, in "LICS 2015", Kyoto, Japan, July 2015, pp. 56–67 [DOI : 10.1109/LICS.2015.16], <https://hal.archives-ouvertes.fr/hal-01168388>
- [25] D. MONTOYA, S. ABITEBOUL, P. SENELLART. *Hup-Me: Inferring and Reconciling a Timeline of User Activity from Rich Smartphone Data*, in "ACM SIGSPATIAL, International Conference on Advances in Geographic Information Systems", Seattle, WA, United States, November 2015, <https://hal.inria.fr/hal-01189093>
- [26] V. ZAYCHIK MOFFIT, J. STOYANOVICH, S. ABITEBOUL, G. MIKLAU. *Collaborative Access Control in WebdamLog*, in "Proceeding of the ACM Sigmod Conference on Data Management", Melbourne, Australia, ACM Sigmod, 2015, <https://hal.inria.fr/hal-01136473>

Scientific Popularization

- [27] S. ABITEBOUL, F. HACHEZ-LEROY. *Humanités numériques*, in "Encyclopédie de l'humanisme méditerranéen", June 2015, <https://hal.inria.fr/hal-01120259>

Other Publications

- [28] S. ABITEBOUL, J. STOYANOVICH. *Data, Responsibly*, 2015, ACM Sigmod Blog, 20 November 2015, <https://hal.inria.fr/hal-01248054>
- [29] D. BAELEDE, S. LUNEL, S. SCHMITZ. *A Sequent Calculus for a Modal Logic on Finite Data Trees*, September 2015, working paper or preprint, <https://hal.inria.fr/hal-01191172>