IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Cachan

Activity Report 2016

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

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. Personal information management. 3
7. Bilateral Contracts and Grants with Industry 4
8. Partnerships and Cooperations 4
8.1. National Initiatives 4
8.2. International Research Visitors 4
9. Dissemination 4
9.1. Promoting Scientific Activities 4
9.1.1. Scientific events organisation 4
9.1.2. Scientific events selection 4
9.1.2.1. Member of the conference program committees 4
9.1.2.2. Reviewer 4
9.1.3. Journal 4
9.1.4. Invited talks 5
9.1.5. Leadership within the scientific community 5
9.1.6. Scientific expertise 5
9.1.7. Research administration 5
9.2. Teaching - Supervision - Juries 5
9.2.1. Teaching 5
9.2.2. Supervision 5
9.2.3. Juries 5
9.3. Popularization 6
10. Bibliography 6

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
7.4. - Logic in Computer Science

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]

Faculty Member

Sylvain Schmitz [ENS Cachan, Associate Professor]

PhD Students

Nathan Grosshans [CNRS]
Anthony Lick [CNRS]
David Montoya [Sinovia, granted by CIFRE]
Karima Rafes
Alexandre Vigny [Univ. Paris VII]
Su Yang [ENSM Paris, from Sep 2016]

Post-Doctoral Fellow

Matthias Niewerth [Inria, until Mar 2016]

Administrative Assistant

Thida Iem [Inria]

Others

Thomas Pellissier-Tanon [ENS Lyon, Intern, until Jun 2016]
Pierre Senellart [Telecom Paristech, Professor]
Victor Vianu [UC San Diego, Professor]

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

Awards

Luc Segoufin together with Mikolaj Bojanczyk, Claire David, Anca Muscholl, and Thomas Schwentick obtained the ACM Alberto O. Mendelzon PODS Test of Time Award in 2016.

6. New Results

6.1. Specification and verification of data-driven systems

Verification of Hierarchical Artifact Systems

Data-driven workflows, of which "business artifacts" are a prime exponent, have been successfully deployed in practice, adopted in industrial standards, and have spawned a rich body of research in academia, focused primarily on static analysis. Over the past few years, we have embarked upon a study of the verification problem for artifact systems. This is a challenging problem because of the presence of unbounded data. In order to deal with the resulting infinite-state system, we developed in earlier work a symbolic approach allowing a reduction to finite-state model checking and yielding a pspace verification algorithm for the simplest variant of the model (no database dependencies and uninterpreted data domain). Subsequently, we extended our approach to allow for database dependencies and numeric data testable by arithmetic constraints. In [19], we make significant progress on several fronts, by considering a much richer and more realistic model than in previous work, incorporating core elements of IBM's successful Guard-Stage-Milestone model. In particular, the model features task hierarchy, concurrency, and richer artifact data. It also allows database key and foreign key dependencies, as well as arithmetic constraints. The results require qualitatively novel techniques, because the reduction to finite-state model checking used in previous work is no longer possible. Instead, the richer model requires the use of a hierarchy of Vector Addition Systems with States. The arithmetic constraints are handled using quantifier elimination techniques, adapted to our setting.

Process-centric views of data-driven workflows.

We also studied the models of *data Petri nets* and *ν-Petri nets*. While these models were introduced in the verification community to analyse protocols and process algebra, they can also be seen as (very limited) data-driven workflows with only unary predicates. Our results this year show that various boundedness problems (e.g. can the database grow unbounded?) are decidable in data Petri nets [22], and pinpoint the exact complexity of safety analysis in ν-Petri nets [23].

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 [12]. We are at the forefront on the complexity analysis for such systems [15], [13], [16], [14].

We investigate in the ANR PRODAQ project a different angle on the static analysis of queries, relying on proof systems. Our first results on the subject [18] provide a sequent calculus for a modal data logic with an optimal proof-search algorithm.

6.2. Personal information management.

Thymeflow We developed Thymeflow, a personal knowledge base with spatio-temporal data [24].

The typical Internet user has data spread over several devices and across several online systems. We demonstrate an open-source system for integrating user's data from different sources into a single Knowledge Base. Our system integrates data of different kinds into a coherent whole, starting with email messages, calendar, contacts, and location history. It is able to detect event periods in the user's location data and align them with calendar events. We will demonstrate how to query the system within and across different dimensions, and perform analytics over emails, events, and locations.

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. International Research Visitors

8.2.1. Visits of International Scientists

Victor Vianu, June 15 to September 15, UC San diego

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. General chair, scientific chair

Serge Abiteboul co-organized the workshop "Data, responsibly" at Dagstuhl on ethics in massive data analysis.

9.1.2. Scientific events selection

9.1.2.1. Member of the conference program committees

Serge Abiteboul: Data engineering (ICDE'2016); In 2016, Serge Abiteboul also participated in Awards committees for the French Academy of Sciences and Inria.

Victor Vianu: FOIKS'16.

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: Associate Editor of ACM TOCL, Editor of the Database Theory column in SIGACT News.

9.1.4. Invited talks

Victor Vianu gave invited talks at the Dagstuhl Workshop on Foundations of Data Management, April 2016; POLARIS Colloquium, Univ. of Lille, June 2016; and Ecole Polytechnique, Saclay, July 2016.

Serge Abiteboul gave invited talks in international meetings at Int'l Symp. on DIStributed Computing in Paris, Didapro 6 in Namur, Dagstuhl Workshop on Foundations of Data Management, Dagstuhl workshop on Data, Responsibly, and Paris Open Source Summit.

9.1.5. Leadership within the scientific community

Serge Abiteboul is 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).

9.1.7. Research administration

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

Serge Abiteboul chaired the selection commitee for the call for proposal Dune "Développement d'Universités Numériques Expérimentales"/

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence : Sylvain Schmitz, Formal Languages, 22.5h, L3, ENS Cachan, France
Master : Sylvain Schmitz, Tree Automata Techniques and Applications, 22.5h, M1, ENS Cachan, France
Master : Sylvain Schmitz, Formal Languages, 30h, M2, ENS Cachan, France
Master : Sylvain Schmitz, Logic, 26.3h, M2, ENS Cachan, France
Master : Sylvain Schmitz, Logical and Computational Structures for Linguistic Modeling, 18h, M2, ENS Cachan, France
Master2 : Serge Abiteboul, Web data management, 15h, MPRI
Master1 : Serge Abiteboul, Initiation to scientific research (lab), 15h
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 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
PhD in progress: Anthony Lick, Proof systems for data queries, 01/09/2016, David Baelde and Sylvain Schmitz
PhD in progress: Su Yang, Analytics in Personal information management systems, 01/09/2016, Serge Abiteboul et Pierre Senellart

9.2.3. Juries

Luc Segoufin was on the Ph.D committee of Florent Capelli, Université de Paris 7, in June 2016.

Sylvain Schmitz was on the Ph.D. committees of Silva Stella, Università degli studi di Torino, in January 2016, and of Michael Blondin, Université de Montréal, in June 2016.

Serge Abiteboul was on the Ph.D. committee of Damian Bursztyn, University of Paris-Saclay, in December 2016.

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. He organized a Workshop Initiation to Wikipedia in Poitiers

Serge Abiteboul is Scientific curator of the exhibit "Terra Data" at Cité des Sciences de La Villette, 2017. He is member of the editorial board of Les big data à découvert, directed Mokrane Bouzeghoub and Rémy Mosseri. He gave invited talks at La Pépinière 4.1 in Nancy, Rencontres du Numérique of ANR, Parole Publique, Sciences à Coeur at Sorbonne Univ., Espace Mendes France at Poitiers,

Serge Abiteboul is Chairman of the Executive Council of the Foundation Blaise Pascal for the promotion of Maths and Computer Science.

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

Articles in International Peer-Reviewed Journals

[11] P. Beame, N. Grosshans, P. McKenzie, L. Segoufin. *Nondeterminism and An Abstract Formulation of Nečiporuk's Lower Bound Method*, in "ACM Transactions on Computation Theory", December 2016, vol. 9, n^{o} 1, pp. 1 - 34 [*DOI :* 10.1145/3013516], https://hal.inria.fr/hal-01426213

[12] F. Jacquemard, L. Segoufin, J. Dimino. *FO2(<,+1,) on data trees, data tree automata and branching vector addition systems*, in "Logical Methods in Computer Science", 2016, vol. 12, n^{o} 2, 32 p. , https://hal.inria.fr/hal-00769249

[13] S. Schmitz. *Complexity Hierarchies Beyond Elementary*, in "ACM Transactions on Computation Theory", February 2016, vol. 8, n^{o} 1 [*DOI :* 10.1145/2858784], https://hal.inria.fr/hal-01267354

[14] S. Schmitz. *Implicational Relevance Logic is 2-ExpTime-Complete*, in "The Journal of Symbolic Logic", June 2016, vol. 81, n^{o} 2, pp. 641–661 [*DOI :* 10.1017/jsl.2015.7], https://hal.inria.fr/hal-01340113

Articles in Non Peer-Reviewed Journals

[15] S. Schmitz. *Automata Column: The Complexity of Reachability in Vector Addition Systems*, in "ACM SIGLOG News", January 2016, vol. 3, n^{o} 1, pp. 3–21 [*DOI :* 10.1145/2893582.2893585], https://hal.inria.fr/hal-01275972

Invited Conferences

[16] J. Leroux, S. Schmitz. *Ideal Decompositions for Vector Addition Systems*, in "STACS 2016 - 33rd Symposium on Theoretical Aspects of Computer Science", Orléans, France, N. Ollinger, H. Vollmer (editors), Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2016, vol. 47, pp. 1–13 [*DOI :* 10.4230/LIPIcs.STACS.2016.1], https://hal.inria.fr/hal-01275991

International Conferences with Proceedings

[17] S. Abiteboul, P. Bourhis, V. Vianu. *A formal study of collaborative access control in distributed datalog*, in "ICDT 2016 - 19th International Conference on Database Theory", Bordeaux, France, W. Martens, T. Zeume (editors), March 2016, https://hal.inria.fr/hal-01290497

[18] D. Baelde, S. Lunel, S. Schmitz. *A Sequent Calculus for a Modal Logic on Finite Data Trees*, in "CSL 2016", Marseille, France, J.-M. Talbot, L. Regnier (editors), Leibniz International Proceedings in Informatics, LZI, September 2016, vol. 62, n^{o} 32, pp. 1–16 [*DOI :* 10.4230/LIPIcs.CSL.2016.32], https://hal.inria.fr/hal-01191172

[19] A. DEUTSCH, Y. LI, V. VIANU. *Verification of Hierarchical Artifact Systems*, in "35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS 2016)", San Francisco, United States, ACM (editor), Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS 2016), June 2016, pp. 179 - 194 [*DOI :* 10.1145/2902251.2902275], https://hal.inria.fr/hal-01389845

[20] J. GOUBAULT-LARRECQ, S. SCHMITZ. *Deciding Piecewise Testable Separability for Regular Tree Languages*, in "ICALP 2016", Rome, Italy, I. CHATZIGIANNAKIS, M. MITZENMACHER, Y. RABANI, D. SANGIORGI (editors), Leibniz International Proceedings in Informatics, July 2016, vol. 55, pp. 97:1–97:15 [*DOI :* 10.4230/LIPIcs.ICALP.2016.97], https://hal.inria.fr/hal-01276119

[21] G. GREFENSTETTE, K. RAFES. *Transforming Wikipedia into an Ontology-based Information Retrieval Search Engine for Local Experts using a Third-Party Taxonomy*, in "Joint Second Workshop on Language and Ontology & Terminology and Knowledge Structures (LangOnto2 + TermiKS) LO2TKS", Portoroz, Slovenia, May 2016, https://hal.inria.fr/hal-01224114

[22] P. HOFMAN, S. LASOTA, R. LAZIĆ, J. LEROUX, S. SCHMITZ, P. TOTZKE. *Coverability Trees for Petri Nets with Unordered Data*, in "FoSSaCS", Eindhoven, Netherlands, Lecture Notes in Computer Science, Springer, 2016, vol. 9634, pp. 445–461 [*DOI :* 10.1007/978-3-662-49630-5_26], https://hal.inria.fr/hal-01252674

[23] R. LAZIĆ, S. SCHMITZ. *The Complexity of Coverability in ν-Petri Nets*, in "LICS 2016", New York, United States, ACM Press, 2016, pp. 467–476 [*DOI :* 10.1145/2933575.2933593], https://hal.inria.fr/hal-01265302

[24] D. MONTOYA, T. PELLISSIER TANON, S. ABITEBOUL, F. M. SUCHANEK. *Thymeflow, A Personal Knowledge Base with Spatio-Temporal Data*, in "25th ACM International Conference on Information and Knowledge Management", Indianapolis, IN, United States, October 2016 [*DOI :* 10.1145/2983323.2983337], https://hal.inria.fr/hal-01355150

Conferences without Proceedings

[25] S. ABITEBOUL, G. MIKLAU, J. STOYANOVICH, G. WEIKUM. *Data, Responsibly (Dagstuhl Seminar 16291)*, in "Dagstuhl seminar", Dagstuhl, Germany, 2016, https://hal.inria.fr/hal-01405693

Scientific Popularization

[26] S. ABITEBOUL. *Analyse des données et choix de société* , in "parole publique", March 2016, https://hal.inria.fr/hal-01273439

Other Publications

[27] J. STOYANOVICH, S. ABITEBOUL, G. MIKLAU. *Data, Responsibly: Fairness, Neutrality and Transparency in Data Analysis*, March 2016, International Conference on Extending Database Technology, https://hal.inria.fr/hal-01290695