Activity Report 2011

Project-Team DAHU

Database and verification

IN COLLABORATION WITH: Laboratoire specification et vérification (LSV)
# Table of contents

1. Members .......................................................... 1
2. Overall Objectives .................................................. 1
3. Scientific Foundations ............................................. 2
4. Application Domains ............................................... 2
5. New Results .......................................................... 2
   5.1. Specification and verification of database driven systems 2
   5.2. Distributed data management 3
   5.3. Tree automata theory 3
6. Partnerships and Cooperations ..................................... 4
   6.1. National Initiatives 4
   6.2. European Initiatives 5
      6.2.1. FP7 Projects 5
         6.2.1.1. FOX 5
         6.2.1.2. Webdam 5
      6.2.2. Major European Organizations with which Dahu has followed Collaborations 5
   6.3. International Initiatives 5
      6.3.1. INRIA International Partners 5
      6.3.2. Visits of International Scientists 6
         6.3.2.1. Senior scientists on sabbatical 6
         6.3.2.2. Internship 6
      6.3.3. Participation In International Programs 6
7. Dissemination ...................................................... 6
   7.1. Animation of the scientific community 6
   7.2. Teaching 7
8. Bibliography .......................................................... 7
Project-Team DAHU

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

Dahu is a common project with LSV and ENS de Cachan. The team was created on January the 1st, 2008.

1. Members

Research Scientists
- Serge Abiteboul [Research Director (DR), HdR]
- Stéphane Demri [Research Director (DR) CNRS, HdR]
- Florent Jacquemard [Research assistant (CR), Inria, HdR]
- Luc Segoufin [Team leader, Research Director (DR), Inria, HdR]

Faculty Member
- Cristina Sirangelo [MCF ENS Cachan]

PhD Students
- Emilien Antoine [Webdam]
- Pierre Bourhis [Allocation couplée, until end of February]
- Nadine Francis [ENS-Cachan, since September]
- Alban Galland [Webdam, until end of June]
- Wojciech Kazana [Webdam]

Post-Doctoral Fellows
- Daniel Deutch [Webdam, until mid of October]
- Yannis Katsis [Webdam, until end of August]
- Bruno Marnette [Webdam, until end of June]
- Szymon Toruńczyk [Fox]

Visiting Scientist
- Victor Vianu [On sabbatical leave from UC San Diego, until mid September]

Administrative Assistant
- Isabelle Biercewicz [Secretary (SAR) 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. Scientific Foundations

3.1. Scientific Foundations

Dahu has strong connections with the Leo project-team in Saclay, the Mostrare project-team in Lille and the Cassis project-team in Nancy.

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, Pierre Bourhis, Luc Segoufin, Szymon Toruńczyk, Victor Vianu.

Modelization and verification of data centric systems. We have intensively studied the Active XML model. It is a high-level specification language tailored to data-intensive, distributed, dynamic Web services. Active XML is based on XML documents with embedded function calls. The state of a document evolves depending on the result of internal function calls (local computations) or external ones (interactions with users or other services). Function calls return documents that may be active, so may activate new subtasks. Our first line of result is a comparison of the specification power of various workflow control mechanisms within the Active XML framework and beyond [23].
AXML is very powerful and many static analysis problems are undecidable. We have also introduced a model of automata designed for modeling infinite runs of systems equipped with static relational databases. The automata model is equipped with finitely many variables, each of which can store values from a linearly ordered domain, such as the natural numbers. The transitions of the automata depend on a conjunctive query involving the database and the current values of the variables. For verifying infinite runs of such automata, an extension of temporal logic is considered, capable of comparing values stored in the variables and the database, at different times of the run. The main contribution of the work is the proof that automated verification of such temporal properties of the system can be carried out in PSpace. For more details, see [35].

Static analysis of query languages. XPath is arguably the most widely used XML query language as it is implemented in XSLT and XQuery and it is used as a constituent part of several specification and update languages. Hence in order to perform static analysis on a system manipulating XML data it is important to master the static analysis for XPath. Most of the important static analysis problems reduce to satisfiability checking: does a given query return a non-empty answer on some data. In general, in the presence of data values, the satisfiability of XPath is undecidable. We have shown that when restricted to its vertical navigational power, XPath becomes decidable [30].

5.2. Distributed data management

Participants: Serge Abiteboul, Emilien Antoine, Daniel Deutch, Alban Galland, Wojciech Kazana, Yannis Katsis, Luc Segoufin, Cristina Sirangelo.

Distributed knowledge base. As a foundation for managing distribution, we have proposed a model of a distributed knowledge base, that handles data and meta-data, as well as access control and localization, in a unique integrated setting. To support automatic reasoning on this knowledge base, we also introduced a novel rule-based language supporting the exchange of rules, namely Webdamlog. This work has been presented [21] and demonstrated [26] at major database conferences.

Probabilistic XML. Data from the Web are imprecise and uncertain. To manage this imprecision in a well-principled way, we have made significant advances in the field of probabilistic databases, and specifically, probabilistic XML. We have introduced new tractable probabilistic models for representing uncertain hierarchical information, and carried out in-depth studies of query evaluation, aggregation, and updates in various probabilistic XML models. These results have matured and some of the results are available in journal articles, e.g., [14].

Enumeration of query answers. 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 degree [19].

Data exchange and Web incomplete information. We have addressed the problem of restructuring data exchanged between communicating applications on the Web. We have proposed and analyzed a new language to specify data restructuring rules (schema mappings). This language generalizes existing mapping dependencies, by allowing a more flexible specification mechanism [20].

Jorge. We also invested a lot of effort in a textbook (undergraduate and graduate level) on Web data management (nicknamed Jorge) to be published at Cambridge University Press [38]. The book is already available on the Webdam Web site http://webdam.inria.fr/Jorge

5.3. Tree automata theory

Participants: Stéphane Demri, Florent Jacquemard, Luc Segoufin.
Most of our results for this section concerns data words and data trees. Those are words and trees where each position contains a data value together with the classical label. Data trees can be seen as a model for XML data. We have studied automata model using registers or memory or allowing constraints that can involve data comparisons in its transitions.

Register Automata. These extend the classical model of finite automata with auxiliary registers storing data values for later comparison.

We have introduced a new model of automata over data trees and shown the decidability of its emptiness problem [30]. These automata were used for obtaining decidability results for the static analysis for some fragments of XPath presented in the previous section.

Automata with counters. In [39], a survey chapter on the verification of infinite-state systems is presented that is focused on the verification of counter systems. Verification problems for vector addition systems and recursive Petri nets are considered. Moreover, we introduce subclasses of counter systems for which reachability questions can be solved in Presburger arithmetic viewed as a means to symbolically represent sets of tuples of natural numbers.

Automata with isomorphism tests among subtrees. We have also considered some models described by tree automata enriched with a feature testing for isomorphisms between subtrees. Such constraints could be used for testing monadic key constraints over XML documents. For these models, the main challenge is to establish the decidability of the non-emptiness of the language specified by a given automaton [18].

Rewriting Controlled by Selection Automata. Motivated by the problem of static analysis of XML update programs, we have studied [33] the combination, called controlled term rewriting systems (CTRS), of term rewriting rules with constraints selecting the possible rewrite positions. These constraints are specified, for each rewrite rule, by a selection automaton which defines a set of positions in a term based on tree automata computations. We have established several decidability and complexity results for several cases of the reachability and regular model checking problems for this tree transformation formalism.

6. Partnerships and Cooperations

6.1. National Initiatives

Dahu is currently participating in one ANR project:

**ENUM** is a research project supported by the ANR blanche (2007-2011) on algorithmic and complexity problems raised by enumerating solutions of a query. The goal is to provide formal methods to understand and compare the complexity of enumerations problems. The partners are University of Paris-7 (with Arnaud Durand), the project-team Mostrare at INRIA-Lille (with Joachim Niehren), the university of Caen (with Etienne Grandjean) and the university of Marseille (with Nadia Creignou). Dahu is involved in the ANR as part of the Paris-7 node. For more information please visit the web pages of ENUM: https://gforge.inria.fr/plugins/wiki/index.php?EnumerationProject&id=267&type=g.

Dahu is also the coordinator of one ARC INRIA

**ACCESS** is an ARC INRIA on Access Control for Web data, a two years project starting in 2010. The goal of this project is to study security and access control techniques for Web data exchange, and in particular the problems of the verification of access control policies (ACP) for XML and of the enforcement of ACP for secure query evaluation and update propagation. As a case study, the results are applied to an XML-based collaborative editing system. The partners are the teams CASSIS and PAREO at the INRIA center of Nancy-Grand-Est, and the team MOSTRARE at the the INRIA center of Lille-Nord-Europe. For more information please visit the web pages: http://acxml.gforge.inria.fr.
6.2. European Initiatives

6.2.1. FP7 Projects

6.2.1.1. FOX

Title: FOX
Type: COOPERATION (ICT)
Defi: FET Open
Instrument: Specific Targeted Research Project (STREP)
Duration: May 2009 - April 2012
Coordinator: Luc Segoufin, INRIA (France)
Others partners: Thomas Schwentick at the university of Dortmund, Mikołaj Bojańczyk at the university of Warsaw, Leonid Libkin at the university of Edinburgh, Georg Gottlob at the university of Oxford, Frank Neven at the university of Hasselt and Maarten Marx at the university of Amsterdam.
See also: http://fox7.eu
Abstract: The objective of FoX is to study the fundamental issues necessary in order to make the data management over the internet more efficient and more reliable.

6.2.1.2. Webdam

Title: WebDam
Type: IDEAS
Instrument: ERC Advanced Grant (Advanced)
Duration: December 2008 - November 2013
Coordinator: Serge Abiteboul, INRIA (France)
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.

6.2.2. Major European Organizations with which Dahu has followed Collaborations

Partner 1: University of Franckfurt, Nicole Schweikardt (Germany)
Subject 1: Logic and complexity.

6.3. International Initiatives

6.3.1. INRIA International Partners

Victor Vianu, UC San Diego, USA.
6.3.2. Visits of International Scientists

6.3.2.1. Senior scientists on sabbatical

Victor Vianu (from Jan 2011 until september 2011)
Subject: AXML
Institution: UC San Diego (USA)

Howard Straubing (from Mai 15 until June 15)
Subject: Tree automata
Institution: Boston College (USA)

6.3.2.2. Internship

Facundo CARREIRO (from Mar 2011 until Jul 2011)
Subject: Arithmetical and Language-based Constraints on Finite Ordered Trees
Institution: Universidad de Buenos Aires (Argentina)

6.3.3. Participation In International Programs

Dahu is coordinator (on the French side) of the project INRIA-DGRSRT (Tunisian universities) 10/I01 on the verification of security properties of Web services, access control policies and firewalls for XML. This project started in 2010, the other partners are the CASSIS team at INRIA Nancy-Grand-Est and the Security team at Sup’Com Tunis.

7. Dissemination

7.1. Animation of the scientific community

Organization of workshops and conferences.

– Serge Abiteboul was General PC chair of the Intl. Conf. on Data Engineering (ICDE) in 2011.
– Luc Segoufin co-organized with Thomas Schwentick (U. Dortmund), Serge Abiteboul and Alin Deutch (UCSD) a Dagstuhl Workshop on Foundation of XML data.

Program Committees.

– Florent Jacquemard: Rewriting Techniques and Application (RTA’11).
– Luc Segoufin: Logic in Computer Science (LICS’11).
– Cristina Sirangelo: Intl. Conf. on Database Theory (ICDT’11), ACM Principles of Database Systems (PODS’11).

Editorial boards. Stéphane Demri is member of the publication board of the review “Technique et Science Informatiques” (among 5 members).

Juries of Ph.D theses.

– Luc Segoufin was reviewer for the Ph.Ds of Tobias Ganzo (U. Aachen, 2011).
– Stéphane Demri was reviewer for the PhD of Rafiq Saleh (U. Liverpool, Sep. 2011) and is examinator for the PhD of François Bobot (LRI, Dec. 2011).
Juries of HDR theses.
- Luc Segoufin was member of the HDR jury of Florent Jacquemard (ENSC, 2011).

Responsibilities.
- Stéphane Demri is a member of the DIM LSC committee (Région IdF) in 2011.
- Luc Segoufin is since 2009 the coordinator of the European project FoX. Since 2010 he is a member of the steering committee of the Intl. Conf. on Database Theory (ICDT). Since 2010 he is part of the “bureau du comité 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/).

Larger audience. Serge Abiteboul participated in a panel at Futurs en Seine, and in debates in Sèvres and Geneva. He has been interviewed by the magazine Enjeux, Les Echos.

Serge Abiteboul presented a keynote talk at the International Workshop for the 30th anniversary of ACM PODS Conference on Principles of Database Systems.

Nomination. Serge Abiteboul has been elected to the Academia Europea. He has also been elected ACM Fellow.

7.2. Teaching

As a Maître de conférence Cristina Sirangelo is teaching in the department of computer science of ENS de Cachan.

As a PhD. student with half of a “mission d’enseignement” Émilien Antoine is teaching in the department of computer science of Université de Paris-Sud 11.

Licence : “Bases de données”, Serge Abiteboul, 20h, L3, ENs Cachan
Master : “Web Data Management”, Serge Abiteboul, 12h, M2, MPRI.
Doctorat : “Tree automata techniques for the verification of infinite state-systems”, Florent Jacquemard, 7h, Liège, Belgique.
Doctorat : “The WWW”, Serge Abiteboul, 3h, the thematic school of the University ES-Sénia, Oran, Algeria.
Doctorat : “Web data management”, Serge Abiteboul, 3h, the Journées Nationales APMEP.

PhD & HdR

HdR : Florent Jacquemard, Extended Tree Automata for the Verification of Infinite Systems, ENS Cachan, November 10 2011
PhD : Pierre Bourhis, On the dynamics of active documents for distributed data management, Université Paris-Sud 11, February 11th 2011, Serge Abiteboul
PhD : Alban Galland, Distributed Data Management with Access Control, Université Paris-Sud 11, September 28th 2011, Serge Abiteboul

8. Bibliography

Major publications by the team in recent years


Publications of the year

Doctoral Dissertations and Habilitation Theses


Articles in International Peer-Reviewed Journal


International Conferences with Proceedings


National Conferences with Proceeding


Scientific Books (or Scientific Book chapters)
