ERC Webdam

Activity Report 2010

1 Executive summary

Two years after its inception, we are happy to report that the Webdam project is solidly on track.

After a planning period where we identified research issues, we continued investigating the main research themes of Webdam. This year we obtained major results in the following themes:

- Querying probabilistic XML. See Section 4.1.
- Data-centric workflows. See Section 4.2.
- Analysis and optimization of distributed data management. See Section 4.3
- Models for Web data management. See Section 4.4.

After difficulties encountered to attract talents the first year, hiring was most fruitful the second year. We could notably bring Prof. Vianu from UCSD for a year (he was already collaborating with Webdam). We also attracted two very talented post docs, Bruno Marnette (PhD Oxford) and Fabian Suchanek (PhD Max Planck) to join Yannis Katsis (PhD San Diego). Also, notably Meghyn Bienvenu, was hired at CNRS at Orsay and joined Webdam.

The plan is to continue with the above research topics. With the arrival of Fabian and Meghyn, we now have the human resources to investigate new directions that we view as essential, around distributed knowledge and ontologies (more semantic Web in essence). Our goal is to continue investigating the directions already established and build in these new directions.

One should note that in 2010, we put a lot of efforts in writing a text book (undergraduate and graduate level) on Web data management to be published at Cambridge University Press [6].

2 The general goal

The Webdam ERC grant (S. Abiteboul) started in December 2008. The goal is to develop a formal model for Web data management. This goal is to

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.

The Webdam project is shared between the Dahu and Leo project-teams, both from INRIA Saclay.

3 Participants

Members

- Serge Abiteboul, manager of the Webdam ERC project [01/12/2008] Dahu Leo
- David Gross-Amblart, assistant professor at University of Bourgogne, seconding [01/09/2010] Leo
- Yannis Katsis, post doc [01/09/2009] Dahu
- Bruno Marnette, post doc [01/09/2010] Dahu
- Philippe Rigaux, professor at Dauphine University, seconding [01/09/2009-01/10/2010] Leo
- Marie-Christine Rousset, professor at University of Grenoble, seconding [01/09/2009] Leo
- Fabian Suchanek, post doc [01/06/2010] Leo
- Victor Vianu, professor at UC San Diego [01/07/2010] Dahu

PhD Students

- Emilien Antoine, Paris Sud University [01/10/2010] Leo
- Pierre Bourhis, Paris Sud University [01/12/2008] Dahu
- Alban Galland, Paris Sud University [01/12/2008] Dahu Leo
- Wojciech Kazana, ENS Cachan [01/02/2010] Dahu
- Evgeny Kharlamov, Free University of Bozen-Bolzano [01/01/2009] Telecom

- Marilena Oita, Telecom ParisTech [01/11/2009] Telecom

Others

- Alin Gabriel Tilea, engineer [01/09/2009-01/06/2010] Leo
- Kristian Lyndbaek, intern [07/10] Dahu

Collaborators

- Meghyn Bienvenu, researcher at Orsay University [01/04/2010] Leo
- Ioana Manolescu, manager of the Leo team, INRIA Saclay [01/12/2008]
- Luc Segoufin, manager of Dahu team, INRIA Saclay [01/12/2008]
- Pierre Senellart, associate professor at Telecom ParisTech [01/12/2008]
- Philippe Rigaux, professor at Paris CNAM [01/10/2010]
- Victor Vianu, professor at UC San Diego [01/12/2008]

Webdam Alumni

- Balder ten Cate went to University of California, Santa Cruz [01/04/2009-01/10/2009]
- Bogdan Marinoiu finished his PhD thesis and went to SAP-Business Object [01/01/2009-01/09/2009]

Visitors

- Balder ten Cate (UCSB) (07/10)
- Yannis Papakonstantinou (UCSD) (07/10)
- Gerome Miklau (U. Washington) (06/2010)
- Tova Milo (Tel Aviv U.) (11/2010)
- Moshe Vardi (Rice U.) (10/2010)
- Hyunjung Park (Stanford U.), winner of the SIGMOD 2010 programming contest (12/10)

4 Research

We next give a more complete list of results that were obtained in 2010.

4.1 Probablistic XML

A major axis of research inside the Webdam project deals with the management of uncertain data in general, and of probabilistic trees (i.e., probabilistic XML) in particular. We have made several advances on this topic.

Aggregate Queries and Continuous Distributions In [4], we provide a systematic study of the complexity of aggregate queries defined with tree-pattern queries with joins (and restrictions of that query language). We show that when only local probabilistic dependencies are considered, some particular aggregate functions (namely, *monoid* ones) are tractable. We also provide a formal model for continuous distributions in probabilistic trees and show that they essentially do not add any complexity to querying tasks, as long as a number of operations (convex sums, differentiation, integration, convolution) can be tractably performed over these distributions, either symbolically or numerically.

Updating Probabilistic XML Initial studies of updates in probabilistic XML suggested that the more complex the probabilistic dependencies in data are, the simpler it is to represent the result of a probabilistic update in that model. We investigated the problem further in [16] and showed the situation is more contrasted: it is actually possible to represent tractably the result of insertions defined by simple queries in a local dependency model, while this is not possible in a model with dependencies expressed by arbitrary conjunctions. Deletions almost always lead to a combinatorial explosion, except when arbitrary propositional formulas are allowed as probabilistic conditions on nodes of a tree.

A more general probabilistic XML model All XML uncertain data models of the literature had the limitation of representing trees of bounded size. In [9], a joint work with the FoX FP7 project, we overcome this introduction with the help of recursive Markov chains, a formalism for probabilistic processes with recursive calls. We explain how to use them to generate probabilistic trees. We show some natural restrictions of RMCs (that are strictly more expressive than existing local probabilistic XML models) yield tractable query answering, in some cases assuming unit-cost arithmetics.

4.2 Data-Centric Workflow Specification Languages

Dahu

There has recently been a proliferation of workflow specification formalisms, notably data-centric, in response to the need to support increasingly ubiquitous processes centered around databases. Prominent examples include e-commerce systems, enterprise business processes, health-care and

scientific workflows. Developing suitable specification mechanisms for such workflows, understanding the expressive power of the different specification formalisms, and performing static analysis are critical issues that we address in some of our recent work [3, 12]. A prototype, named AXART, based on a model presented in [3] has been demonstrated at the VLDB 2010 conference [2].

A powerful framework for specifying and studying data-centric workflows is provided by Active XML, a high-level specification language tailored to data-intensive, distributed, dynamic Web-services. In brief, Active XML consists of 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). Functions can be naturally used to model tasks in a workflow. They return documents that may be active, so may in turn activate new sub-tasks, thus having the ability to naturally specify a hierarchy of tasks. We have previously studied the verification of temporal properties of runs of Active XML systems, specified in a tree-pattern based temporal logic, Tree-LTL, that allows expressing a rich class of semantic properties of the workflow. The main results establish the boundary of decidability and the complexity of automatic verification of Tree-LTL properties. In our most recent work [3], we focus on *comparing* the specification power of various workflow control mechanisms within the Active XML framework and beyond. This is intrinsically difficult because of the lack of a standard yardstick for expressiveness. In [3], we develop a flexible framework for comparing workflow specification languages, in which the pertinent aspects to be taken into account are defined by *views*. We use it to compare the expressiveness of several workflow specification mechanisms based on automata, pre/post conditions, and temporal constraints.

The AXART system extends the model studied in [3] with the ability to have an explicit hierarchy of distributed tasks with associated control. A demonstration of the system was presented in [2], based on an example taken from the movie industry, that specifies the workflow involved in applying for a role in a film.

Another prominent formalism for specifying data-centric workflows is IBM's Business Artifacts. In collaboration with UC San Diego and IBM, we studied in previous work the verification problem for a restricted class of such workflows. However, the early results suffer from an important limitation: they fail in the presence of even very simple data dependencies or arithmetic, both crucial to real-life business processes. In [12], we extend the artifact model and verification results to alleviate this limitation. We identify a practically significant class of business artifacts with data dependencies and arithmetic, for which verification is decidable.

4.3 Analysis of Distributed Data Management

With the evolution of the Web and the emergence of universal standards for data exchange, data management is becoming increasingly distributed. As a first step in setting theoretical foundations for distributed data management systems, we study the equivalence of such systems [7]. To model these systems, we use the Active XML framework. The resulting model is expressive enough to capture distributed systems that are recursive, utilize asynchronous communication and operate on data streams.

As our model is quite general, the equivalence problem is undecidable. We exhibit restrictions of the model, in terms of query languages used in the function calls and the presence or absence of external function calls, for which equivalence can be effectively decided. We study the computational complexity of the equivalence problem, and for a limited class of distributed systems present the axiomatization of equivalence.

4.4 Social networks

Leo

Use of the web to share personal data is increasing rapidly with the emergence of Web 2.0 and social networks applications. However, users have yet to trust all the different hosts of their data and face difficulty with updates. To overcome these problems, we are studying a model of distributed knowledge base with access control and cryptographic functionality. The model allows exchanging documents, access control statements, keys and instructions in a distributed setting. We are considering different implementations of this model that can be used to leverage technologies such as DHT or Gossiping. Such implementation will be demonstrated in the ICDE 2011 conference [8]. The model allows us to reason on the knowledge base, in particular on the access control and the distribution policies. The reasoning is based on trusted knowledge, the statements, whom authenticity can be easily checked. In particular we describe authentication based on signature or on url and a wild range of data distribution. Some preliminary result on the reasoning part have been presented during the Datalog2.0 workshop [1]. The precise description of the model is yet a very active and relatively mature topic of the team.

In such a social network, participants may bring conflicting opinions. We have studied the problem of trying to corroborate information coming from a very large number of participants. We have proposed and evaluated various algorithms towards this goal [13].

4.5 Ontology management

Leo

The data on the Web is not limited to HTML pages. Nowadays, an increasing part of the Web is taken by the *Semantic Web*. The Semantic Web can be seen as a gigantic, distributed and interlinked entity-relationship graph. It contains machine-interpretable information from encyclopedias, gazetteers, government censuses, music databases, libraries and numerous other sources. Each such sub-graph is commonly known as an *ontology*.

Our work in this area is threefold. First, we work on contributing data to the Semantic Web. Together with researchers from the Max Planck Institute in Germany, we are developing techniques to enrich the YAGO ontology[1] by a temporal and spatial dimension. Second, we investigate the interaction of ontologies and Web services. Together with researchers from Stefano Ceri's group in Milan, we are working on bridging the gap between the Semantic Web and the Web service technology developed in the Search Computing project. Third, we are working on algorithms that can detect and merge equivalent entities, classes and properties in different ontologies.

4.6 Web archiving

In the course of the Webdam project, and especially in the setting of the PhD thesis of Marilena Oita, we investigate how actual data from the Web can be archived in a timely manner, going further than what is currently done by archival crawlers. In particular, we researched how Web feeds can be used to archive Web pages that contain temporal data objects, such as blog posts or news items. We use RSS or Atom feeds to extract these Web objects and to detect change in the context of an incremental crawl. To detect change on crawled Web pages that have a Web feed associated, we designed an algorithm that extracts the information of interest (the data object), with the aim of analyzing changes effectively, without being tricked by possible changes in the surrounding boilerplate. See [20] for further details.

4.7 Query enumeration

Dahu

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*. For this, it might be interesting to compute adequate index structures. The ultimate goal being to obtain

[1] See article by F.M. Suchanek and G. Kasneci and G. Weikum in WWW07.

index structures easily computable (say in linear time in the size of the database), that allow constant delay in the enumeration process. We are interested in understanding under what assumptions (on both the query language and the kind of the database) such algorithms exist. This work is in particular in the setting of the PhD thesis of Wojciech Kazana.

5 Dissemination

5.1 Program committees

- Dahu Leo Serge Abiteboul is General Program Chair of the International Conference on Data Engineering 2011 in Hannover, Germany (one of the main conferences on database systems).
- Pierre Senellart served as the tutorial co-chair of ICDE 2010, the program co-chair of the industrial track of EDBT 2011, and in the program committees of CIKM 2010, BDA 2010, and ICDT 2011.
- Dahu Victor Vianu served on the Program Committees of IEEE Symp. on Logic in Computer Science (LICS) 2010, Int'l. Conf. on Very Large Databases (VLDB) 2010, Int'l. Symp on Foundations of Information and Knowledge Systems (FoIKS) 2010, Alberto Mendelzon Workshop on Foundations of Data Management (AMW) 2010
- Leo Fabian Suchanek served on the Program Committee of the World Wide Web Conference (WWW) 2011.

5.2 Organization of workshops and conferences

Dahu Serge Abiteboul co-organized with Andreas Oberweis (KIT, Germany) and Jianwen Su (UCSB, USA) the Dagstuhl Workshop on Enabling Holistic Approaches to Business Process Lifecycle Management (04/2010) [5].

Dahu Leo Pierre Senellart organized (together with Serge Abiteboul) the ACM SIGMOD 2010 programming contest. Teams of contestants from degree-granting institutions had to develop an efficient distributed query engine on top of an in-memory index. The competition received much attention, with 29 teams from 23 different institutions over the world [14].

5.3 Education and editing

Two important editing activities continued:

1. Leo S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset and P. Senellart are writing a text book entitled *Web Data Management*. It covers the recent advances in the modeling, querying, integration and indexing of very large data sets. The focus is on Web scale data

management, and the text deals with some of the most important issues that arise in this context (e.g., heterogeneity, volume, distribution in large networks, etc.). The target audience are graduate and PhD students, engineers and practitioners seeking for an in-depth presentation of the languages, techniques and tools required to build large-scale and distributed information systems. We plan a first edition in spring 2010, along with a presentation of the book material (or part of) during a Summer School co-organized by WebDam that was held in Les Houches in May 2010. A preliminary version is available [6]. The book is to be published by Cambridge University Press.

2. Dahu The initiative by S. Abiteboul, R. Hull and V. Vianu of a second electronic volume of *Foundations of Databases*. Two new parts are being considered: semistructured data (L. Libkin) and data integration (A. Deutsch). Both are fully related to the activity of Webdam.

5.4 Invited Presentations and tutorials

- Serge Abiteboul gave a keynote presentation on Web information management and knowledge bases at the 10th International Conference on Web Engineering, Vienna (07/2010); an invited presentation on Web data management at the Datalog 2.0 Workshop, held in March 2010, at Oxford University [1]; an invited presentation on Object Databases at the Dagstuhl workshop on Relationships, Objects, Roles, and Queries in Modern Programming Languages (04/2010); an invited presentation on Workflow Specification Languages for Active Documents at the Fox Workshop in Amsterdam (05/2010); an invited presentation on WebdamExchange: A Model for Data Access on the Web. Workshop on Formal Methods for Web Data Trust and Security; invited presentations at Centre d'Alembert and French Academy of Sciences.

- Victor Vianu was invited to give a presentation at the Workshop on Automata and Logic for Data Manipulating Programs, Paris, December 2010.

- Fabian Suchanek was an invited expert at the Search Computing Workshop in Milan.

- Fabian Suchanek was invited as a keynote speaker for the First International Conference on Integrated Computing Technology 2011.

- Fabian Suchanek contributed to a tutorial on "Harvesting the Web of Data" at the Conference on Information and Knowledge Management (CIKM) 2010.

5.5 Journal editing

- Victor Vianu: Editor-in-chief of JACM, Area editor for ACM Trans. on Computational Logic (logical aspects of databases), Editor of the Database Theory Column of SIGACT News
- Serge Abiteboul is a member of the steering committee of Proceedings of the VLDB Endowment (PVLDB) Journal, a journal that just started.
- Pierre Senellart is Information Director of the Journal of the ACM.

6 Awards

The following awards were received for works performed before joining Webdam:

- Victor Vianu received the 2010 Alberto O. Mendelzon Test of Time Award of the ACM Symp. on Principles of Database Systems (PODS) for the article *Typing for XML Transformers*, joint with Tova Milo and Dan Suciu.
- Meghin Bienvenu received the AFIA Prize for her thesis.
- Fabian Suchanek received the Otto-Hahn-Medal, the dissertation award of the Max Planck Society.
- Fabian Suchanek was also selected for a research group leader stipend by the Max Planck Society.
- Fabian Suchanek won the ACM Dissertation Award Honorable Mention for his dissertation.

References

[1] Serge Abiteboul, Meghyn Bienvenu, Alban Galland, and Marie-Christine Rousset. Distributed Datalog Revisited. In *Datalog 2.0 Workshop*, Oxford Royaume-Uni, 2011.

[2] Serge Abiteboul, Pierre Bourhis, Bogdan Marinoiu, and Alban Galland. AXART - Enabling Collaborative Work with AXML Artifacts. In *International Conference on Very Large Data Bases, demonstration*, Singapour, 2010.

[3] Serge Abiteboul, Pierre Bourhis, and Victor Vianu. Comparing Workflow Specification Languages: A Matter of Views. In *ICDT*, Uppsala Suède, 2011.

[4] Serge Abiteboul, T-H. Hubert Chan, Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Aggregate Queries for Discrete and Continuous Probabilistic XML. In *International Conference on Database Theory (ICDT)*, pages 50–61, Lausanne Suisse, 2010.

[5] Serge Abiteboul, Agnes Koschmider, Andreas Oberweis, and Jianwen Su. Proceedings of the Dagstuhl Seminar 10151 "Enabling Holistic Approaches to Business Process Lifecycle Management", 2010.

[6] Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, and Pierre Senellart. *Web Data Management*. Cambridge University Press, 2011 (to appear).

[7] Serge Abiteboul, Balder Ten Cate, and Yannis Katsis. On the Equivalence of Distributed Systems with Queries and Communication. In *ICDT*, Uppsala Suède, 2011.

[8] Emilien Antoine, Alban Galland, Kristian Lyngbaek, Amélie Marian, and Neoklis Polyzotis. Social Networking on top of the WebdamExchange System. In *International Conference on Data Engineering, demonstration*, Hannover Allemagne, 2011.

[9] Michael Benedikt, Dan Olteanu, Evgeny Kharlamov, and Pierre Senellart. Probabilistic XML via Markov Chains. In *International Conference on Very Large Data Bases*, pages 770–781, Singapour, 2010.

[10] Diego Calvanese, Evgeny Kharlamov, Werner Nutt, and Dmitriy Zheleznyakov. Evolution of DL-Lite Knowledge Bases. In *Proc. of ISWC The 9th International Semantic Web Conference, (ISWC)*, pages 112–128, Chine, 2010.

[11] Diego Calvanese, Evgeny Kharlamov, Werner Nutt, and Dmitriy Zheleznyakov. Updating ABoxes in DL-Lite. In *Proc. of AMW The 3dr Alberto Mendelzon Workshop on Foundations of Data Management (AMW)*, Argentine, 2010.

[12] Elio Damaggio, Alin Deutsch, and Victor Vianu. Artifact Systems with Data Dependencies and Arithmetic Constraints. In *Proc. International Conference on Database Theory*, 2011.

[13] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating Information from Disagreeing Views. In *International Conference on Web Search and Data Mining (WSDM)*, New York City États-Unis, 2010.

[14] Clément Genzmer, Volker Hudlet, Hyunjung Park, Daniel Schall, and Pierre Senellart. The SIGMOD 2010 Programming Contest: A Distributed Query Engine. *SIGMOD Record*, 2010.

[15] Georg Gottlob and Pierre Senellart. Schema Mapping Discovery from Data Instances. *Journal of the ACM*, 2010.

[16] Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Updating Probabilistic XML. In *Proc. of Extending Database Technology Workshop on Updates in XML*, Lausanne Suisse, 2010.

[17] Hady Lauw, Ralf Schenkel, Fabian Suchanek, Martin Theobald, and Gerhard Weikum. Harvesting Knowledge from Web Data and Text. In *CIKM*, Toronto Canada, 2010.

[18] Bruno Marnette and Floris Geerts. Static analysis of schema-mappings ensuring oblivious termination. In *International Conference on Database Theory*, 2010.

[19] Bruno Marnette, Giansalvatore Mecca, and Paolo Papotti. Scalable data exchange with functional dependencies. In *VLDB*, 2010.

[20] Marilena Oita and Pierre Senellart. Archiving Data Objects using Web Feeds. In *International Workshop on Web Archiving*, Vienna Autriche, 2010.

[21] Nicoleta Preda, Gjergji Kasneci, Fabian Suchanek, Thomas Neumann, and Wenjun Yuan. Active knowledge: Dynamically enriching rdf knowledge bases by web services (angie). In *International Conference on Management of Data (SIGMOD 2010)*. ACM, 2010.

[22] Remi Tournaire, Alexandre Termier, Jean-Marc Petit, and Marie-Chirstine Rousset. Combining logic and probabilities for discovering mappings between taxonomies. In *KSEM*, 2010.

[23] Remi Tournaire, Alexandre Termier, Jean-Marc Petit, and Marie-Chirstine Rousset. Probamap: a scalable tool for discovering probabilistic mappings between taxonomies. In *AKBC*, 2010.

[24] Dmitriy Zheleznyakov, Diego Calvanese, Evgeny Kharlamov, and Werner Nutt. Updating TBoxes in DL-Lite. In *Proc. of DL Description Logic Workshop (DL)*, Canada, 2010.