ERC Webdam

Activity Report 2009

February 17, 2010

1 Executive summary

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

The project started with a planning period whose goal was to identify the research issues upon which most effort will be focused. This was done before, during and after a Webdam workshop organized for this purpose. We have already made progress on some of the research themes we consider essential for Webdam, including:

- **The specification of task sequencing in data-driven workflows** The lack of explicit control for task sequencing has been a major criticism of data-centric models such as Active XML. We proposed the AXML artifact model to capture data and workflow on management of activities in distributed settings. We introduced and compared various mechanisms for specifying sequencing of tasks.

- **Querying probabilistic information**. This is of critical importance on the Web, where data is imprecise and uncertain. We proposed the first in-depth study of aggregate queries on probabilistic XML.

Several new topics were initiated this year:

- **Access control in a Web environment** This is again a critical issue in the context of the Web. For instance, lack of access control to information is seen as a main drawback of social networks.

- **Web archiving** The archiving of Web pages has already been studied and there are even archiving systems such as the WayBack machine of Internet Archive. We are working on less understood aspects such as the archiving of web services (also forms) and that of streams, e.g. RSS streams and tweets.

One difficulty encountered so far concerns recruiting. It has proved harder than expected to attract top talent to the project. We were successful in attracting several talented junior researchers (unfortunately Balder ten Cate had to leave a few months after joining for family reasons). One should note that, in terms of collaboration, there is also significant synergy with the EC Fox project (managed by Luc Segoufin).

The plan for 2010 is to continue with the above research topics. Dependent on human resources, we will also investigate new directions that we view as important, around the issues of trust and beliefs, updates and transactions. We are currently missing qualified researchers for the last topic. We thus see recruiting as an essential activity for next year.

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 Gemo project-teams, both from INRIA Saclay.

3 Participants

Members

- Serge Abiteboul, manager of the Webdam ERC project [01/12/2008]
- Yannis Katsis [01/09/2009]
- Philippe Rigaux, professor at Dauphine University, seconding at Webdam [01/09/2009]
- Marie-Christine Rousset, professor at University of Grenoble, seconding at Webdam [01/09/2009]

PhD Students

- Pierre Bourhis, Paris Sud University [01/12/2008]
- Alban Galland, Paris Sud University [01/12/2008]
- Wojciech Kazana, ENS Cachan [01/02/2010]

- Evgeny Kharlamov, Free University of Bozen-Bolzano [01/01/2009]
- Marilena Oita, Telecom ParisTech [01/10/2009]

Engineers

- Alin Gabriel Tilea [01/09/2009]

Collaborators

- Ioana Manolescu, manager of the Gemo 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]
- 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 (one month or more)

- Bruno Marnette, PhD student at Oxford (2 months)
- Amélie Marian, assistant professor at Rutgers University (2 months)
- Alkis Polizotis, professor at UC Santa Cruz (1 month)
- Sihem Amer-Yahia, Yahoo Research (1 month)

4 Research

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

4.1 Verification of AXML Systems

Active XML 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 sub-tasks. In [13, 12]. we 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 application. The main results establish the boundary of decidability and the complexity of automatic verification of Tree-LTL properties.

Towards a data-centric workflow approach, we introduced in [2] an *artifact model* to capture data and workflow management activities in distributed settings. We argue that the model, built on Active XML, captures the essential features of service calls and business artifacts as described informally by Nigam and Caswell in 2003. To illustrate, we considered the *monitoring* of distributed systems and the *verification* of temporal properties for them.

4.2 Probablistic XML

Various known models of probabilistic XML can be represented as instantiations of the abstract notion of p-documents. In addition to ordinary nodes, p-documents have distributional nodes that specify the possible worlds and their probabilistic distribution. Particular families of p-documents are determined by the types of distributional nodes that can be used as well as by the structural constraints on the placement of those nodes in a p-document. Some of the resulting families provide natural extensions and combinations of previously studied probabilistic XML models. In [9], the expressive power of families of p-documents has been investigated and a first study of the complexity of updates has been given. An ongoing work considers a more systematic approach to the problem of updating probabilistic XML. The evaluation of aggregate functions such as count, sum, avg, for probabilistic XML is the topic of [6]. A joint work with the FP7 FoX project, still in progress, deals with tractable extensions of the probabilistic XML models proposed so far, especially for trees of unbounded size.

4.3 Monitoring

We have worked on the conception and implementation of tools for monitoring Peer to Peer Systems. A system named P2PMonitor has been developed for this purpose. It is a P2P system itself, with peers exchanging messages by Web-service calls. We focused on a problem closely related to monitoring: view maintenance over active documents. Indeed, the monitoring problem can be seen as aggregating streams into an active document and incrementally evaluating a tree-pattern query over this active document. We have developed algorithmic datalog-based foundations for such an incremental query processing [3]. We have also studied theoretical issues raised in this context, such as query satisfiability over active documents and stream relevance for given queries [4].

The problem of determining the dynamic relevance of a service in the presence of binding patterns is the topic of an ongoing joint work with Michael Benedikt from University of Oxford.

4.4 Searching in P2P

We have pursued the work on a peer-to-peer platform for building and managing warehouses of Web resources. Our previous work addressed the issue of indexing extensional XML data (trees). We are currently working on indexing also graph data, more precisely, XML documents with references to other documents or including function calls.

4.5 Social networks

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 functionalities. 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.

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 [17].

4.6 Distributed XML Design

A distributed XML document is an XML document that spans several machines or Web repositories. We assume that a distribution design of the document tree is given, providing an XML tree some of whose leaves are "docking points", to which XML subtrees can be attached. These subtrees may be provided and controlled by peers at remote locations, or may correspond to the result of function calls, e.g., Web-services. If a global type τ, e.g. a DTD, is specified for a distributed document T, it would be most desirable to be able to break this type into a collection of local types, called a local typing, such that the document satisfies τ if and only if each peer (or function) satisfies its local type. In [7], we lay out the fundamentals of a theory of local typing and provide formal definitions of three main variants of locality: local typing, maximal local typing, and perfect typing, the latter being the most desirable.

5 Dissemination

5.1 PhD Thesis

Bogdan Marinoiu [22], Analysis and verification of distributed systems.

5.2 Participation in conferences

- Serge Abiteboul was General Program Chair of the Very Large Data Base Conference 2009 (the main conference on database systems).
- Ioana Manolescu: SIGMOD'09 (PC member), WWW'09 (Co-chair of the Web Engineering track)
- Pierre Senellart: VLDB'09 (Submission chair), WWW'09 (PC member), ICDE'10 (PC member)
- Victor Vianu: PODS'09 (PC member), VLDB'09 (PC member), FOIKS'10 (PC member)
- Ioana Manolescu and Pierre Senellart participated in Repeatability & Workability Evaluation track of SIGMOD 2009 [21].

5.3 Invited Presentations

- Serge Abiteboul: Time'09 [2]
- Ioana Manolescu: XML Symposium & Database and Programming Languages workshops (VLDB'09)
- Victor Vianu: ICDT'09 [27]

5.4 Webdam Workshops

We organized a Workshop on Brainstorming on Foundations of Web Data Management (August 28th, 2009, Post VLDB). Serge Abiteboul and Pierre Senellart were program chairs. Serge Abiteboul and Alban Galland were organizers.

Balder ten Cate organized in Cachan a Workshop on Modal Logic (July 21st 2009).

The Webdam kickoff meeting was organized in Cachan on January 7th, 2009.

Journal editing

- Victor Vianu: Editor-in-chief of JACM, Area editor for ACM Trans. on Computational Logic (logical aspects of databases)
- 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 prestigious Journal of the ACM.

6 Education and editing

Two important editing activities started:

1. S. Abiteboul, P. Rigaux, M.-C. Rousset and P. Senellart are writing a text book entitled *Web Data Management and Distribution*. 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 will be held in Chamonix in may 2010. A preliminary version is available [11].
2. 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.

Awards Serge Abiteboul has been elected member of the Academy of Sciences.

References

[1] Serge Abiteboul, Omar Benjelloun, and Tova Milo. Active XML. In *Encyclopedia of Database Systems*, pages 38–41. 2009.

[2] Serge Abiteboul, Pierre Bourhis, Alban Galland, and Bogdan Marinoiu. The AXML artifact model. In *Intern. Conf. on Temporal Representation and Reasoning (TIME)*, 2009.

[3] Serge Abiteboul, Pierre Bourhis, and Bogdan Marinoiu. Efficient maintenance techniques for views over active documents. In *EDBT*, pages 1076–1087, 2009.

[4] Serge Abiteboul, Pierre Bourhis, and Bogdan Marinoiu. Satisfiability and relevance for queries over active documents. In *PODS*, pages 87–96, 2009.

[5] Serge Abiteboul, Bogdan Cautis, and Tova Milo. Reasoning about XML update constraints. *Journal of Computer and System Sciences (JCSS)*, 2009.

[6] Serge Abiteboul, T-H. Hubert Chan, Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Aggregate Queries for Discrete and Continuous Probabilistic XML. In *Proc. ICDT*, 2010.

[7] Serge Abiteboul, Georg Gottlob, and Marco Manna. Distributed XML design. In *PODS*, pages 247–258, 2009.

[8] Serge Abiteboul, Ohad Greenshpan, Tova Milo, and Neoklis Polyzotis. Matchup: Autocompletion for mashups (demo). In *ICDE*, pages 1479–1482, 2009.

[9] Serge Abiteboul, Benny Kimelfeld, Yehoshua Sagiv, and Pierre Senellart. On the expressiveness of probabilistic XML models. *VLDB Journal*, 18(5):1041–1064, October 2009.

[10] Serge Abiteboul and Neoklis Polyzotis. Searching shared content in communities with the data ring. *IEEE Data Eng. Bull.*, 32(2):44–51, 2009.

[11] Serge Abiteboul, Philippe Rigaux, Marie-Christine Rousset, and Pierre Senellart. Web data management and distribution, 2009.

[12] Serge Abiteboul, Luc Segoufin, and Victor Vianu. Modeling and verifying active XML artifacts. *IEEE Data Eng. Bull.*, 32(3):10–15, 2009.

[13] Serge Abiteboul, Luc Segoufin, and Victor Vianu. Static analysis of active XML services. *ACM trans. on Database Systems (TODS)*, 34(4), 2009.

[14] Loredana Afanasiev and Balder ten Cate. On Core XPath with Inflationary Fixed Points. *FICS*, page 11, 2009.

[15] Cédric du Mouza, Witold Litwin, and Philippe Rigaux. Large-scale Indexing of Spatial Data in Distributed Repositories: the SD-Rtree. *VLDB Journal*, 18(4):933–958, 2009.

[16] Cédric du Mouza, Witold Litwin, Philippe Rigaux, and Thomas Schwarz. AS-Index: A Structure For String Search Using n-grams and Algebraic Signatures. In *Intl. Conf. on Information and Knowledge Management (CIKM)*, 2009. Honk-Kong, China.

[17] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating information from disagreeing views. In *WSDM (Web Search and Data Mining)*, New York, USA, February 2010.

[18] A. Gheerbrant and Balder ten Cate. Craig Interpolation for Linear Temporal Languages. In *Computer Science Logic: 23rd International Workshop, CSL 2009, 18th Annual Conference of the EACSL, Coimbra, Portugal, September 7-11, 2009, Proceedings*, page 287. Springer, 2009.

[19] Georg Gottlob and Pierre Senellart. Schema mapping discovery from data instances. *Journal of the ACM*, September 2009. To appear.

[20] Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Updating probabilistic XML. In *Proc. Updates in XML*, Lausanne, Switzerland, March 2010.

[21] Stefan Manegold, Ioana Manolescu, Loredana Afanasiev, Jianlin Feng, Gang Gou, Marios Hadjieleftheriou, Stavros Harizopoulos, Panos Kalnis, Konstantinos Karanasos, Dominique Laurent, Mihai Lupu, Nicola Onose, Christopher Ré, Virginie Sans, Pierre Senellart, Tianyi Wu, and Dennis Shasha. Repeatability & workability evaluation of SIGMOD 2009. *SIGMOD Record*, 38(3):40–43, September 2009.

[22] Bogdan Marinoiu. *Analysis and verification of distributed systems*. PhD thesis, Universit de Paris Sud, 2009.

[23] Balder ten Cate, Laura Chiticariu, Phokion Kolaitis, and Wang-Chiew Tan. Laconic Schema Mappings: Computing the Core with SQL Queries. *VLDB*, 2009.

[24] Balder ten Cate and Gaelle Fontaine. An Easy Completeness Proof for the Modal mu-calculus on Finite Trees. *FICS*, 2009.

[25] Balder ten Cate, Tadeusz Litak, and Maarten Marx. Complete Axiomatizations of XPath Fragments. *JAL*, 2009.

[26] Aparna Varde, Fabian M. Suchanek, Richi Nayak, and Pierre Senellart. Knowledge discovery over the deep Web, semantic Web and XML. In *Proc. DASFAA*, pages 784–788, Brisbane, Australia, April 2009. Tutorial.

[27] Victor Vianu. Automatic verification of database-driven systems: a new frontier. In *ICDT*, pages 1–13, 2009.