Number: 226513
Acronym: Webdam
Title: Foundations of Web Data Management

ERC Grant

Mid Term Activity Report

Period covered: 01/12/2008 to 31/05/2011
Start date of project: 01/12/2008
PI: Serge Abiteboul
Host: INRIA Saclay–Île-de-France
Date of preparation: 30 June 2011
Date of submission (SESAM): 30 June 2011
Duration: 30 Months

Goals The Webdam ERC grant (S. Abiteboul) started in December 2008. The goal is to develop a formal model for Web data management that would open new horizons for the development of the Web in a well-principled way, enhancing its functionality, performance, and reliability. More specifically, the aim is to create a universally accepted formal framework for describing complex and flexible interacting Web applications featuring notably data exchange, sharing, integration, querying and updating. It is also to develop formal foundations enabling peers to concurrently reason about global data management activities and cooperate in solving specific tasks with desired quality of service.

1 Summary of the major achievements

After 2.5 years (at mid term), we are happy to report that the Webdam project is solidly on track and progressing.

After a planning period where we identified research issues, we started investigating the main research themes of Webdam.

Major results were obtained on the following themes that are central to the project:

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. We believe this work will become central to the research of Webdam. We develop this in more detail in Section II.1.

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. We develop this in more detail in Section II.2.

Workflow for distributed applications The sequencing of tasks in a distributed setting is a complex issue that is little understood. We are pursuing fundamental work in the context of data-centric artifacts for modeling workflows in such environments, as well as for performing static analysis on these workflows. This is considered in Section II.3.

Progress was also achieved in a number of other directions perhaps less central to Webdam but essential for better understanding the many facets of Web data management:

The Semantic Web A main challenge facing the Web today is that of capturing its semantics. To follow up on this aspect, we worked on ontology alignment (Section II.4.2) and ontology evolution (Section II.4.3). We also worked on Web archiving (Section II.4.4), and on the management of deep Web data (Section II.4.5).

Distribution and optimization Another main issue that could hinder the development of the Web is that of poor performance of applications managing distributed data. Because of the scale of the Web, any model

or approach that would ignore optimization issues is bound to fail. This is why it is important for Webdam to pursue research on optimization. We proposed a general approach to the analysis and optimization of distributed data management (Section II.4.1). We introduced a model and efficient algorithms for monitoring (Section II.4.7). At a more fundamental level, we studied query enumeration (Section II.4.6).

Design We proposed pioneering work on distributed XML design (Section II.4.8).

From the point of view of the visibility of Webdam, one should mention:

Books Webdam put a lot of effort in a textbook (undergraduate and graduate level) on Web data management (nicknamed Jorge) to be published at Cambridge University Press [57]. The book is already available freely on the Webdam Web site:

`http://webdam.inria.fr/Jorge`

A previous book on Foundations of Data Management (nicknamed Alice) is also available freely on the Webdam Web site:

`http://webdam.inria.fr/Alice`

Webdam is pursuing leading an effort to complement the online version of Alice with new topics on semistructured data and on data integration, two central directions for the Web.

EC FoX In terms of collaboration, there is a significant synergy with the FP7 FoX project (managed by Luc Segoufin) on the Foundations of XML, also a key topic for the Web.

Workshops A workshop was organized in the early stage of the project (August 2009) to make more precise the research issues upon which most effort was to be focused. An internal workshop was then organized in March 2011 to evaluate the progress of the project. Webdam participated in the organization of a major international workshop on data-centric workflows [62] (in Dagstuhl, April 2010). Webdam is preparing another major international workshop on distributed data management [61] together with the FoX European project (in Dagstuhl, October 2011).

2 Publishable brief summary

The Webdam ERC grant is a 5-year project that started in December 2008. The goal is to develop a formal model for Web data management that would open new horizons for the development of the Web in a well-principled way, enhancing its functionality, performance, and reliability.

Information of interest may be found on the Web in a variety of forms, in many systems, with different access protocols. For instance, a standard user may have information on many devices (smartphone, laptop, TV box, etc.), many systems (mailers, blogs, web sites, etc.), many social networks (Facebook, Picasa, etc.). This same user may have access to more information from family, friends, associations, companies, etc., or organizations (tax, health, etc.). The control and management of this diversity are today beyond the skill of casual users. Facing similar issues, companies see the cost of managing and integrating information skyrocketing. The thesis of Webdam is that managing this diversity of data can be achieved using a *distributed knowledge base* handling both data and meta-data, as well as access control and localization information, in a unique holistic setting. We believe that complex Web data management tasks currently requiring deep expertise will be greatly facilitated by the automatic reasoning of the inference engine of the knowledge base. We have obtained fundamental results in that direction and started experimenting with a prototype system.

Another critical dimension of the problem is the imprecise and uncertain nature of data on the Web. Dealing with this aspect in a well-principled way led us to investigate probabilistic databases. This resulted in significant advances, including a novel approach to probabilistic XML (XML is the standard for data exchange on the Web). We have carried out in-depth studies of query evaluation, aggregation, and updates in various probabilistic XML models.

Finally, when supporting complex activities in a Web setting, one typically has to organize the cooperation between possibly many systems, and notably the sequencing of their tasks. The specification of such sequencing, sometimes referred to as choreography, is little understood. We are pursuing an original approach that models tasks with pieces of data, that are called *business artifacts* (following IBM terminology). The evolution of an artifact is constrained by rules on the evolution of the data. Using this approach, we are pursuing fundamental works in order to understand the intrinsic nature of workflows involving distributed systems.

Webdam also put a lot of effort in developing a textbook (advanced undergraduate or graduate level) on Web data management to be published at Cambridge University Press [57]. The book is already available on the Webdam Web site.

3 Major difficulties encountered

Complexity of the model A lot of work was devoted to the Active XML model during the first couple of years and important results obtained. When considering issues such as trust, access rights, or provenance in the context of social data, the tree aspect of Active XML model turned out to complicate some of the issues. So, we refocused on the relational model and Datalog-style languages for some of the more recent works, e.g., the work around Webdamlog. See Section II.1. We plan to work in the two coming years on reconciling the two approaches.

Localization In a first phase, the Webdam project was involving researchers of the Institut National de Recherche en Informatique et Automatique (INRIA), from the teams Gemo/Leo at Orsay University and Dahu at Ecole Nationale Supérieure de Cachan (ENS Cachan). The project also rapidly involved researchers at l'Ecole Nationale Supérieure des Télécommunications (Télécom ParisTech), around Pierre Senellart, notably on probabilistic data. It turned out to be more complicated than expected to focus the group in such a distributed environment. The research was concentrated after a couple of years at ENS Cachan and Télécom ParisTech.

Human Resources These form the main asset of a project such as Webdam. The first year we encountered difficulties for attracting talent. One major researcher we brought in (Balder ten Cate from Amsterdam) had to leave after a few months for personal reasons. On the other hand, we were very successful the second year. We notably brought on board Prof. Vianu from UCSD for a year (he was already collaborating with Webdam). We also attracted talented post docs, Daniel Deutch (PhD Tel Aviv), Bruno Marnette (PhD Oxford) and Fabian Suchanek (PhD Max Planck Institute). Also, Meghyn Bienvenu was hired at CNRS at Orsay and joined Webdam. With the arrival of Fabian and Meghyn, we finally had the human resources to investigate new important directions, in particular with respect to knowledge management. Webdam is now very proud of the team that has been put together.

At the end of 2011, it is expected that many researchers will leave Webdam, some to join new permanent positions, F. Suchanek (Max Planck), D. Deutch (Ben Gurion University, Israel), D. Gross-Amblard (Prof. University Rennes), others to return to their universities, V. Vianu (UCSD), or move to industry, B. Marnette (Winton Capital Management). This is an inherent issue: Talented people are offered (permanent) positions elsewhere and often accept them. We see this difficulty as an opportunity for Webdam to grow by developing collaboration with some of the members who left. Of course, we also intend to bring in new talents and effort is being invested to recruit them.

4 List of keywords

Distributed data, Database, Data extraction, Data integration
Data model, Query languages
Uncertainty, Inconsistency, Probabilistic data,
Semistructured data model, XML, Active XML, Probabilistic XML
Workflow, Business artifact
Social Networks, Content recommendation, Corroboration
World Wide Web, Web searching, Web indexing, Web applications
Web services, Web languages
Knowledge, Knowledge representation, Metadata, Datalog
Distributed query processing and optimization.

5 List of attachments

- Project output records
- Attachment I: Participants
- Attachment II: Research results in more detail
- Attachment III: Dissemination

Number: 226513
Acronym: Webdam
Title: Foundations of Web Data Management

Project output records

1 Publications

In the bibliography entries, the keywork *ack* specifies that the corresponding publication acknowledges the ERC funding of Webdam. The keyword *open* specifies that the access to this publication is open followed by a link to the electronic version.

The publications marked open are all reachable from the Webdam publications page.

Journals

[1] Serge Abiteboul, Bogdan Cautis, and Tova Milo. Reasoning about XML update constraints. *Journal of Computer and System Sciences*, 2009. open `http://www.math.tau.ac.il/~milo/projects/di_xml_semidata/papers/GemoReport-463.pdf`.

[2] Serge Abiteboul, Benny Kimelfeld, Yehoshua Sagiv, and Pierre Senellart. On the expressiveness of probabilistic XML models. *Journal on Very Large Databases*, 18(5):1041–1064, October 2009. ack, open `http://pierre.senellart.com/publications/abiteboul2009expressiveness.pdf`.

[3] Serge Abiteboul and Neoklis Polyzotis. Searching shared content in communities with the data ring. *IEEE Data Engineering Bulletin*, 32(2):44–51, 2009. ack, open `http://sites.computer.org/debull/A09June/alkis_dataengi1.pdf`.

[4] Serge Abiteboul, Luc Segoufin, and Victor Vianu. Modeling and verifying active XML artifacts. *IEEE Data Engineering Bulletin*, 32(3):10–

15, 2009. ack, open http://hal.inria.fr/docs/00/42/94/84/PDF/AbiteboulSegoufinVianu09.pdf.

[5] Serge Abiteboul, Luc Segoufin, and Victor Vianu. Static analysis of active XML services. *ACM Transactions on Database Systems*, 34(4), 2009. open http://www.lsv.ens-cachan.fr/~segoufin/Papers/File/axml.pdf.

[6] Cédric du Mouza, Witold Litwin, and Philippe Rigaux. Large-scale Indexing of Spatial Data in Distributed Repositories: the SD-Rtree. *Journal on Very Large Databases*, 18(4):933–958, 2009. open http://www.springerlink.com/index/0h53343652743r21.pdf.

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

[8] Georg Gottlob and Pierre Senellart. Schema Mapping Discovery from Data Instances. *Journal of the ACM*, 2010. ack, open http://hal.inria.fr/inria-00537238/PDF/paper.pdf.

[9] David Gross-Amblard. Query-preserving watermarking of relational databases and XML documents. *ACM Transaction on Database Systems*, 36:3, 2011. open http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.6.8734.

[10] Wojciech Kazana and Luc Segoufin. First-order query evaluation on structures of bounded degree. *Journal on Logical Methods in Computer Science*, 2011. ack, open http://arxiv.org/abs/1105.3583.

[11] Julien Lafaye, Jean Béguec, David Gross-Amblard, and Anne Ruas. Blind & squaring-resistant watermarking of vectorial building layers. *GeoInformatica*, 2011. ack, to appear.

[12] 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. open http://pierre.senellart.com/publications/manegold2009repeatability.pdf.

[13] Balder ten Cate, Laura Chiticariu, Phokion Kolaitis, and Wang-Chiew Tan. Laconic Schema Mappings: Computing the Core with SQL Queries. *Journal on Very Large Databases*, 2009. ack, open http://www.vldb.org/pvldb/2/vldb09-595.pdf.

[14] Balder ten Cate, Tadeusz Litak, and Maarten Marx. Complete Axiomatizations of XPath Fragments. *Journal of Applied Logic*, 2009. open `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.141.247&rep=rep1&type=pdf`.

Conferences

[15] Serge Abiteboul, Meghyn Bienvenu, Alban Galland, and Emilien Antoine. A rule-based language for Web data management. In *Proceedings of the Symposium on Principles of Database Systems*, 2011. ack, open `http://hal.inria.fr/docs/00/58/28/91/PDF/pods17a-abiteboul.pdf`.

[16] Serge Abiteboul, Pierre Bourhis, Alban Galland, and Bogdan Marinoiu. The AXML artifact model. In *Proceedings of the International Conference on Temporal Representation and Reasoning*, 2009. ack, open `http://www.labri.fr/perso/anca/docflow/publications_files/ABGM.pdf`.

[17] Serge Abiteboul, Pierre Bourhis, and Bogdan Marinoiu. Efficient maintenance techniques for views over active documents. In *Proceedings of the International Conference on Extending Database Technology*, pages 1076–1087, 2009. open `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.165.9831&rep=rep1&type=pdf`.

[18] Serge Abiteboul, Pierre Bourhis, and Bogdan Marinoiu. Satisfiability and relevance for queries over active documents. In *Proceedings of the Symposium on Principles of Database Systems*, pages 87–96, 2009. open `http://leo.saclay.inria.fr/publifiles/gemo/GemoReport-10019.pdf`.

[19] Serge Abiteboul, Pierre Bourhis, and Victor Vianu. Comparing Workflow Specification Languages: A Matter of Views. In *Proceedings of the International Conference on Database Theory*, 2011. ack, open `http://www.edbt.org/Proceedings/2011-Uppsala/papers/icdt/a9-abiteboul.pdf`.

[20] Serge Abiteboul, T-H. Hubert Chan, Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Aggregate Queries for Discrete and Continuous Probabilistic XML. In *Proceedings of the International Conference on Database Theory*, pages 50–61, 2010. ack, open `http://hal.inria.fr/inria-00537632/PDF/main.pdf`.

[21] Serge Abiteboul, Georg Gottlob, and Marco Manna. Distributed XML Design. In *Proceedings of the Symposium on Principles of Database Systems*, pages 247–258, 2009. ack, open `http://fox7.eu/wp-content/uploads/pods2.pdf`.

[22] Serge Abiteboul, Balder Ten Cate, and Yannis Katsis. On the Equivalence of Distributed Systems with Queries and Communication. In *Proceedings of the International Conference on Database Theory*, 2011. ack, open http://www.edbt.org/Proceedings/2011-Uppsala/papers/icdt/a13-abiteboul.pdf.

[23] Yael Amsterdamer, Daniel Deutch, Tova Milo, and Val Tannen. On provenance minimization. In *Proceedings of the Symposium on Principles of Database Systems*, pages 141–152, New York, NY, USA, 2011. ACM. ack, open http://www.cs.bgu.ac.il/~deutchd/publications/pods2011a.pdf.

[24] Yael Amsterdamer, Daniel Deutch, and Val Tannen. Provenance for aggregate queries. *The Computing Research Repository*, abs/1101.1110, 2011. ack, open http://arxiv.org/pdf/1101.1110v1.

[25] Michael Benedikt, Georg Gottlob, and Pierre Senellart. Determining Relevance of Accesses at Runtime. In *Proceedings of the Symposium on Principles of Database Systems*, 2011. ack, open http://fox7.eu/wp-content/uploads/pods16d-benedikt.pdf.

[26] Michael Benedikt, Dan Olteanu, Evgeny Kharlamov, and Pierre Senellart. Probabilistic XML via Markov Chains. In *Proceedings of the International Conference on Very Large Databases*, pages 770–781, 2010. ack, open http://hal.archives-ouvertes.fr/hal-00537778/PDF/main.pdf.

[27] Diego Calvanese, Evgeny Kharlamov, Werner Nutt, and Dmitriy Zheleznyakov. Evolution of DL-Lite Knowledge Bases. In *Proceedings of International Semantic Web Conference*, pages 112–128, 2010. ack, open http://hal.archives-ouvertes.fr/hal-00537785/PDF/CalvaneseEtAl2010EvolutionDLLite.pdf.

[28] Elio Damaggio, Alin Deutsch, and Victor Vianu. Artifact systems with data dependencies and arithmetic. In *Proceedings of the International Conference on Database Theory*, pages 66–77, New York, NY, USA, 2011. ACM. ack, open http://www.edbt.org/Proceedings/2011-Uppsala/papers/icdt/a8-damaggio.pdf.

[29] Cédric du Mouza, Witold Litwin, Philippe Rigaux, and Thomas Schwarz. AS-Index: A Structure For String Search Using n-grams and Algebraic Signatures. In *Proceedings of the International Conference on Information and Knowledge Management*, 2009. ack, open http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.161.2287&rep=rep1&type=pdf.

[30] Zoé Faget, Philippe Rigaux, David Gross-Amblard, and V. Thion Goasdoué. Modeling synchronized time series. In *Proceedings of the International Database Engineering and Applications Symposium*, pages 82–89, 2010. ack, open `http://cedric.cnam.fr/~thionv/Publications/2010/IDEAS2010_Faget_Gross_Rigaux_Thion.pdf`.

[31] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating Information from Disagreeing Views. In *Proceedings of the International Conference on Web Search and Data Mining*, 2010. ack, open `http://hal.inria.fr/inria-00429546/PDF/document.pdf`.

[32] David Gross-Amblard, Philippe Rigaux, Lylia Abrouk, and Nadine Cullot. Fingering Watermarking in Symbolic Digital Scores. In *Proceedings of the International Conference on Music Information Retrieval*, 2009. open `http://www.lamsade.dauphine.fr/scripts/FILES/publi1101.pdf`.

[33] Bruno Marnette and Floris Geerts. Static analysis of schema-mappings ensuring oblivious termination. In *Proceedings of the International Conference on Database Theory*, 2010. ack, open `http://disi.unitn.it/~p2p/RelatedWork/Matching/icdt10_Marnette.pdf`.

[34] Bruno Marnette, Giansalvatore Mecca, and Paolo Papotti. Scalable data exchange with functional dependencies. In *Proceedings of the International Conference on Very Large Databases*, 2010. ack, open `http://www.vldb.org/pvldb/vldb2010/pvldb_vol3/R09.pdf`.

[35] Balder ten Cate and Gaelle Fontaine. An Easy Completeness Proof for the Modal mu-calculus on Finite Trees. In *Proceedings of the International Workshop on Fixed Points in Computer Science*, 2009. ack, open `http://users.soe.ucsc.edu/~gaelle/fossacsend2.pdf`.

[36] Remi Tournaire, Alexandre Termier, Jean-Marc Petit, and Marie-Christine Rousset. Combining logic and probabilities for discovering mappings between taxonomies. In *Proceedings of the International Conference on Knowledge Science, Engineering and Management*, 2010. ack, open `http://www.springerlink.com/content/d110g2t62n5gq8mm/`.

[37] Victor Vianu. Automatic verification of database-driven systems: a new frontier. In *Proceedings of the International Conference on Database Theory*, pages 1–13, 2009. open `http://www.edbt.org/Proceedings/2009-StPetersburg/icdt/papers/p0001-Vianu.pdf`.

Workshops

[38] Serge Abiteboul, Meghyn Bienvenu, Alban Galland, and Marie-Christine Rousset. Distributed Datalog Revisited. In *Proceedings of the Datalog*

2.0 International Workshop, 2011. ack, open http://hal.inria.fr/inria-00540814/PDF/datalog20-serge.pdf.

[39] Serge Abiteboul, Alban Galland, and Neoklis Polyzotis. A model for web information management with access control. In *Proceedings of the International Workshop on the Web and Databases*, 2011. ack, open webdb2011.rutgers.edu/papers/Paper\%2013/WebDB11.pdf.

[40] Loredana Afanasiev and Balder ten Cate. On Core XPath with Inflationary Fixed Points. In *Proceedings of the International Workshop on Fixed Points in Computer Science*, page 11, 2009. ack, open http://cs.ioc.ee/fics09/fics09proc.pdf.

[41] Yael Amsterdamer, Daniel Deutch, and Tova Milo. On the optimality of top-k algorithms for interactive web applications. In *Proceedings of the International Workshop on the Web and Databases*. ACM, 2011. ack, open http://webdb2011.rutgers.edu/papers/Paper\%2011/InteractiveWebApp.pdf.

[42] Diego Calvanese, Evgeny Kharlamov, Werner Nutt, and Dmitriy Zheleznyakov. Updating ABoxes in DL-Lite. In *Proceedings of Alberto Mendelzon Workshop on Foundations of Data Management (AMW)*, 2010. ack, open http://hal.archives-ouvertes.fr/hal-00537787/PDF/main.pdf.

[43] Bogdan Cautis and Evgeny Kharlamov. Challenges for View-Based Query Answering over Probabilistic XML. In *Proceedings of the Alberto Mendelzon International Workshop on Foundations of Data Management (AMW)*, 2011. ack, open http://hal.inria.fr/inria-00591913/PDF/main.pdf.

[44] A. Gheerbrant and Balder ten Cate. Craig Interpolation for Linear Temporal Languages. In *Proceedings of International Workshop on Computer Science Logic*. Springer, 2009. ack, open http://www.springerlink.com/content/l0l324k014410278/.

[45] Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Updating Probabilistic XML. In *Proceedings of Workshop on Updates in XML; at the International Conference on Extending Database Technology*, 2010. ack, open http://hal.archives-ouvertes.fr/hal-00537793/PDF/main.pdf.

[46] Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Value Joins are Expensive over (Probabilistic) XML. In *Proceedings of the Workshop on Logic in Databases ; at the International Conference on Extending Database Technology*, 2011. ack, open http://hal.inria.fr/inria-00591905/PDF/main.pdf.

[47] Evgeny Kharlamov and Dmitriy Zheleznyakov. Understanding Inexpressibility of Model-Based ABox Evolution in DL-Lite. In *Proceedings of the Alberto Mendelzon International Workshop on Foundations of Data Management*, 2011. ack, open `http://hal.inria.fr/inria-00591918/PDF/main.pdf`.

[48] Marilena Oita and Pierre Senellart. Archiving Data Objects using Web Feeds. In *Proceedings of the International Workshop on Web Archiving*, 2010. ack, open `http://hal.inria.fr/inria-00537962/PDF/iwawienna.pdf`.

[49] Marilena Oita and Pierre Senellart. Deriving Dynamics of Web Pages: A Survey. In *Proceedings of the Temporal Workshop on Web Archiving*, 2011. ack, open `http://hal.inria.fr/inria-00588715/PDF/survey.pdf`.

[50] Fabian Suchanek, Alessandro Bozzon, Emanuele Della Valle, Alessandro Campi, and Stefania Ronchi. Towards an Ontological Representation of Services in Search Computing. In Stefano Ceri and Marco Brambilla, editors, *New Trends in Search Computing*. Springer, 2011. ack, open `http://hal.inria.fr/inria-00591790/PDF/seco_yago.pdf`.

[51] Remi Tournaire, Alexandre Termier, Jean-Marc Petit, and Marie-Christine Rousset. Probamap: a scalable tool for discovering probabilistic mappings between taxonomies. In *Proceedings of the International Workshop on Automated Knowledge Base Construction*, 2010. open `akbc.xrce.xerox.com/IMG/pdf/CameraReady_TournaireR.pdf`.

[52] Dmitriy Zheleznyakov, Diego Calvanese, Evgeny Kharlamov, and Werner Nutt. Updating TBoxes in DL-Lite. In *Proceedings of International Workshop on Description Logics*, 2010. ack, open `http://hal.archives-ouvertes.fr/hal-00537790/PDF/main.pdf`.

Demonstrations (in proceedings of international conferences)

[53] Serge Abiteboul, Pierre Bourhis, Bogdan Marinoiu, and Alban Galland. [Demo] AXART - Enabling Collaborative Work with AXML Artifacts. In *Proceedings of the International Conference on Very Large Database*, 2010.

[54] Serge Abiteboul, Ohad Greenshpan, Tova Milo, and Neoklis Polyzotis. Matchup: Autocompletion for mashups (demo). In *Proceedings of the International Conference on Database Engineering*, pages 1479–1482, 2009.

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

[56] Johannes Hoffart, Fabian Suchanek, Klaus Berberich, Edwin Lewis Kelham, Gerard De Melo, and Gerhard Weikum. [Demo] YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages. In *Proceedings of the World Wide Web Conference - Demos*, 2011.

Book

[57] Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, and Pierre Senellart. *Web Data Management*. Cambridge University Press, 2011 (to appear). ack open `http://webdam.inria.fr/Jorge`.

Book Chapter

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

[59] Evgeny Kharlamov and Pierre Senellart. Modeling, Querying, and Mining Uncertain XML Data. In Andrea Tagarelli, editor, *XML Data Mining: Models, Methods, and Applications*. IGI Global, 2011. ack.

Others

[60] Serge Abiteboul, Klemens Böhm, Christoph Koch, and Kian-Lee Tan, editors. *Proceedings of the International Conference on Data Engineering*. IEEE Computer Society, 2011.

[61] Serge Abiteboul, Alin Deutsch, Thomas Schwentick, and Luc Segoufin, editors. *Proceedings of the Dagstuhl Seminar 11421 "Foundations of distributed data management"*, 2011, to appear.

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

[63] Serge Abiteboul, Volker Markl, Tova Milo, and Jignesh Patel. Special issue: Best papers of international conference on very large databases 2009. *Journal on Very Large Databases*, 20(2):155–156, 2011.

[64] Serge Abiteboul, Volker Markl, Tova Milo, Jignesh Patel, and Philippe Rigaux, editors. *Proceedings of the 27th International Conference on Very Large Databases*, 2009.

[65] Pierre Bourhis. *On the Dynamics of Active Documents for Distributed Data Management*. PhD thesis, Université Paris-Sud, 2011.

[66] David Gross-Amblard. *Tatouage des bases de données*. Habilitation à Diriger des Recherche, Université de Bourgogne, 12 2010.

[67] Evgeny Kharlamov. *A Probabilistic Approach to XML Data Management*. PhD thesis, Faculty of Computer Science, Free University of Bozen-Bolzano, 2011.

[68] Hady Lauw, Ralf Schenkel, Fabian Suchanek, Martin Theobald, and Gerhard Weikum. Harvesting Knowledge from Web Data and Text. In *Proceedings of the International Conference on Information and Knowledge Management*, 2010.

[69] Bogdan Marinoiu. *Analysis and verification of distributed systems*. PhD thesis, Université Paris Sud, 2009.

[70] Nicoleta Preda, Gjergji Kasneci, Fabian Suchanek, Thomas Neumann, Wenjun Yuan, and Gerhard Weikum. Active Knowledge: Dynamically Enriching RDF Knowledge Bases by Web Services. In *Proceedings of the ACM SIGMOD Conference on the Management of Data*, 2010. Tutorial.

[71] Fabian Suchanek, Martin Theobald, Gerhard Weikum, Hady Lauw, and Ralf Schenkel. Semantic Knowledge Bases from Web Sources. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2011. Tutorial.

[72] Fabian Suchanek, Aparna Varde, Richi Nayak, and Pierre Senellart. The Hidden Web, XML and Semantic Web: A Scientific Data Management Perspective. In *Proceedings of the International Conference on Extending Database Technology*, 2011. ack. Tutorial.

2 Research expeditions

Not applicable

3 Awards and recognitions

3.1 Research Prizes

- Victor Vianu received the 2010 Alberto O. Mendelzon Test of Time Award of ACM PODS for the article *Typing for XML Transformers*,

joint with Tova Milo and Dan Suciu.

- Daniel Deutch won the ICDT 2011 Best Student Paper Award
- Meghyn Bienvenu received the Association Francaise pour l'Intelligence Artificielle Prize for her thesis in 2010.
- Fabian Suchanek received the Otto-Hahn-Medal, the Dissertation Award of the Max Planck Society in 2010.
- Fabian Suchanek was also selected for a research group leader stipend by the Max Planck Society in 2010.
- Fabian Suchanek won the ACM Dissertation Award Honorable Mention for his dissertation in 2010.
- The demo of the YAGO ontology [56], to which Fabian Suchanek contributed, won the Best Demo Award at the WWW 2011 conference.

3.2 Memberships of a learned society

- Serge Abiteboul has been elected in 2009 member of the French Academy of Sciences.
- Marie-Christine Rousset has been elected in 2011 senior member of the Institut Universitaire de France.

3.3 Memberships of editorial boards

- Victor Vianu is Editor-in-Chief of the *Journal of the ACM* and Area Editor for *ACM Trans. on Computational Logic* (logical aspects of databases).
- Serge Abiteboul is a member of the steering committee of the *Proceedings of the VLDB Endowment* (*PVLDB*) journal, a recently launched journal.
- Pierre Senellart is Information Director of *JACM*.

3.4 Others

- Serge Abiteboul served as General Program Chair of [1] the International Conference on Data Engineering (ICDE) 2011 in Hanover, Germany (one of the main conferences on database systems) [60].

[1] In this text, the first time a conference or journal is encountered, we give the name in full and the acronym. We then use the acronym.

He served as General Program Chair of the International Conference on Very Large Databases (VLDB) 2009, the main conference on database systems) in Lyon, France [64, 63].

These are two out of the three top conferences in database systems. The third one is the ACM SIGMOD International Conference on the Management of Data (SIGMOD) where Webdam was also strongly represented; see Attachment III.

- Ioana Manolescu served as Co-Chair of the Web Engineering Track in the World Wide Web Conference (WWW) 2009. She served as Co-Chair of the "Systems, Experiments, Applications" Track in ICDE 2011.

- Pierre Senellart was Submission Chair of VLDB 2009, Tutorial Co-Chair of ICDE 2010, and Program Chair of the Industrial Track of the International Conference on Extending Database Technology (EDBT) 2011.

Members of Webdam also gave invited presentations in a number of conferences and workshops. They have been regularly invited to serve in the program committees of the best conferences. See Attachment III.

4 Patents, licencing, intellectual property

We are in the process of licencing the software used in two demonstrations [53, 55].

5 Dissemination to non-academic audience

Serge Abiteboul made the following presentations:

- The Future of Communications and the Web, Futur en Seine, 2011.
- Le Web de demain: promesses et écueil, Conférences de la Communauté Israélite de Genève, 2010.
- Web et révolution numérique: promesses et écueils, Sèvres Débat, 2010.
- Demain, la République du Web : une utopie ? La Cantine, 2009.
- Web data management, Colloque Une histoire de DIM, Domaines d'Intérêt Majeur, Paris 2009.
- Web et Industrie, Rencontres INRIA-Industrie, 2009.

- Gestions de données distribuées, Perspectives IT pour le Codir des Clubs OLG, 2009.

He participated in a radio program for the 20 years of the Web: Le Téléphone Sonne, Alain Bedouet, France Inter, 2009.

He gave interviews for

- Pour la Science, 2009.
- Le Nouvel Economiste, 2009,
- L'informaticien, 2009
- Le Sévrien 2009
- Science et Vie, Nos 3 questions, 2009
- Specif, 2009

6. Other significant output

See Attachment III.

I Participants

In this section, we detail the human participation in Webdam. When there is a single date, this indicates that the person is still working for Webdam.

Researchers and professors

- Serge Abiteboul, senior researcher *principal investigator* [01/12/2008]
- David Gross-Amblard, assistant professor at University of Bourgogne, seconding [01/09/2010]
- Yannis Katsis, post doc [01/09/2009]
- Bruno Marnette, post doc [01/09/2010-31/07/2011]
- Philippe Rigaux, professor at Dauphine University, seconding at Webdam [01/09/2009-01/10/2010]
- Marie-Christine Rousset, professor at University of Grenoble, seconding at Webdam [01/09/2009]
- Fabian Suchanek, post doc [01/06/2010-31/07/2011]
- Daniel Deutch, post doc [01/02/2011]
- Balder ten Cate, researcher [01/04/2009-01/10/2009]
- Victor Vianu, professor at UC San Diego [01/07/2010]

PhD Students

- Yael Amsterdamer, Tel Aviv University [18/02/2011]
- Emilien Antoine, Paris Sud University [01/10/2010]
- Pierre Bourhis, Paris Sud University [01/12/2008-01/03/2011]
- Alban Galland, Paris Sud University [01/12/2008]
- Wojciech Kazana, ENS Cachan [01/02/2010]
- Evgeny Kharlamov, Free University of Bozen-Bolzano [01/01/2009-31/12/2010]
- Bogdan Marinoiu, Paris Sud University [0112/2008-01/09/2009]
- Marilena Oita, Télécom ParisTech [01/10/2009]

Other full-time members

- Alin Gabriel Tilea, engineer [01/09/2009-01/06/2010]
- Kristian Lyngbaek, intern (Delft University) [07/2010-05/2010]
- Jules Testart, intern (McGill University) [05/2011-06/2011]

Collaborators

- Meghyn Bienvenu, CNRS researcher at Orsay University [01/04/2010]
- Ioana Manolescu, INRIA senior researcher, manager of the Gemo team, INRIA Saclay [01/12/2008]
- Luc Segoufin, INRIA senior researcher, manager of the Dahu team, INRIA Saclay [01/12/2008]
- Philippe Rigaux, professor at Paris CNAM [01/10/2010]
- Pierre Senellart, associate professor at Télécom ParisTech [01/12/2008]
- Victor Vianu, professor at UC San Diego [01/12/2008-01/07/2010]

Webdam Alumni

- Webdam members are very successful in academia. Balder ten Cate obtained an assistant professor position at UC Santa Cruz. Pierre Bourhis is a postdoc at Oxford University. Some members who are still at webdam, obtained positions they will join in the near future. David Gross-Amblard obtained an associate professor position at Rennes University; Daniel Deutch obtained an assistant professor position at Ben Gurion University, Israel. Fabian Suchanek obtained a research group leader position at Max Planck Institute, Germany. B. Marnette is joining Winton Capital Management.
- Members of Webdam are also successful in industry, Bogdan Marinoiu, moved to SAP (working on Data warehouses), Philippe Rigaux is moving to the Internet Memory Foundation, with whom we collaborate; see Section II.4.4.

Visitors (one month or more)

- Balder ten Cate Professor at UC Santa Cruz (1 month)
- Bruno Marnette, PhD student at Oxford (2 months)
- Amélie Marian, Assistant Professor at Rutgers University (2 months)
- Tova Milo, Professor at Tel Aviv U. (1 month)

- Yannis Papakonstantinou, Professor at UC San Diego (1 month)
- Alkis Polizotis, Professor at UC Santa Cruz (1 month)
- Sihem Amer-Yahia, Yahoo Research (1 month)

We also had many shorter visits, notably, Gerome Miklau (U. Washington) and Moshe Vardi (Rice U.).

II Research results in more detail

In this section, we detail the scientific contributions. Most of the cited papers may be found at: `http://webdam.inria.fr/`

II.1 The Web as a distributed knowledge base

The management of data in a heavily distributed setting such as the Web poses a number of challenging issues such as how to find data, how to control data access, and how to integrate data coming from different sources. Webdam is investigating a holistic approach to support these complex data management tasks which is based on reasoning over a *distributed knowledge base*.

WebdamExchange We propose a knowledge-base model, called WebdamExchange, for sharing information on the Web, where the information is hosted on different machines that may use different access control and distribution schemes. The model uses logical statements for specifying data, access control, distribution and knowledge about other peers. The statements can be communicated, replicated, queried, and updated, while keeping track of time and provenance. This unified basis allows applications to reason about which data is accessible, where it resides, and how to retrieve it securely.

A system supporting this model has been implemented and demonstrated in the ICDE 2011 conference [55]. The demonstration illustrates how users can keep control over their data even in a social network that facilitates exchanges. In particular, it shows how users within very different data-distribution schemes (centralized, DHT, gossiping in an unstructured P2P, etc.) and different access control schemes, can transparently collaborate while keeping a good control over their own data even when some of their data resides in standard website or social network systems such as Facebook. The demonstration also illustrates how users can even control data using a device with limited storage and processing capabilities such as a smart phone.

Webdamlog In WebdamExchange, the peers' policies are hard-wired (coded in Java, in the prototype system). One would like users to be able to specify their own policies or customize existing ones. For that, we want to use a declarative language. Indeed, we strongly believe that the specification of modern distributed applications should be based on a reasonably simple (declarative) language that would hide most of the unnecessary details of distributed data management. This is in the spirit of work from Berkeley University on declarative programming for distributed systems, notably around the Dedalus language.

Preliminary work on this topic was presented at the Datalog 2.0 workshop [38]. Then in [15], we introduced a novel Datalog-style rule-based language, called Webdamlog. In this language, peers exchange messages (i.e. logical facts) as well as rules. The model is formally defined, and its interest for distributed data management is illustrated through a variety of examples. We validate the semantics of our model by showing that under certain natural conditions, our semantics converges to the same semantics as the centralized system with the same rules. Indeed, we can show this is even true when updates are considered. Another major contribution of this work is a study of the impact on expressiveness of 'delegations" (the installation of rules by a peer in some other peer) and explicit timestamps.

This research is part of the PhD thesis of Alban Galand which is nearing completion and that of Emilien Antoine which recently started, both under the direction of Serge Abiteboul.

II.2 The Web as a probabilistic world

The information found on the Web is typically uncertain, imprecise, possibly inconsistent. Also we may wrongly interpret it. This leads to issues such as data quality or trust. This is a major theme for Webdam. Our approach is based on probabilities and more precisely, on *probabilistic trees* (i.e., probabilistic XML). We have made important progress in this topic, studying the modeling, querying, and more generally management, of probabilistic XML.

Models for Probabilistic XML We started by proposing a general model for probabilistic trees that encompasses all previously proposed models (including local and global probabilistic dependencies) and studying the respective expressiveness of these [2]. With this framework, it is possible to represent in a possibly compact manner all discrete and finite probability distributions over trees. This naturally leads to the question of representing continuous distributions, or discrete infinite distributions (with trees of unbounded size). In [20], we provide a formal model for continuous distributions in probabilistic trees and shows 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. In [26], a joint work with the FoX FP7 project, we overcome the limitation that trees described in probabilistic XML have a bounded size with the help of recursive Markov chains (RMCs), a formalism for probabilistic processes with recursive calls. We explain how to use them to generate probabilistic trees. We show that 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.

Querying Probabilistic XML Previous work (notably by Cohen, Kimelfeld, and Sagiv) had investigated the complexity of, and algorithms for, evaluating tree-pattern queries (and monadic second-order logic) over classical probabilistic XML models. In addition to extending these results to our more general settings (e.g., in [26]), we have extended query capabilities over probabilistic XML in a number of directions. In [46], we extend tree-pattern queries with value joins and show that for simple classes of such queries, there is a dichotomy between tractable and intractable queries over probabilistic XML models with local dependencies. In [20], 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. In [43], we investigate the problem of answering queries using views for probabilistic XML.

Managing Probabilistic XML We looked at other issues of interest pertaining to probabilistic XML. Initial studies of updates in probabilistic XML [2] 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 investigate the problem further in [45] and show 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. Finally, in [59], we show how some data mining problems can be solved using probabilistic XML querying techniques.

This research is part of the PhD thesis of Evgeny Kharlamov, under the direction of Werner Nutt (Free University of Bozen-Bolzano) and Pierre Senellart. As part of his thesis, Evgeny Kharlamov passed 1.5 years in Paris, within Webdam.

Corroboration In an environment with many independent participants, one typically finds conflicting opinions. In [31], we study the problem of corroborating information coming from a large number of participants. We propose and evaluate various algorithms towards this goal that provide improvements over known techniques such as voting, that often already perform reasonably well in practice.

II.3 Sequencing tasks on the Web

The orchestration of distributed activities (under the responsibility of a conductor) and their choreography (when they are fully autonomous) are complex issues that are essential for a wide range of applications including notably, e-commerce systems, business processes, health-care and scientific workflows. The difficulty is to guarantee consistency or more generally, quality of service, and to statically verify critical properties of the system. This is a theme where Webdam is very active with pioneering works and fundamental results.

There has recently been a proliferation of workflow specification formalisms. One such approach, proposed at IBM by Nigam and Caswell in 2003, is that of *data-centric workflow*, that places data at the center of the process, controlling sequencing by constraining the evolution of this data. This is also the philosophy of Webdam, with evolving trees (Active XML) at the center of the workflow.

In brief, Active XML consists of XML documents (the standard of the Web for data exchange) 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.

Verifying temporal properties of runs In [5, 4], we study the verification of temporal properties of runs of Active XML systems, specified in a tree-pattern-based temporal logic, namely Tree-LTL, 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.

In collaboration with UC San Diego and IBM organized around Prof. Victor Vianu, we studied the verification problem for a restricted class of data-centric workflows (in the spirit of IBM's Business Artifacts). However, these 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 [**?**], 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.

Alternative specification approaches There are different approaches to workflow specification such as automata-based, logic-based, or predicate-based control of function calls. In [19], we reconcile them by showing that they can be seen as closely related points of view on the evolution of data.

Specifically, we propose in [19] a flexible framework for comparing workflow specification languages, in which the pertinent aspects to be taken into account are defined by *views*. We use these views to compare the expressiveness of different workflow specification mechanisms based on automata, pre/post conditions, and temporal constraints. (Note the intrinsic difficulty of the problem because of the lack of a standard yardstick for measuring expressiveness.)

The Active XML Artifact model In [16], we introduce the *Active XML artifact model* to capture data and workflow management activities in distributed settings. We argue that the model is a natural extension of the Business Artifact model of Nigam and Caswell.

A prototype named AXART based on this model has been developed and demonstrated at VLDB 2010 [53]. It uses an application taken from the movie industry, that specifies task sequencing when managing actors applications for roles in films. Because the system builds on [16], it allows for complex organization of tasks. Because it borrows from [19], the system supports different ways of expressing workflow constraints, a rather unique feature.

This research is part of the PhD thesis of Pierre Bourhis, under the direction of Serge Abiteboul.

II.4 Other topics

II.4.1 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 [22]. 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 a full axiomatization of equivalence.

II.4.2 Ontologies and ontology alignment

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 progresses on four research avenues. 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 [2] by a temporal and spatial dimension [56]. 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 [50]. Third, we are working on algorithms that can detect and merge equivalent entities, classes and properties in different ontologies. With more and more ontologies becoming available on the Web, it is becoming important to discover their overlap in order to harness their synergies. Last, we are working on ownership proofs for ontological data, in the spirit of [9].

II.4.3 Evolution of ontologies

Description Logics (DLs) provide excellent mechanisms for representing structured knowledge by means of ontologies, and as such they constitute the foundations for the various fragments of OWL, the standard ontology language of the Semantic Web. As a response to the dynamicity of the Web, we have studied the problem of evolution for ontologies expressed in different DLs. We propose some fundamental principles that evolution should respect in [27], and study evolution for ontologies expressed in OWL 2 QL, a tractable fragment of OWL 2. We review known model and formula-based approaches for evolution and exhibited their limitations. Building upon the insights gained, we propose two novel formula-based approaches and develop polynomial time algorithms for them [27, 42, 52]. We also consider in [47] model-based approaches to analyze why these approaches raise difficulties when considering ontology evolution.

II.4.4 Web archiving

In the course of the Webdam project, we investigated 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 study how Web feeds can be used for archiving Web pages containing 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,

[2] See article by F.M. Suchanek and G. Kasneci and G. Weikum in WWW 2007.

we propose 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 [48] for further details. We also present in [49] a survey on techniques used for timestamping and detecting changes of pages on the Web.

This research is in cooperation with the Internet Memory Foundation and is part of the PhD thesis of Marilena Oita, under the direction of Pierre Senellart.

II.4.5 Deep Web data

A part of the Web is hidden to traditional Web search engines, because its data lie behind Web form interfaces. Understanding how to exploit in an automatic manner data from the deep Web is a challenging task, and we have laid some foundations of deep Web data management in Webdam. In [8], we propose a formal model for discovering the optimal schema mapping between two deep Web sources, allowing to understand how the data presented in one source can be derived from the other. In [25], we investigate the notion of *relevance* of a form interface to a query, i.e., when using a particular form can be useful in a query evaluation task.

This work is in collaboration with the ERC Diadem project (G. Gottlob, University of Oxford).

We also consider the problem of determining the dynamic relevance of a service in the presence of binding patterns in an ongoing joint work with Michael Benedikt from University of Oxford. There are also deep connections with the work mentioned further in Section II.4.7.

II.4.6 Query enumeration

In many applications, the output of a query may have a huge size and enumerating all the answers may already consume too much of the allowed resources. When this is the case, it may be appropriate to first output a small subset of the answers and then, on demand, output a subsequent small number of answers and so on until all possible answers have been exhausted. To make this even more useful, it is preferable to be able to minimize the time necessary to output the first answers as well as the time between consecutive sets of answers; this second time interval is known as the *delay*. This problem can be tackled by computing appropriate auxiliary index structures. The goal is to construct in reasonable time (say in linear time in the size of the database), index structures allowing extremely short delay in the enumeration process (say constant time). We are interested in understanding under what assumptions (on both the query language and the nature of the database) such index and algorithms exist. This approach was shown possible in [10] within the following scenario: solutions to any query

expressible in first-order logic over a relational structure of bounded degree can be enumerated, after a linear preprocessing, with a constant delay.

This research is part of the PhD thesis of Wojciech Kazana, under the direction of Luc Segoufin.

II.4.7 Monitoring

We have worked on the conception and implementation of tools for monitoring Peer to Peer Systems. Before Webdam started, a system named P2PMonitor has been developed for this purpose. It is a P2P system itself, with peers exchanging messages by Web-service calls. In Webdam, we have studied what we believe is at the heart of such distributed monitoring: the maintenance of views 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. In [17], we develop algorithmic Datalog-based foundations for such an incremental query processing. We also study theoretical issues raised in this context, such as whether a query over some active documents can be satisfied and whether a particular input stream is relevant for a given query [18].

This research is part of the PhD theses of Bogdan Marinoiu and Pierre Bourhis, both under the direction of Serge Abiteboul.

II.4.8 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 [21], 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.

III Dissemination

III.1 PhD Theses

Bogdan Marinoiu (2009) *Analysis and verification of distributed systems* [69].

Pierre Bourhis (2011) *On the dynamics of active documents for distributed data management* [65].

Evgeny Kharlamov (2011) *A Probabilistic Approach to XML Data Management* [67].

III.2 Habilitation thesis

David Gross-Amblard (2010) *Database Watermarking* [66].

III.3 Participation in conference programs

Program committees

Meghyn Bienvenu has served on the Program Committees of the International Joint Conference on Artificial Intelligence (IJCAI) 2011, the AAAI Conference on Artificial Intelligence 2010 and 2011, and the European Conference of Artificial Intelligence (ECAI) 2010. She is also a member of the Program Committee of the International Description Logic Workshop (2009, 2010, 2011), in addition to serving on the workshop's Steering Committee (elected 2009-2012).

David Gross-Amblard has served on the Program Committees of Bases de Données Avancées (BDA 2010, BDA 2011).

Ioana Manolescu served on the Program Committee of SIGMOD 2009.

Ioana Manolescu and Pierre Senellart participated in the Repeatability & Workability Evaluation Track of SIGMOD 2009 [12].

Pierre Senellart served on the Program Committees of BDA 2010, the Conference on Information and Knowledge Management (CIKM) 2010, ICDE 2010, the International Conference on Database Theory (ICDT) 2011, and WWW 2009 & WWW 2010.

Fabian Suchanek served on the Program Committee of WWW 2011.

Victor Vianu served on the Program Committees of ACM Symposium on Principles of Database Systems (PODS) 2009, VLDB 2009, IEEE Symposium on Logic in Computer Science (LICS) 2010, VLDB 2010, International Symposium on Foundations of Information and Knowledge Systems (FoIKS) 2010, Alberto Mendelzon Workshop on Foundations of Data Management (AMW) 2010 and 2011.

III.4 Participation in conference organization

Webdam participated in the organization of *two major workshops* centered around its research themes both in the Leibniz Center for Informatics (Dagstuhl):

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) [62].

Serge Abiteboul and Luc Segoufin are co-organizing with Alin Deutsch (UCSD, USA) and Thomas Schwentick (TU Dortmund, Germany) the Dagstuhl Workshop on Foundations of Distributed Data Management (10/2011) [61].

Pierre Senellart organized (together with Serge Abiteboul) the 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 worldwide [7].

Fabian Suchanek will organize the PhD Workshop at CIKM 2011, together with Anisoara Nica (Sybase Inc.). The workshop aims to give PhD students an opportunity to present their dissertation research at a relatively early stage. The workshop focuses on databases, information retrieval and knowledge management.

III.5 Invited presentations and tutorials

- Serge Abiteboul gave a keynote presentation on "Active XML Artifact" at the International Symposium on Temporal Representation and Reasoning (TIME) 2009 [16]. He gave a keynote presentation on "Web Information Management and Knowledge Bases" at the 10th International Conference on Web Engineering, Vienna (07/2010)

  He gave an invited presentation on "Web Data Management" at the Datalog 2.0 Workshop, held in March 2010, at Oxford University [38]; 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" at the Workshop on Formal Methods for Web Data Trust and Security; an invited presentation on "Web Data Management" at Centre d'Alembert and French Academy of Sciences.

He participated in a panel on "Distributed Data Management" at ICDE2011.

He gave a course on "Web Data Management" at the 4th Winter School in Hot Topics in Distributed Computing (La Plagne 2011).

- Ioana Manolescu gave an invited presentation at XML Symposium & Database and Programming Languages workshops (VLDB'09).
- Pierre Senellart and Fabian Suchanek presented a tutorial on "Data Management over the Deep Web and the Semantic Web" at EDBT 2011 [72].
- Fabian Suchanek was an invited expert at the Search Computing Workshop in Milan. He contributed to a tutorial on "Harvesting the Web of Data" at CIKM 2010, and on "Web-scale Knowledge Bases" at IJCAI 2011.
- Serge Abiteboul and Victor Vianu gave two of the invited presentations (out of five) at the PODS special workshop organized to celebrate the 30th anniversary of the conference (06/2011). They spoke respectively on "Trees, semistructured data, and other strange ways to go beyond tables" and on "Database Theory: Back to the Future".

III.6 Webdam workshops

Besides the two Dagstuhl workshops, Webdam organized the following events:

- A Webdam kickoff meeting at ENS Cachan in January 2009.
- A brainstorming workshop at Télécom ParisTech in August 2009 (Post VLDB) with members of the advisory board[3].
- A Workshop on Modal Logic (organized by Balder ten Cate) at ENS Cachan in July 2009.
- A mid-term Webdam Workshop at Télécom ParisTech, in March 2011.

III.7 Books

Two important activities were carried out:

[3]This informal board consists of Prof. Peter Buneman, University of Edinburgh, Prof. Stefano Ceri, Politecnico di Milano, Prof. Georg Gottlob, Oxford University, Prof. Rick Hull, IBM Watson Research Center, Prof. Tova Milo (Chair), Tel Aviv University, Prof. Moshe Vardi, Rice University.

1. S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset and P. Senellart have written a textbook entitled *Web Data Management and Distribution*, to be published by Cambridge University Press in 2011 [57]. 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 consists of graduate and PhD students, engineers and practitioners seeking an in-depth presentation of the languages, techniques and tools required to build large-scale and distributed information systems. Parts of the book served as material for various courses and events, including a summer school co-organized by Webdam in Les Houches in May 2010.

2. S. Abiteboul, R. Hull, and V. Vianu initiated a second electronic volume of Foundations of Databases. Two new parts are being considered: semistructured data (Leonid Libkin) and data integration (Alin Deutsch). Both are fully related to the activity of Webdam.