



Deliverable D2.1
First Year Report on Workpackage 2
Dissemination level: PU
Date: 11 June 2010

1 Overview of WP2

Objectives of the WP. The key objectives were:

1. Develop new techniques and algorithms for schema mapping and Web data exchange; and
2. Develop new techniques and algorithms for improving query evaluation by making use of schema information.

The first objective is represented by tasks T2.1 (Schema Mappings for XML) and T2.2 (Data Exchange for XML Documents), and the second by tasks T2.3 (Query Languages and Evaluation Techniques) and T2.4 (Processing text-centric XML).

Main results. The key achievements of the year are in the area of schema mappings and XML data exchange, and in efficient query processing. For data exchange, we (1) provided a complete classification of classes of schema mappings based on the complexity of their static analyses; (2) further classified schema mappings based on the behaviour of query answering algorithms, and (3) identified a large and practically relevant class of XML schema mappings that admits particularly efficient static analysis and query answering algorithms.

For query processing, we produced a family of algorithms for the most efficient (linear-time) evaluation of XPath. Previous algorithms either omitted key features of XPath such as value comparisons, or ran in quadratic time, which is too inefficient for large documents. The existence of linear-time algorithms for all of XPath was an open problem which has now been solved.

Dissemination. Our results have been published in [2, 1, 12, 18, 5, 20, 19, 9, 17, 15, 8, 14]; papers [4, 6] are currently under submission. Publication venue include top database and theory conferences and leading computer science and logic leading journals, as well as a keynote at DL (Description Logics 2009) [11], an invited tutorial at PODS (Principles of Database Systems 2009) [10], a keynote at RTA 2010 (Rewriting Techniques and Applications) [3], as well as talks at MEMICS 2009 (a school for PhD students in Brno) and GANDALF 2010 (a conference in Amalfi, Italy).

Collaboration. The work package involved researchers from Edinburgh, Amsterdam, Warsaw, and Paris. The work on tasks T2.1 and T2.2 was mainly done in Edinburgh, in collaboration with researchers from Warsaw and Paris. The work on task T2.3 was done in Warsaw and Paris, and involved Warsaw/Paris collaboration. The work on task T2.4 was done in Amsterdam.

Justification of resources. FoX funding has been used to support research positions in this project at different FoX sites and to support travel to conferences where WP2 research papers have been published and presented.

A total amount of 76 person-month (PM) has been assigned to WP2. This year we have devoted 20.16 PM to WP2 as detailed below. The fact that we used less than one-third of the PMs associated with WP 1 is explained by the fact that the research positions could not be filled immediately at the start of the project.

Tony Tan has been appointed by Edinburgh (UEDIN) in July 2009 to work on FoX. He has been working with Leonid Libkin on several issues related to static analyses of XML, and also started collaboration with Dortmund on projects related to WP1. Part of the FoX funding was used to support Leonid Libkin who was involved in all aspects of T2.1 and T2.2. Altogether they contributed 5.3 PM to WP2.

Paweł Parys has been appointed by Warsaw (UWAR) in July 2009. For 4 months, he has been working on T2.3. For the remaining 6 months, he has been working on WP3. Diego Figueira spent 3 months in Warsaw, 1 of them working on WP2.

Maria-Hendrike Peetz was appointed by Amsterdam (UVA) in November 2009. She has been working with Maarten Marx on Task 2.4 of WP2. Altogether they have contributed 6.2 PM on WP2.

At INRIA Serge Abiteboul, Luc Segoufin, and Cristina Sirangelo worked on T2.1 and T2.3. Altogether they contributed 3.66 PM to WP2.

2 Description of the new results

2.1 Tasks T2.1 and T2.2

Last year we saw a significant progress towards the key objective of tasks T2.1 and T2.2: *“Develop new techniques and algorithms for schema mapping and Web data exchange”*. As these tasks are closely interrelated and summarised by the same key objective, the description of the progress related to these tasks is combined in one subsection.

The main contributions to XML schema mappings and data exchange are the following papers: [2, 7] published in the premier database theory conference (ACM PODS), [1], published in another top-tier database conference (ICDT), and [12] published in a well-known international journal.

There is also a new paper [6] under submission.

The key goals of our work were:

1. Analyse the complexity of reasoning about XML schema mappings and identify tractable yet practical fragments;
2. Further analyse those fragment in terms of query answering over schema mappings they define, and
3. Develop practical algorithms for XML data exchange and query answering over exchanged data.

Significant progress has been achieved towards items 1 and 2, and some initial investigations have been done on item 3. In [2], we provided a complete classification of features available in XML schema mappings, and gave a detailed study of the complexity of static analysis problems related to them. We identified features of schema mappings that lead to problems of very high complexity; removing them, we produced a class of mappings that is not only tractable, but is also very common in practice (some empirical studies suggest that up to 70% of real life XML schemas are those used in this class).

In [1] we extended the results of [2] to query answering. We showed a certain tradeoff between between features of query languages and of mappings, further classifying cases of good complexity. The languages we looked at in [1] were rather limited as they were essentially XML-to-relational languages.

We started a new research programme whose goal is to remedy this deficiency in [7]. The idea is to extend languages previously studied in connection with data exchange to fully fledged XML query languages, and extend the standard data exchange query answering semantics, based on certain answers, to them. We did so with a rather expressive subset of XQuery, and gave a proper definition of the certain answers semantics. The key result is that it works well (giving us tractable query evaluation) with the restricted class of mappings we identified in [2, 1].

In [6] we made the first steps towards addressing the problem of the practicality of XML data exchange under the tractability restrictions. We noticed that the restrictions fit in nicely with one of the common techniques for storing XML data in relational databases. We then provided a fully relational implementation of the key tasks of XML data exchange and proved its correctness and efficiency.

To summarise, for tasks T2.1 and T2.2, we have an understanding of the features of mappings and queries that lead to tractability and thus potentially efficient implementation of XML data exchange. The key remaining open questions are: (a) providing algorithms for efficient construction of target instances in data exchange, and (b) extending the relational implementation of data exchange to more expressive real-life queries.

2.2 Task T2.3

There are three main subgoals of this task, described below.

Distributed query processing. In some on-going work, the INRIA site is investigating the underlying foundations of distributed optimisation based on mixing local query processing with exchanges of data, i.e., send and receive functions for transmitting data. The problem quickly becomes undecidable even for positive queries because of the recursion introduced by different peers exchanging information in a cyclic manner. We are investigating the frontiers of undecidability. This work is in preliminary stages, there are currently no published results.

Optimisations. In [4] (a journal version of two papers, including the paper [18] supported by FoX) we studied the complexity of XPath evaluation. The main result is an algorithm that evaluates a query Q on a document D in time $|D| \cdot poly(Q)$ when the query is in XPath 1.0, and in time $|D| \cdot exp(Q)$ when the query is in regular XPath. The key contribution is the linear time data complexity, which is important when querying large documents. Previous algorithms would either exclude key features such as attribute comparisons, or would run in quadratic time, which is unacceptable performance on large datasets.

This line of work was revisited in [5], where similar (but slightly more powerful) results were obtained using a new automaton model. The idea is to introduce an intermediate step in the evaluation algorithm, where the query is compiled into an automaton (strictly speaking, a transducer), and then the transducer is evaluated by the automaton. This approach is consistent with the “Optimisation” sub-task of 2.3, which searches for query primitives among the very numerous available XML languages. In this case, the query primitive would be an automaton. We hope that this automaton model can be used to efficiently evaluate other XML languages.

Property testing for very fast querying. The goal of this sub-task was to extend the technique of property testing, known in algorithmics, to XML languages. It is known that property testing can be done for regular word languages, we wanted to extend this to regular tree languages. After investigating the problem closely, we discovered that any nontrivial solution would touch on some open problems in algorithmics, open problems whose solutions would probably require techniques from outside our area of expertise. That is why we have suspended work on the approach to very fast querying that involves property testing. Nevertheless, we still intend to study very fast querying, understood as sublinear algorithms, in the remaining years of the project.

2.3 Task T2.4

This task is concerned with processing *text-centric* XML. Key research goals of this task are

- queries which combine constraints on the structure and on the textual content of a document;
- the uncertainty introduced by text analytics, and used to provide a good relevance ranked list of results.

Task T2.4 is a partly experimental (as usual in Information Retrieval), partly theoretical task. On the experimental side, we created a large corpus (16K files, 25M elements, 5.4 Gigabyte) of document-centric XML with a very rich and complex schema: the proceedings of the EU parliament from 1999. Together with a number of schema-mappings and test-queries this collection will be added to the FoX workbench. Also on the experimental side we created large repositories of document centric XML filled with government data and are experimenting with large scale data multi-lingual integration [17].

An important aspect of Task T2.4 is creating awareness of the benefits of rich document-centric XML by key problem holders. We directed our efforts to publishers of parliamentary data [13, 16, 15, 9].

Theoretically, we investigated conservativity requirements on XML tree patterns extended with full text search and used tree pattern minimisation to ensure consistent answer behaviour of content and structure query ranking algorithms [19].

To summarise, we have created a solid testbed needed for the experimental part of T2.4 and we created the infrastructure for integrating parliamentary information from various sources. This infrastructure will be used in the first and last two subtasks of T2.4. The work [19] sets the framework for achieving the two above stated research goals.

References

- [1] S. Amano, C. David, L. Libkin, and F. Murlak. On the tradeoff between mapping and querying power in xml data exchange. In *Proc. of Intl. Conf. on Database Theory (ICDT)*, 2010.
- [2] S. Amano, L. Libkin, and F. Murlak. XML schema mappings. In *Proc. of ACM Symp. on Principles of Database Systems (PODS)*, pages 33–42, 2009.
- [3] M. Bojanczyk. Automata for data words and data trees. In *Proc. of Rewriting Techniques and Applications (RTA)*, 2010.
- [4] M. Bojanczyk and P. Parys. Xpath evaluation in linear time. In *Submitted to JACM*.

- [5] M. Bojanczyk and P. Parys. Efficient evaluation of nondeterministic automata using factorization forests. In *Proc. of ACM Symp. on Principles of Database Systems (PODS)*, 2010.
- [6] R. Chirkova, L. Libkin, and J. Reutter. XML data exchange via relations. submitted.
- [7] C. David, L. Libkin, and F. Murlak. Certain answers for XML queries. In *Proc. of ACM Symp. on Principles of Database Systems (PODS)*, 2010.
- [8] S. Grijzenhout, V. Jijkoun, and M. Marx. Opinion mining in Dutch Hansards. In *Proceedings Workshop From Text to Political Positions (t2pp 2010)*, 2010.
- [9] R. Kaptein and M. Marx. Focused retrieval and result aggregation with political data. *Information Retrieval*, 2010.
- [10] L. Libkin. The finite model theory toolbox of a database theoretician. In *Proc. of ACM Symp. on Principles of Database Systems (PODS)*, pages 65–76, 2009.
- [11] L. Libkin and C. Sirangelo. Open and closed world assumptions in data exchange (invited talk). In *Proc. of Workshop on Description Logics*, 2009.
- [12] L. Libkin and C. Sirangelo. Data exchange and schema mappings in open and closed worlds. *JCSS*, 2010. to appear.
- [13] M. Marx. Advanced information access to parliamentary debates. *Journal of Digital Information*, 10(6), 2009. Special issue on *Information Access to Cultural Heritage*.
- [14] M. Marx. Logical foundations of XML and XQuery. In S. T. et al., editor, *Reasoning Web 2009*, number 5689 in LNCS, pages 111–157. Springer, 2009.
- [15] M. Marx, N. Aders, and A. Schuth. Digital sustainable publication of legacy parliamentary proceedings. In *Proceedings 11th International Digital Government Research Conference (dg.o 2010)*, 2010.
- [16] M. Marx and A. Nusselder. What you say is who you are. how open government data facilitates profiling politicians. In *Proceedings Open Knowledge Conference, London 2010*, volume 575 of *CEUR Workshop Proceedings*, 2010.
- [17] M. Marx and A. Schuth. DutchParl. A Corpus of Parliamentary Documents in Dutch. In *Proceedings Language Resources and Evaluation (LREC) 2010*, pages 3670–3677, 2010.

- [18] P. Parys. Xpath evaluation in linear time with polynomial combined complexity. In *Proc. of ACM Symp. on Principles of Database Systems (PODS)*, 2009.
- [19] H. Peetz and M. Marx. Tree patterns with full text search. In *Proceedings WebDB 2010*, 2010.
- [20] B. ten Cate, T. Litak, and M. Marx. Complete axiomatizations for XPath fragments. *Journal of Applied Logic*, 8:153–172, 2010.