



Deliverable D1.4
First Year Report on Work Package 1
Dissemination level: PU
Date: 11 June 2010

1 Overview

Objective of the WP. The key objectives of this work package are:

1. to develop a software library that contains tools for handling XML schemas, and which is suitable for testing the models and tools developed in the project;
2. to expand the focus of XML schema formalisms from the structure-centric to a data-oriented approach to XML and to develop models for handling data in XML documents;
3. to develop models and tools for the management of distributed repositories of XML documents and schemas stored across multiple sites; and
4. to develop models and methods for the management of XML repositories in the presence of other data models.

The four objectives correspond to the four tasks

T1.1 Developing a foundational Schema Software Library

T1.2 Reasoning about models and constraints involving data values

T1.3 Management of XML repositories

T1.4 Combining multiple data models

Main results. The key achievements of the year are in the development of the software library, reasoning about models and constraints involving data, and schema management for large, distributed documents.

The most spectacular results have been achieved in task T1.2. The understanding of the interplay between structural and data-oriented constraints has been improved considerably in a series of published research papers. The single most striking achievement is the decidability of an important fragment of XPath in the presence of value comparisons [6]. This is an important step toward the goals of task T1.2 as it tells us that, in principle, we can reason about the most frequently used part of XPath

in the presence of data values. This breakthrough result was prepared by [5] and accompanied by [3]. Other completed work that improved our understanding with respect to the goals of task T1.2 are [7, 17].

Very good results have also been produced with respect to tasks T1.3 and T1.1. With respect to work that has already been finished, the most remarkable results in task T1.3 yielded important insights into the schema aspects of distributing large documents (and their schema) over several sites were found [1, 11]. As this is a fundamental issue for schema-aware distributed XML repositories, it constitutes a big step towards the goals of the task.

In both tasks, T1.2 and T1.3, many further investigations are under way, but not yet published. In task T1.1, most of the work was concerned with the design of the Schema Software Library which is still in progress and the implementation and design of prototypical software for the basic schema library. The key achievements related to this task that have already made their way to publications consist of algorithms that allow to bring XML schemas that do not yet conform to the W3C standard automatically into a correct form [2, 8].

Dissemination. The results of this work package have been published (or accepted for publication) in the research papers [5, 6, 7, 17, 3, 1, 11]. These papers were or will be presented at top-level conferences in the respective fields (PODS, ICDT, LICS, CSL).

Furthermore, some publications and submissions from other work packages contained significant parts relevant for work package 1 [2, 8, 4].

The schemas of the European Parliament data collected for FoXLib are described in [16, 14]. The Amsterdam PI gave a tutorial on the foundations of XML at the 2009 Reasoning Web summer school [13], and gave several seminars promoting the use of XML and XML technology in the fields of cultural heritage and open governmental data (e.g. at the Dutch Parliament, Data Archiving and Networked Services (DANS) and the IISG (International Institute of Social History)) [10, 14, 9, 15, 12]. At various occasions, members of the project gave presentations on particular results of the project at other universities or for research organisations, including University of Hannover and University of Koblenz, and the Dortmund section of the German *Gesellschaft für Informatik*.

Collaboration. The work package involved researchers from all seven sites and many investigations were joint efforts involving two or more sites.

For the development of FoXLib (Task T1.1), extensive collaboration between researchers of Dortmund (TUDO) and Hasselt (UHAS) was needed. On several occasions video conferences were organised, as well as a visit by Geert Jan Bex in Dortmund for more in-depth discussions. Also, partially related to this task, Tomasz Idziaszek (Warsaw) visited UHasselt

in 2009 for three months resulting in [8]. Marx (Amsterdam) visited the TUDO group to make them familiar with the parliamentary data “use case” mentioned in WP 2.

There is an intense ongoing collaboration between Tony Tan (Edinburgh) and the group in Dortmund on the issue of data constraints (task T1.2). Tony Tan visited the TUDO group two times. Henrik Björklund, from Umeö university, visited the TUDO group for some work on task T1.2.

The initial research paper on the main achievement of task T1.3 [1] is the product of a collaboration between the groups in Oxford and Paris. The topic was afterwards taken up by the Dortmund site in [11].

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 WP1 research papers have been published and presented.

A total amount of 80 person-months (PM) has been assigned to WP1. This year we have devoted 34.92 PM to WP1 as detailed below.

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. Altogether they contributed 1.1 PM to WP1.

Matthias Niewerth has been appointed by Dortmund (TUDO) in June 2009. He has been working with Wim Martens and Thomas Schwentick on several issues related to tasks T1.2 and T1.3. Altogether they contributed 14 PM to WP1.

Diego Figueira has been appointed by INRIA in May 2009. He has been working under the supervision of Luc Segoufin to work on issues related to WP1 and WP3. During this year he contributed 6.27 PM on WP1.

In Hasselt (UHAS) Tomasz Idziaszek contributed 3 PM to task T1.1. Moreover Geert Jan Bex, Frank Neven and Wouter Gelade (UHasselt) contributed 6.25 PM to task T1.1.

In Amsterdam (UVA) Maarten Marx contributed 4 PM to the development of FoXLib.

Georg Gottlob in Oxford also contributed 0.3 PM to WP1.

2 Description of the new results

Task T1.1: Developing a foundational Schema Software Library

This task has seen very many activities along various lines. On the foundational side, the research papers [2, 8] contributed important algorithms for the management of schemas with the current W3C standards in mind.

These investigations were motivated by the fact that some of the restrictions stipulated by the XML Schema standard are very often violated in practice or make it difficult to combine and change schemas. In [2] it was investigated how schemas that violate the *Unique Particle Attribution (UPA)* constraint of XML Schema can be “repaired” automatically. In [8] the *Element Declaration Consistent (EDC)* constraint was studied and algorithms were invented that translate schemas not obeying this constraint into schemas that respect it.

The remaining research in this task was more closely related to the development of a Schema Software Library (to which we refer as *FoXLib* in the sequel).

A second line of work involved preliminary prototypical implementations of existing algorithms (most of them previously developed by FoX participants) in the context of students’ projects at the Technical University of Dortmund. As these implementation efforts are part of teaching (and we think that this strong connection between teaching and research is very useful) we usually do not expect that the results are directly usable for FoXLib. However, these implementations give valuable insights that we use for the specification and implementation of the actual library. These student projects mainly deal with prototypical implementations for the basic library mentioned in the DoW.

In one of these projects, a group of 13 Diplom-students (*Projektgruppe 530* at the TU Dortmund) designed a pattern-based schema language *BonXai* aimed to reflect most features of XML Schema in a more user friendly way and implemented a software library with operations for the translation between XML Schema, BonXai and other schema languages. Two further students are extending the Java object model of the above mentioned first step to incorporate the operations (a), (b), (c), and (e) of the basic library (DoW). They are expected to finish their work at the end of September.

Besides algorithms, FoXLib also contains a corpus of real life data reflecting the intricacies of a dynamic distributed collection of XML documents, schemas, transformations, updates and queries. The University of Amsterdam created the core of this corpus based on the proceedings of the European parliament [16, 14].

The specification of the basic library (Deliverable D1.1) is under way. However, its development has been considerably delayed by the fact that Geert Jan Bex, the initially designated FoX Software Quality Manager, has left FoX after six months. His responsibilities have been subsequently assumed by Frank Neven, UHasselt, and some of his intended activities will be carried out at TUDO. These rearrangements and the information transfer from Geert Jan Bex to the new responsible persons required a considerable effort and resulted in the aforementioned delay of this specification.

Even though the specification is not *finished* yet, some of the key parts

have already been specified by Geert Jan Bex during the first six months and further work has been done since then. Some key parts of the library have even already been implemented and tested which makes us confident that we will be able to catch up with the schedule within the next two years. Deliverable D1.1 will be finished within the next few months.

The specification will be disseminated to the other members of the project during the second training event in September.

Task T1.2: Reasoning about models and constraints involving data values

The work in this task splits up into two parts: the development of a rich modelling language and the investigation of decidable or even efficiently tractable sub-languages. Some of the results obtained in this task contribute to both parts. They propose language features that may become interesting for the full language and, at the same time, they study the complexity of reasoning tasks for these models. The results serve as a basis for the design of useful fragments.

An important step towards the full language is documented in the research paper [3]. It proposes a new kind of automata that capture an extension of XPath (*Regular XPath*) and it demonstrates that they are very suitable for showing decidability over documents obeying certain restrictions concerning data. Furthermore, and most important for T1.2, this formalism allows to uniformly classify different other formalisms studied before and therefore sets the basis for a better understanding of similarities/differences between different models.

Another important aspect of the full language is being studied in Edinburgh. A very interesting way of specifying constraints on data values by *set constraints* is proposed. These constraints are closely related to linear numerical constraints and thus immediately yield algorithms for reasoning tasks.

However, the key achievement of the task is constituted by the breakthrough results on satisfiability of XPath. As XPath is used as a sublanguage in all standard XML languages (XML Schema, XQuery, XSLT) it is very important to understand its properties with respect to reasoning tasks. It was previously known that its satisfiability problem is undecidable in general. However, in two successive papers [5, 6] it turned out that for the most relevant part of XPath (where navigation is in forward direction with respect to the document only) satisfiability can actually be decided.

Further achievements of this task mainly contribute to the study of sub-languages and include the following.

- The main result of [7] shows that very simple fragments of linear temporal logics for XML and for data words that can only navigate with transitive operators are undecidable.

- An important logic in the context of XML reasoning is two-variable predicate logic. The reasoning on several variants of this logic, based on the allowed ways in which formulas refer to the relation between positions in a tree and the relationships between data values, had already been studied. In [17] it is shown that automated reasoning is possible even if one allows to compare data values with respect to the linear order (in the setting of data strings when positions can be compared by their relative order, too).¹

Currently, many further investigations are going on, some of them are almost completed. We want to

- continue our general investigations on XPath and XML Schema,
- lower the complexities by exhibiting new fragments and models, and
- extend the expressive power (as, e.g., the current formalisms can deal with non unary key constraints)

Task T1.3: Management of XML repositories

The goal of this task is to develop tools and models as a basis for a schema-aware XML Repository management system (XRMS). An initial study revealed that, even though there are many, commercial or prototypical, systems that deal with XML documents, there is no established notion of an XRMS, let alone of a schema-aware XRMS. This is in sharp contrast to the world of relational databases where the notion of database management systems (DBMS) is essential and quite standardised. To this end, at TUDO, we started a Ph.D. project (carried out by an external student, Thomas Timm) that aims at the development of a concise notion of an XRMS with a particular emphasis on schema-awareness. It shall specify a complete XRMS and implement key parts of it in a prototypical manner. We hope to integrate most of these parts into FoXLib. This work involves a broad study of the literature, existing systems and the thorough investigation of use cases. The Ph.D. project is co-supervised by a colleague from TUDO, Prof. Jannach, to guarantee high-quality results from a Software Engineering point of view.

A particular operation that a schema-aware XRMS has to support is the combination of schemas (as already mentioned in the DoW). The aforementioned research paper [8] develops useful algorithms that produce a new schema obeying the aforementioned EDC constraint when two schemas are combined (e.g., by union or intersection).

The investigations mentioned so far belong to the first phase of task 1.3, in that they study single-site repositories. The second phase is devoted to the investigation of distributed repositories and has already yielded

¹This paper is a joint achievement with a project funded by the German DFG under grant SCHW 678/4-1.

important results during the first year of the project ([1], mentioned below, was even already triggered by the proposal and was submitted before the actual funding started). The main achievement so far is the study of the implications to schema management from documents that are distributed over several sites of an XML repository. The goal is to partition the schema in a way that allows the local sites to typecheck their part of the document (to guarantee that the entire document meets its specification) and to be as nonrestrictive as possible.

In the landmark paper [1], the setting is introduced, algorithms that compute designs are developed and the complexities of the underlying reasoning tasks are pinpointed. The research paper [11] gives improved algorithms for some of the cases studied in [1] and studies a setting that conforms to the W3C standard more closely in that it considers schemas respecting the UPA constraint. It resolves most of the complexities in that setting and shows that the particularly attractive case of *perfect typings* has a tractable solution.

Task T1.4: Combining multiple data models

Task T1.4 will be mainly carried out in the second half of the project. During the first six months, the presence of multiple data models was considered an important aspect in the initial study on existing XML management systems mentioned in T1.3.

Furthermore, [4] achieves tractable XML Data Exchange by using a relational database system. The paper reveals important insights that will be useful for the investigation of data exchange across multiple data models.

References

- [1] S. Abiteboul, G. Gottlob, and M. Manna. Distributed XML design. In *Proceedings of the Twenty-Eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2009*, pages 247–258, 2009.
- [2] G. J. Bex, W. Gelade, W. Martens, and F. Neven. Simplifying XML schema: Effortless handling of nondeterministic regular expressions. In *ACM SIGMOD International Conference on Management of Data, SIGMOD 2009*, pages 731–744, 2009.
- [3] M. Bojańczyk and S. Lasota. An extension of data automata that captures XPath. In *Proceedings of the 25th Annual IEEE Symposium on Logic in Computer Science, LICS 2010*, 2010.
- [4] R. Chirkova, L. Libkin, and J. Reutter. XML data exchange via relations. submitted.

- [5] D. Figueira. Satisfiability of downward XPath with data equality tests. In *Proceedings of the Twenty-Eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2009*, pages 197–206, 2009.
- [6] D. Figueira. Forward-XPath and extended register automata on data-trees. In *Proc of Int Conference on Database Theory (ICDT)*, 2010.
- [7] D. Figueira and L. Segoufin. Future-looking logics on data words and trees. In *Mathematical Foundations of Computer Science (MFCS) 2009*, pages 331–343, 2009.
- [8] W. Gelade, T. Idziaszek, W. Martens, and F. Neven. Simplifying XML Schema: Single-type approximations of regular tree languages. to be presented at ACM Symposium on Principles on Database Systems (PODS 2010), 2010.
- [9] S. Grijzenhout, V. Jijkoun, and M. Marx. Opinion mining in Dutch Hansards. In *Proceedings Workshop From Text to Political Positions (t2pp 2010)*, 2010.
- [10] R. Kaptein and M. Marx. Focused retrieval and result aggregation with political data. *Information Retrieval*, 2010.
- [11] W. Martens, M. Niewerth, and T. Schwentick. Schema design for xml repositories: Complexity and tractability. In *Proceedings of the 29th Symposium on Principles of Database Systems (PODS)*, 2010.
- [12] M. Marx. Advanced information acces to parliamentary debates. *Journal of Digital Information*, 10(6), 2009. Special issue on *Information Access to Cultural Heritage*.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] T. Schwentick and T. Zeume. Two-variable logic with two order relations. In *Conference for Computer Science Logic (CSL)*, 2010.