NWO **Vrije Competitie 2009**
Exacte Wetenschappen

1 Project Details

1.a) Project Title. Data Exchange for Document Centric XML

1.b) Project Acronym. DEX

1.c) Principal Investigator. Dr. Maarten Marx
Intelligent Systems Lab Amsterdam (ISLA)
University of Amsterdam
E-mail: maartenmarx@uva.nl
URL: http://ilps.science.uva.nl/

2 Summary and Dutch language abstract

2.a) Summary. This project is about a topic in database theory, data exchange. Data exchange is about the problem of taking data structured under a source schema and materializing an instance of a target schema that reflects as accurately as possible the source data. The logical foundations of this problem in the case of relational data have been formulated in 2003 and applied in IBM's data exchange product Clio. In later years this theory has been extended to nested relational and data-centric XML. A key feature of this type of data is its invariance for sibling order. Much XML data that is exchanged in practice however is not invariant for sibling order and the existing theory and tools are not applicable.

The current project thus aims to extend the theory of data exchange to this kind of XML data, called *document centric* XML. To specify common mappings in documents centric XML, the formalism of tree patterns is not sufficient and needs to be extended with conditional axis steps. These are a syntactic variant of the until operator from temporal logic. For this expansion of the mapping rule formalism we study the main topics in data exchange: expressivity, minimization and definability; static analysis problems like schema mapping containment and consistency of mappings; computation of certain answers.

2.c) Keywords. XML, Data Exchange, Schema Mapping, Textual Databases, Database Applications, Logic.

3 Classification

The project falls within the discipline of *Computer Science*. Relevant research themes from the *Nationale Onderzoeksagenda Informatie- en Communicatietechnologie 2005-2010 (NOAG-ict)* are:

3.2 Data Explosie. ICT-disciplines: Algorithms and Computation Theory; Hypermedia, Hypertext and Web; Management of Data.

4 Composition of the Research Team

Name	Title	Role	Expertise	Affiliation
Marx, M.	Dr.	Applicant	XML, Logic, Knowledge Representation	ISLA, IvI, U. Amsterdam
ten Cate, B.	Dr.	Adviser	Data Exchange, XML, Logic	U. Santa Cruz, USA
NN		PhD Student	Databases, Theoretical CS	ISLA, IvI, U. Amsterdam (to be funded by NWO)

The project will be directed by Marx, who will act as the daily supervisor for the PhD student. Prof. Dr. Maarten de Rijke (ISLA, IvI, UvA) will be the formal promotor of the PhD student. Dr Balder ten Cate (formerly at UvA, and IBM-Almaden, USA) will act as liaison with the IBM-Almaden group.

5 Research School

The research will be embedded in the Dutch Research School for Information and Knowledge Systems (SIKS).

6a Description of the proposed research

6.1 Scientific research question and the intended research results

6.1.1 Background to data exchange

Data exchange has been defined as the problem of taking data structured under a source schema and materializing an instance of a target schema that reflects as accurately as possible the source data [17]. In recent years, data exchange applications have become more and more widespread, particularly due to the proliferation of web data in various formats (relational, XML, RDF, etc) and the emergence of e-business applications that need to communicate data yet remain autonomous. Vertical search engines and mashup sites on the web are a key application area for data exchange: data from various sources is collected and transformed into a common target schema [20]. The surplus value of a mashup derives from the established links (joins) between the previously disparate data-sources [13].

The importance of data exchange in practice is witnessed by the involvement of industry: IBM developed the Clio system [19]; Microsoft focused on (meta)-model management [14, 15].

Even though data exchange is an old and common data management problem, its most foundational aspects had not been studied until recently. There are two reasons for this. First, most of the early research on databases concentrated on the stand-alone relational model, and much less on interoperability, integration, and exchange. Second, there was no solid foundation, nor even a proper formal model, for the problem of data exchange.

Such a model was finally proposed in 2003 by Fagin, Kolaitis, Miller and Popa [17], and was quickly adopted as the appropriate model for data exchange. Data exchange and the closely related area of data integration are now well-established research disciplines in the database community.

Even though the advent of XML was a driving force behind the renewed interest in data integration and exchange [20], the theory of data exchange for XML still has several open problems whose counterparts in the relational case have been solved successfully. Arenas and Libkin [11] transformed the relational framework of Fagin et al to *data-centric* XML and showed the first important results, which are followed up by a number of further papers from them and co-workers [8, 9, 16].

With this proposal we aim to extend this theory to *document-centric* XML. We now explain the difference between the two.

XML has two main use-cases: as a markup language for documents (derived from its SGML heritage), and as a language for exchanging data [33]. A MySQL dump and an RSS feed are examples of the latter, the works of Shakespeare in XML is an example of the former. XML data created for these different use-cases also has different characteristics. *Data-centric XML* closely resembles a nested relational database: document order is irrelevant; content is stored in attribute-value pairs, often uniquely determined by their ancestor path to the root; the data without the markup is useless.

In documents marked up with XML (aptly called *document-centric XML*) the textual content is enriched with markup, but still valuable without it. A crucial difference with data-centric XML is that information is encoded in the (document) order of the XML elements. This has two consequences: (1) preserving order of elements in transformations is a needed feature, and (2) the mapping rule language is not expressive enough (see Figure 1).

The reliance on the ordering in document-centric XML causes that the XML structure in them is often flat, resembling more a time line than a tree. The field of model checking and verification has shown that temporal logic is a robust and expressive formalism for specifying properties of time lines. Temporal logic and XPath are closely connected [1]. The temporal expression *Until A is true, B holds* is needed for very common transformations, see the example in Figure 1. But it is known that the tree patterns used in current XML mapping languages cannot express this construct [12, 22].

We thus see a large application area which is not covered by the existing XML data exchange theory. This situation provides a clear research plan: extend the expressive power of mappings with temporal patterns and study their logical and algorithmic properties within the data exchange framework.

6.1.2 Scientific research question and intended results

Our scientific aim is to extend the theory of data exchange to document-centric XML.
We operationalize document-centric XML as non-recursive XML files which need positive temporal patterns for specifying mappings. We focus on mappings which preserve the order of the data.

The established body of literature on data exchange suggests four areas of research with each a number of important open problems.

R1 Mapping rule language We extend the tree patterns used in mapping rules with conditional transitive closure over child, next sibling and document order axis [1], creating *conditional tree patterns (CTP)*. The following results are then desired:

- **Expressivity** Characterise the expressive power of conditional tree patterns as in [12]. A desirable result would be that they capture exactly the first order part of positive transitive closure logic [26].
- **Minimization** Tree patterns have the important minimization property [10]. Does that transfer to CTP? Does minimization remain a PTIME complete problem?
- **Typology of mappings** Relational schema mappings are naturally partitioned using syntactic criteria into LAV, GAV and GLAV type of rules. [4] provided semantic characterizations for these three classes. We intend to obtain similar characterizations for XML mapping rules created from tree patterns and conditional tree patterns.
- **Definability** An important theoretical yardstick for the quality of a logical system is whether its syntax and semantics are balanced. Beth's definability property or Craig's interpolation are often used as criteria. For data exchange systems, a closely related definability question is whether all queries which are determined by a set of views can also be defined (in query language L) using these views. For the relational case, Segoufin and Vianu showed that this property holds in just a few cases [24]. The property holds for views expressed in the semi-join algebra or the packed fragment [2]. We intend to develop a similar definability theory for XML data exchange. The fact that conditional tree patterns are inside the packed fragment gives good hope for positive results.

R2 Model Management Data exchange occurs often in dynamic environments in which schemas evolve and new data sources (with new schemas) come into existence. This makes maintaining and managing a collection of mappings a complex task in which the following two computational problems are important (both are solved for relational mappings in [4]). We intend to develop algorithms for the corresponding problems in XML data exchange.

- **Schema mapping containment** What is the complexity of deciding that a mapping rule logically follows from a set of mapping rules? For st-tgd's the problem is known to be NP complete.
- **Schema mapping translation** studies the complexity of deciding whether an X-type schema mapping can also be expressed using Y-type schema mappings. We like to have optimal algorithms which in case of a positive answer also construct the desired mappings.

R3 Static analysis is about tools which help developers of schema mappings. In the large space of possible mappings we want to find syntactic conditions on the inputs which guarantee tractable solutions for questions about the existence of target solutions (consistency and absolute consistency of mappings). For tree patterns and data centric XML, much is known from the work of Libkin et al [8, 11]. We intend to extend their results to conditional tree patterns. Document centric XML transformations often have a unique desired target. Thus we also intend to map the complexity of this problem: given source and target schema's S and T, and a mapping Σ, decide whether for each source t valid w.r.t S there exists a *unique* target valid w.r.t. T which is a solution for t under Σ.

R4 Complexity of computing certain answers We intend to find combinations of mappings and query languages for which the problem of computing the certain answers has polynomial data complexity. Results in [9] indicate that this is hard but possible in restricted cases.

The **scoping exchange pattern** has a flat source file in which all elements below the root are siblings. It maps this into a hierarchical target in which the scope of the arguments is made explicit by nesting. A common example consists of XHTML files in which headings and paragraph elements are all represented as sibling elements. The **scope** of the heading elements `H1, H2, ...` is made explicit in the nesting of the target file.
Scoping transformation rules have the following logical form: for elements labelled `Hi` collect next-sibling elements to be nested **until** you see an element labelled `Hi`. Rules of this form cannot be expressed using mapping rules based on tree patterns [11], not even by nested mappings.

Figure 1: The scoping data exchange pattern.

6.2 Research approach and methodology

The intended results of the project are mainly theoretical and we intend to disseminate our results using the PODS and ICDT conferences on foundations of database systems.

We ground the project in actual document-centric XML data exchange practice with the following instruments. We use the format of the mapping scenarios from STBenchmark, a benchmark of mapping systems [7], to describe common document-centric XML transformations, like the one in Figure 1. We collect these scenarios from the XSLT literature, from the experiences in the PoliticalMashup project at the UvA, and from our network of researchers building vertical search engines. This family of use cases will direct us in the vast landscape of possible data exchange settings towards those areas with applicational interest.

The research methodology for our theoretical questions follows the proven path for XML research: combine tree automata theory with tools from modal and temporal logic with the traditional tools of the database theoretician, finite model theory and complexity theory.

More specifically, results on expressivity will be based on semantic characterization theorems in terms of simulations and tree automata [4, 12, 31]. Work on minimization is based on [10] and the improved algorithm of [23]. The typology of mappings will follow the syntactic categories within the source-to-target tuple generating dependencies and their semantic characterization will be based on closure properties related to those in [4].

[2] established the connection between Beth definability and rewriting queries determined by views [24]. [25] describes an implemented rewriting algorithm based on computing Craig interpolants from proofs using Smullyan's Tableaux.

For static analysis, we follow the methodology developed by Arenas and Libkin and co-workers [8, 11]. Especially useful for us are lower complexity bounds and results on dangerous features of mapping patterns and target schemas.

Computing certain answers is done by materializing a universal model and processing the query on it [17]. This model is created by the chase procedure. For document-centric XML transformations which must preserve the order of the leaves, we will develop an order-preserving chase. Recent work [9, 16] contains rather negative results for query languages which return complex data structures like trees. Fortunately existing well studied query languages for document-centric XML like NEXI are simpler and just return sequences of XML-nodes [32].

6.3 Scientific importance and urgency of the research proposed

In the last "Claremont report on Database Research" [6], which aims to set the research topics of the next five years, the *interplay of structured and unstructured data* was one of the five key future research themes. Within this theme, web data, incompleteness, dynamicity and the desire to link data were all mentioned as important challenges.

The importance and impact of theoretical work on data exchange is shown by a remarkable fact: This year a purely theoretical result —the semantic characterization of a number of relational mapping rules— was invited from ICDT to the Communications of the ACM [4].

Research in XML data exchange is more difficult than in the relational case because of the far greater range of parameter settings for which the problem can and should be studied. For data-centric XML, the level of knowledge approaches that of the relational case. This is mainly achieved by considering settings with much structure, closely resembling the relational case.

But most real world XML data exchange concerns the much less structured document-centric XML, for instance a news-aggregator or a website harvesting and mediating product reviews. Such systems now have to code their transformations directly in XQuery or XSLT or even in an all-purpose programming language. Creating such mappings is a laborious and error-prone process and this drove the development of declarative relational mapping formalisms like LAV and GAV which are now built into commercial systems like Clio [20].

The proposed research is thus of scientific importance because it is needed to understand data exchange in a truly semi-structured setting. It is urgent because in practice XML data is still transformed using mappings hidden inside programming code, a technique that does not scale.

6.4 Relationship of the proposed research with comparable research being conducted elsewhere

Theoretical research on XML data exchange is carried out in a number of groups: the most important are the IBM-Almaden/U. Toronto group with the Clio system, the group around Lenzerini in Italy, Libkin's theoretical database group in Edinburgh, the group of Gottlob and Benedikt in Oxford, and the group of Arenas in Chile. ISLA contributed with innovative foundational characterization results [2, 4].

ISLA has good working connections with the IBM-Almaden group and participates in the EU-funded FP7 project *Foundations of XML (FoX)* in which also the groups of Libkin and Gottlob participate. Data exchange is an important topic within FoX.

The focus on document-centric XML is unique within the theoretical database community.

6.5 How the proposed research ties in with the current research of the group in which the project member to be attracted shall work

The PhD student will work within the subgroup of ISLA specialized in semi-structured data led by Maarten Marx. The former group member, Balder ten Cate is an adviser to the project. Marx and ten Cate have been working on foundational problems for XML since 2004 [1, 2, 3, 27, 28, 29, 30, 31], and work on data exchange and data integration since 2007 [2, 4].

Marx participates with one PhD student in the EU-funded FP7 project "Foundations of XML (FoX)" in which data exchange, dynamicity, and incomplete information are leading research themes. The Phd student of this project will be able to participate in the training and events organised by the FoX-consortium.

ISLA has much practical experience with data exchange for XML. Much of ISLA's research can be tagged as '*Information Retrieval (IR) meets Language Technology meets Semi-Structured Data*'. ISLA participates since its beginning in INEX, the evaluation forum for IR-research on document-centric XML-data and co-developed the important NEXI query language [21]. Answering NEXI queries over millions of XML documents coming from multiple sources is a task which requires massive data exchange.

Data exchange also plays a prominent part in the PoliticalMashup project of Marx [18].

6b Application Perspective

Section 6.3 argued why data exchange for document-centric XML needs its own theory and tools based on that theory: in summary, developing and managing sets of mappings written in non-declarative programming languages does not scale.

For relational and data-centric XML data, a data exchange tool exists, Clio, which comes with IBM's DB2 system. Many of our intended results extend those of [17], the paper which formalized the theory behind Clio. We thus expect that our results can smoothly be applied in Clio and so extend the applicability of Clio's tools from relational and data centric XML to document centric XML.

Through regular coordination with the IBM-Almaden group (the developers of Clio), and the grounding of our work in practical use cases, we ensure that our theoretical work has applicational value. We transfer our developed knowledge via existing connections with IBM-Almaden. We strengthen this connection by a research visit of the prospective PhD student.

7 Literature

Five most important publications by the research team with respect to the proposal

[1] M. Marx. Conditional XPath. *ACM Transactions on Database Systems (TODS)*, 30(4):929–959, 2005.

[2] M. Marx. Queries determined by views: pack your views. In *Proceedings PODS '07*, pages 23–30, New York, NY, USA, 2007. ACM.

[3] M. Marx and M. de Rijke. Semantic Characterizations of Navigational XPath. *ACM SIGMOD Record*, 34(2):41–46, 2005.

[4] B. ten Cate and P. Kolaitis. Structural characterizations of schema-mapping languages. In *Proc. ICDT*, pages 63–72, 2009. Also in *Communications of the ACM*, Vol 53,1 2010 (101–110).

[5] B. ten Cate, W. Conradi, M. Marx, and Y. Venema. Definitorially complete description logics. In *Proceedings KR 2006*, pages 79–89, 2006.

Other references

[6] R. Agrawal, A. Ailamaki, P. A. Bernstein, E. A. Brewer, M. J. Carey, S. Chaudhuri, A. Doan, D. Florescu, M. J. Franklin, H. Garcia-Molina, J. Gehrke, L. Gruenwald, L. M. Haas, A. Y. Halevy, J. M. Hellerstein, Y. E. Ioannidis, H. F. Korth, D. Kossmann, S. Madden, R. Magoulas, B. C. Ooi, T. O'Reilly, R. Ramakrishnan, S. Sarawagi, M. Stonebraker, A. S. Szalay, and G. Weikum. The Claremont report on database research. *Commun. ACM*, 52(6):56–65, 2009.

[7] B. Alexe, W. Tan, and Y. Velegrakis. Stbenchmark: towards a benchmark for mapping systems. *Proc. VLDB Endow.*, 1(1):230–244, 2008. ISSN 2150-8097.

[8] S. Amano, L. Libkin, and F. Murlak. XML schema mappings. In *Proc. PODS '09*, pages 33–42, 2009.

[9] S. Amano, C. David, L. Libkin, and F. Murlak. On the tradeoff between mapping and querying power in xml data exchange. In *Proceedings ICDT '10*, pages 155–164, New York, NY, USA, 2010. ACM.

[10] S. Amer-Yahia, S. Cho, L. Lakshmanan, and D. Srivastava. Tree pattern query minimization. *The VLDB Journal*, 11:315–331, 2002. ISSN 1066-8888. URL `http://dx.doi.org/10.1007/s00778-002-0076-7`.

[11] M. Arenas and L. Libkin. XML data exchange: Consistency and query answering. *J. ACM*, 55(2): 1–72, 2008. ISSN 0004-5411. Expanded version of abstract in Proc. PODS 2005.

[12] M. Benedikt, W. Fan, and G. Kuper. Structural properties of XPath fragments. *Theoretical Computer Science*, 336(1):3–31, 2005.

[13] T. Berners-Lee. Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006.

[14] P. Bernstein. Applying model management to classical meta-data problems. In *Proceedings Conference on Innovative Data Systems Research (CIDR'03)*, pages 209–220, 2003.

[15] P. Bernstein and S. Melnik. Model management 2.0: manipulating richer mappings. In *Proceedings SIGMOD '07*, pages 1–12, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-686-8.

[16] C. David, L. Libkin, and F. Murlak. Certain answers for xml queries. In *Proceedings PODS '10*, pages 191–202, New York, NY, USA, 2010. ACM.

[17] R. Fagin, P. Kolaitis, R. Miller, and L. Popa. Data exchange: Semantics and query answering. *Theoretical Computer Science*, 336(1):89–124, 2005.

[18] T. Gielissen and M. Marx. Exemelification of parliamentary debates. In *Proceedings of the 9th Dutch-Belgian Information Retrieval Workshop (DIR 2009), Twente, The Netherlands*, pages 19–25, 2009.

[19] L. Haas, M. Hernández, H. Ho, L. Popa, and M. Roth. Clio grows up: from research prototype to industrial tool. In *Proceedings SIGMOD '05*, pages 805–810, 2005. ISBN 1-59593-060-4.

[20] A. Halevy, A. Rajaraman, and J. Ordille. Data integration: The teenage years. In *Proceedings VLDB '06*, pages 9–16, 2006.

[21] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Articulating information needs in XML query languages. *ACM Trans. Inf. Syst.*, 24(4):407–436, 2006.

[22] M. Marx and M. de Rijke. Semantic characterizations of XPath. In *TDM'04 workshop on XML Databases and Information Retrieval.*, Twente, The Netherlands, June 21, 2004.

[23] P. Ramanan. Efficient algorithms for minimizing tree pattern queries. In *SIGMOD '02: Proceedings of the 2002 ACM SIGMOD international c onference on Management of data*, pages 299–309, New York, NY, USA, 2002. ACM.

[24] L. Segoufin and V. Vianu. Views and queries: determinacy and rewriting. In *Proceedings PODS 05*, pages 49–60, 2005.

[25] I. Seylan, E. Franconi, and J. de Bruijn. Effective query rewriting with ontologies over dboxes. In *Proceedings IJCAI 09*, pages 923–925, 2009.

[26] B. ten Cate. The expressivity of XPath with transitive closure. Available at `http://staff.science.uva.nl/~bcate/papers/regxpath.pdf`, 2005.

[27] B. ten Cate. The expressivity of XPath with transitive closure. In *Proceedings PODS*, pages 328–337, 2006.

[28] B. ten Cate and C. Lutz. The complexity of query containment in expressive fragments of XPath 2.0. In *Proceedings PODS'07*, 2007.

[29] B. ten Cate and M. Marx. Axiomatizing the logical core of XPath 2.0. *Theory of Computing Systems*, 44(4):561–589, 2009.

[30] B. ten Cate and M. Marx. Navigational xpath: calculus and algebra. *SIGMOD Record*, 36(2):19–26, 2007.

[31] B. ten Cate and L. Segoufin. XPath, transitive closure logic, and nested tree walking automata. In *Proceedings PODS*, pages 251–260, 2008.

[32] A. Trotman and B. Sigurbjörnsson. Narrowed Extended XPath I (NEXI). In *Advances in XML Information Retrieval*, pages 16–40, 2005.

[33] World-Wide Web Consortium. Extensible Markup Language (XML) 1.0 (second edition) W3C. `http://www.w3.org/TR/REC-xml`.

8 Budget

We request additional travel budget for an extended internship of the student to a leading DB industrial research center: IBM-Almaden. The visit to IBM Almaden is planned to ground our theoretical work in practice and to help the application of our developed theory in IBM's Clio product.

(1) PhD student, postdoc, and/or other personnel		
a) appointment of research personnel		= PhD
b) additional travel budget	Visit IBM Almaden USA (3 months)	= € 5.700
c) project related equipment/software		= –
d) other related activities		= –
Total b,c,d		€ 5.700