

Foundations of distributed data management

Edited by

Serge Abiteboul¹, Alin Deutsch², Thomas Schwentick³, and
Luc Segoufin⁴

1 INRIA - Orsay Cedex, FR, serge.abiteboul@inria.fr

2 University of California – San Diego, US, deutsch@cs.ucsd.edu

3 TU Dortmund, DE, thomas.schwentick@udo.edu

4 ENS – Cachan, FR, luc.segoufin@inria.fr

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 11421 “Foundations of distributed data management”.

Seminar 16.–21. October, 2011 – www.dagstuhl.de/11421

1998 ACM Subject Classification C.2.4 Distributed Systems, H.2 Database Management, H.3.5 Online Information Services

Keywords and phrases XML Query language, Distribution, Incompleteness

Digital Object Identifier 10.4230/DagRep.1.10.37

Edited in cooperation with Tom Ameloot


1 Executive Summary

Serge Abiteboul

Alin Deutsch

Thomas Schwentick

Luc Segoufin

License  Creative Commons BY-NC-ND 3.0 Unported license
© Serge Abiteboul, Alin Deutsch, Thomas Schwentick, and Luc Segoufin

Description of the Seminar’s Topic

The Web has brought fundamentally new challenges to data management. Web data management differs from traditional database management in a number of ways. First, Web data differ in their structure: trees with links (usually described by mark-up languages such as XML) instead of tables. Also, Web data are by nature distributed, often on a large number of autonomous servers. Finally, Web data are typically very dynamic and imprecise.

Unlike for the classical relational database model, there is still no commonly accepted model for data management over the Web. The lack of a clean, simple, mathematical model further prevents us from designing general solutions to typical data management problems, such as building indexes, optimizing queries, and guaranteeing certain properties of applications.

As witnessed by the two seminars that previously occurred in Dagstuhl on this topic (Seminar 01361 in 2001 and Seminar 05061 in 2005, both entitled “Foundations of Semistructured Data”), most of the recent research efforts have concentrated on adapting traditional database techniques to the XML setting. In particular, foundational research on XML focused on the tree structure of XML documents, applying well-developed techniques based on logic and



Except where otherwise noted, content of this report is licensed
under a Creative Commons BY-NC-ND 3.0 Unported license

Foundations of distributed data management, *Dagstuhl Reports*, Vol. 1, Issue 10, pp. 37–57

Editors: Serge Abiteboul, Alin Deutsch, Thomas Schwentick, and Luc Segoufin



DAGSTUHL
REPORTS

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

automata for trees. These lines of research have been very successful. However, they do not address all the facets of Web data. In particular distribution, dynamicity, incompleteness and reliability had received limited attention in past work, but play a central role in a Web setting. The aim of Seminar 11421 was to bring together researchers covering this spectrum of relevant areas, to report on recent progress in terms of both results as well as new, relevant research questions. It was organized at the initiative of members of the EU funded research projects FoX (fox7.eu) and Webdam (webdam.inria.fr) that are acknowledged for their support.

The seminar focused on the following key aspects of Web data management.

Semistructured data/query languages, with particular emphasis on XML/XPath and RDF/SPARQL. Semistructured data is the preferred way to organize information on the Web, and de facto and de jure standards are emerging. The study of XML data management and XML query languages remains a constant in the entire line of seminars culminating with 11421. Four notable additions over the predecessor seminars deserve mentioning. First is the emphasis on XML with Data values (and in its simplified version, on Data Words), where the data labeling of an XML tree (a word) is drawn from an infinite domain. Query evaluation, and static analysis tasks in this setting are considerably harder, often undecidable, and finding the right limitations of the logics used for querying and of the data models is still an open research challenge. Second is the RDF data model, with its associated query language SPARQL. They are used for modeling/querying semantic Web data (ontologies), but also classical semistructured data. Since the standardization process is still ongoing, the work performed by researchers in our extended community has significant potential for impact. Indeed, one of the seminar participants, Marcelo Arenas, sits on the standard working group and is a leader in studying the semantics, query evaluation complexity, as well as optimization potential of the SPARQL language. Third is the emphasis on static typing. This is applied to XML data, schema inference, and the experimental evaluation of large collections of XML schemas found in real life (the work conducted in the framework of the above mentioned project FoX is relevant). In addition, the new SPARQL language is in need of foundational contributions towards type inference. Fourth is the work on related languages which, while not being under consideration as official standards, have established themselves as quasi-standards for querying graph databases in the database theory community. These include variations on the language of regular path queries, in which reachability queries in the graph are expressed using various classes of regular expressions over the alphabet of edge labels.

Incomplete and Probabilistic Databases. Information found on the Web is often incomplete, or uncertain due to contradictory facts across distinct data sources. Blindly applying classical query evaluation techniques to such databases leads to inconsistent answers. In the past, the database community has proposed a revolutionary way to view such information, namely as a set of possible databases, sometimes with an associated probability distribution. Query evaluation becomes a more refined task, in which query results are classified as *possible*, i.e. they belong to the answer over some possible database, or *certain*, i.e. they belong to the query answer over all possible databases. When the set of possible databases is accompanied by a probability distribution, the likelihood of possible answers can be derived. Not surprisingly, query evaluation in this setting is harder than in the standard relational setting, and work on finding the trade-offs between evaluation complexity and query language expressivity is always challenging. For Web data management, with its in-flux design for the data models and query languages, answering these questions is particularly timely.

Data Exchange is concerned with the (materialized or virtual) migration of data between data sources. Since in a Web setting such data sources are likely autonomous and have

distinct schemas even when modeling similar real-life concepts, it is proposed to specify declaratively how data from the source database relates to the data published into the target database. These specifications are known as *schema mappings*, and they are exploited for various tasks, ranging from actually migrating the data from source to target, to leaving the data at the source but migrating queries from the target schema to the source schema. Seminar 11421 gave particular attention to the case of data exchange for XML data (prior work confines itself mostly to relational data sources), and for incomplete data (prior work focuses solely on complete data). It also addressed the problem of inferring schema mappings from examples given by less sophisticated users, who simply associate source/target data pairs and expect a tool to automatically generate the mappings. The seminar included a tutorial by Phokion Kolaitis, co-founder of the data exchange field.

Distribution of data across sources (typically within peer-to-peer networks), as well as of the computation performed by queries, and more generally, processes on top of such data, is another prominent topic of Web data management. The seminar explored recent answers to the long-standing challenge of coming up with models of computation that enable expressive languages that are semantically clean, efficiently executable and nevertheless admit automatic optimization. The above mentioned, highly visible, European research project Webdam proposes a vision inspired by the quintessentially declarative Datalog language from classical relational database research. A notable related approach is motivated by the area of declarative networking, which has gained the attention of the systems community in past years, and more recently of the theory community, which is now carrying out foundational research to complement and enhance the existing systems contributions. Such models as the relational transducer networks are being proposed to formalize famous (but so far informally stated) conjectures about expressivity and evaluation complexity of declarative networking programs. The seminar was also interested in general questions on Peer-to-Peer networks.

Static verification of temporal properties is key to increasing the reliability and facilitating the design of various classes of processes powered by an underlying (collection of) databases. Notable examples include electronic commerce Web sites, declarative networking programs, and general business processes. In all these cases, the underlying data is dynamic, its evolution in time governed by large collections of declarative rules, whose interference with each other and global effect are impossible to predict without automatic verification tools. Of particular interest is the verification of properties pertaining to the temporal evolution of the system, which are naturally expressed in various temporal logic flavors.

Crowdsourcing is another highly relevant recent development in the Web data management arena, one in which practice has pressed ahead of foundational work, which is now attracting the interest of the theory community. The seminar dedicated particular attention to this topic, reserving a long talk slot for a survey.

Organization of the Seminar and Activities

The workshop brought together 51 researchers from complementary areas of database theory, logic, and theoretical computer science in general, all with an established record of excellence in Web data management. The participant pool comprised both senior and junior researchers, including several advanced PhD students.

Participants were invited to present their own work, and/or survey state-of-the-art advances and challenges in the field. Thirty-four talks were given, which included four (60-90 minute) tutorials and thirty regular (30 minute) talks. All presentations were scheduled

prior to the workshop, and due to the flood of volunteered talks, the organizers had to cap the number of slots. Talks were chosen so as to represent well the aspects of Web data management described above. The talks are listed below, classified by the covered topics. The classification is necessarily rough, as many talks crossed the boundaries between areas, in keeping with the seminar's intent. To the organizers' pleasant surprise, some of the results established surprising bridges between fields previously seen as unrelated (such as Machine Learning and Data Exchange), and brought in techniques from novel areas (such as Nominal Sets).

Crowdsourcing

- Tova Milo, Research Challenges in Crowdsourcing [tutorial]

Webdam Project Overview

- Emilien Antoine, Social Networking with WebdamExchange and WebdamLog
- Meghyn Bienvenu, A rule-based Language for Web Data Management

Web Data Management: Static Analysis

- Tom Ameloot, Relational Transducers for Declarative Networking
- Marie-Christine Rousset, Alignment-based Trust for Resource Finding in Semantic P2P Networks
- Alin Deutsch, Feasible Verification of Expressive Business Processes
- Anca Muscholl, Some Decidability Results on Distributed Games
- Evgeny Kharlamov, Evolution of DL-Lite Knowledge Bases
- Sophie Tison, Views and Updates

XPath with Data and Data Words

- Diego Figueira, XPath Decidability [tutorial]
- Mikołaj Bojańczyk, Temporal Logic on Changing XML Documents
- Szymon Toruńczyk, Automata-based Verification Over Linearly Ordered Data Domains
- Thomas Zeume, Two-Variable Logic, Orders and Successors

Probabilistic Data

- Pierre Senellart, PARIS: Probabilistic Alignment of Relations, Instances and Schemas
- Robert Fink, Aggregation in Probabilistic Databases via Knowledge Computation
- Serge Abiteboul, Finding Optimal Probabilistic Generators for XML Collections

FoXLib Project Overview

- Frank Neven, An Overview of FOXLIB
- Maarten Marx, Collections of XML, Schemas and Queries in FoxLib

SPARQL and Regular Expressions

- Wim Martens, The Complexity of Evaluating SPARQL Property Paths
- Giorgio Ghelli, Type-checking for SPARQL
- Juan Reutter, Parameterized Regular Expressions and their Languages

Schema Mappings and Data Exchange

- Phokion G. Kolaitis, Characterizing Schema Mappings with Examples [tutorial]
- Filip Murlak, XML Data Exchange
- Balder ten Cate, Learning Schema Mappings
- Marcelo Arenas, Data Exchange Beyond Complete Data

Incompleteness

- Leonid Libkin, A New Look at Incompleteness in Relations, XML, and Beyond [tutorial]

Nominal Sets and Access Paths

- Mikołaj Bojańczyk, Nominal Sets: An Introduction
- Sławomir Lasota, Nominal Sets: Automata
- Pierre Bourhis, Querying Access Paths

Queries

- Nicole Schweikardt, Expressiveness and Static Analysis of Extended Conjunctive Regular Path Queries
- Wojtek Kazana, Query Enumeration with Constant Delay
- Frank Neven, Deciding Twig-Definability of Node-Selecting Tree Automata
- Stijn Vansummeren, A New Characterization of the Acyclic Conjunctive Queries and Its Application to Structural Indexing
- Dan Olteanu, Factorized Representation of Query Results

Concluding Remarks and Future Plans

Due to the rich coverage of the area of foundations of Web data management, as achieved by both the presentations and the informal interactions, the organizers regard the seminar as a great success.

The weeklong format was well-suited to such an ambitious topic. The topic was well-received, as witnessed by the high rate of accepted invitations, and the exemplary degree of involvement by the participants. These volunteered such a high number of exceptional-quality talks that the organizers were faced with not being able to accommodate demand.

Bringing together researchers from different areas of data management, programming languages, theoretical computer science and logic fostered valuable interactions and led to fruitful collaborations, as reflected also by the very positive feedback from the audience.

The organizers wish to express their gratitude toward the Scientific Directorate of the Center for its support of this seminar, and hope to continue this seminar series on Web data management.

2 Table of Contents

Executive Summary

<i>Serge Abiteboul, Alin Deutsch, Thomas Schwentick, and Luc Segoufin</i>	37
---	----

Overview of Talks


Deduction in the Presence of Distribution, Contradictions and Uncertainty <i>Serge Abiteboul</i>	44
Deciding Eventual Consistency for a Simple Class of Relational Transducer Networks <i>Tom Ameloot</i>	44
Distributed Knowledge Base: Webdam System <i>Emilien Antoine</i>	45
A Rule-based Language for Web Data Management <i>Meghyn Bienvenu</i>	45
Efficient evaluation for a temporal logic on changing XML documents <i>Mikołaj Bojańczyk</i>	46
Nominal Sets <i>Mikołaj Bojańczyk</i>	46
Feasible Verification of Expressive Business Processes <i>Alin Deutsch</i>	46
Satisfiability for XPath <i>Diego Figueira</i>	47
Aggregation in Probabilistic Databases via Knowledge Compilation <i>Robert Fink</i>	47
Query Enumeration with Constant Delay <i>Wojtek Kazana</i>	48
Evolution of Knowledge Bases: DL-Lite case <i>Evgeny Kharlamov</i>	48
Schema Mappings and Data Examples <i>Phokion G. Kolaitis</i>	49
A new look at incompleteness in relations, XML, and beyond <i>Leonid Libkin</i>	49
The Complexity of Evaluating Path Expressions in SPARQL <i>Wim Martens</i>	50
Collections of XML, Schemas and Queries in FoXLib <i>Maarten Marx</i>	50
Asking the Right Questions in Crowd Data Sourcing <i>Tova Milo</i>	51
Solutions in XML data exchange <i>Filip Murlak</i>	51
Controlling distributed systems <i>Anca Muscholl</i>	52

Deciding Twig-definability of Node Selecting Tree Automata <i>Frank Neven</i>	52
Factorised Representations of Query Results <i>Dan Olteanu</i>	53
Alignment-based Trust for Resource Finding in Semantic P2P Networks <i>Marie-Christine Rousset</i>	53
Expressiveness and static analysis of extended conjunctive regular path queries <i>Nicole Schweikardt</i>	54
Finding Optimal Probabilistic Generators for XML Collections <i>Pierre Senellart</i>	54
Learning Schema Mappings <i>Balder Ten Cate</i>	55
Verification over linearly ordered data domains <i>Szymon Toruńczyk</i>	55
A new characterization of the acyclic conjunctive queries, and its application to structural indexing. <i>Stijn Vansummeren</i>	56
Two-Variable Logic, Orders and Successors <i>Thomas Zeume</i>	56
Participants	57

3 Overview of Talks

3.1 Deduction in the Presence of Distribution, Contradictions and Uncertainty


Serge Abiteboul (ENS – Cachan, FR)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Serge Abiteboul

We study deduction, captured by datalog-style rules, in the presence of contradictions, captured by functional dependencies (FDs). We start with a simple semantics for datalog in the presence of functional dependencies that is based on inferring facts one at a time, never violating the FDs, until no further facts can be added. This is a non-deterministic semantics, that may lead to several possible worlds. We present a proof theory for this semantics and compare it to previous work on datalog with negation. We also discuss a set-at-a-time semantics, where in each iteration, all facts that can be inferred are added to the database, and then choices are made between contradicting facts. We then proceed to our main goal of defining a semantics for the distributed setting. Note that contradictions naturally arise in a distributed setting since different peers may have conflicting information, opinions or recommendations. In the distributed case, we propose and study a concrete semantics for (an important fragment of) a previously proposed distributed datalog idiom, namely Webdamlog, that we enrich to account for FDs. Here again, we compare the semantics with previously studied semantics and in particular Webdamlog with negation in the absence of FDs. Finally, we note that in a distributed environment, it is natural to settle contradictions by introducing probabilities. We consider a simple adaptation of the distributed semantics to a probabilistic setting and show that it captures an intuitive way of resolving contradictions. We propose a sampling algorithm for evaluating queries under this semantics.

3.2 Deciding Eventual Consistency for a Simple Class of Relational Transducer Networks

Tom Ameloot (Hasselt University, BE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Tom Ameloot

Joint work of Ameloot, Tom; Van den Bussche, Jan
Main reference T.J. Ameloot, J. Van den Bussche, “Deciding Eventual Consistency for a Simple Class of Relational Transducer Networks,” in Proceedings of the 15th International Conference on Database Theory, 2012, to appear.

Networks of relational transducers can serve as a formal model for declarative networking, focusing on distributed database querying applications. In declarative networking, a crucial property is eventual consistency, meaning that the final output does not depend on the message delays and reorderings caused by the network. Here, we show that eventual consistency is decidable when the transducers satisfy some syntactic restrictions, some of which have also been considered in earlier work on automated verification of relational transducers. This simple class of transducer networks computes exactly all distributed queries expressible by unions of conjunctive queries with negation.

3.3 Distributed Knowledge Base: Webdam System

Emilien Antoine (INRIA Saclay – Orsay, FR)

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Emilien Antoine

Joint work of Abiteboul, Serge; Antoine, Emilien; Bienvenu, Meghyn; Galland, Alban

Main reference S. Abiteboul, M. Bienvenu, A. Galland, E. Antoine, “A rule-based language for web data management,” Proceedings of the 30th ACM Symposium on Principles of Database Systems (PODS’11), pp. 293–304, Athens, Greece, 2011.

URL <http://hal.inria.fr/docs/00/58/28/91/PDF/pods17a-abiteboul.pdf>

URL <http://dx.doi.org/10.1145/1989284.1989320>

As an extension to the talk about WebdamLog which is a model for a distributed declarative database in a peer to peer environment, I present (1) the WebdamExchange system for access control management on top of WebdamLog and (2) some clues about the implementation of a WebdamLog engine. WebdamExchange provides also an architecture for communication, to deal with heterogeneity of peers on the Internet, introduced as Webdam for the peer.

References

- 1 Serge Abiteboul and Meghyn Bienvenu and Alban Galland and Emilien Antoine. *A rule-based language for web data management*. Proceedings of the Symposium on Principles of Database Systems, pp. 293–304, Athens, Greece, 2011.
- 2 Serge Abiteboul and Alban Galland and Neoklis Polyzotis. *A Model for Web Information Management with Access Control*. Proceedings of the International Workshop on the Web and Databases, Athens, Greece, 2011.

3.4 A Rule-based Language for Web Data Management

Meghyn Bienvenu (INRIA Saclay – Orsay, FR)

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Meghyn Bienvenu

Joint work of Abiteboul, Serge; Antoine, Emilien; Bienvenu, Meghyn; Galland, Alban

Main reference S. Abiteboul, M. Bienvenu, A. Galland, E. Antoine, “A rule-based language for web data management,” Proceedings of the 30th ACM Symposium on Principles of Database Systems (PODS’11), pp. 293–304, Athens, Greece, 2011.


URL <http://hal.inria.fr/docs/00/58/28/91/PDF/pods17a-abiteboul.pdf>

URL <http://dx.doi.org/10.1145/1989284.1989320>

There is a new trend to use Datalog-style rule-based languages to specify modern distributed applications, notably on the Web. In this talk, I will introduce such a language (called Webdamlog) for a distributed data model where peers exchange messages (i.e. logical facts) as well as rules. After illustrating the language, I will mention some results concerning the connection with centralized Datalog semantics and the impact on expressiveness of “delegations” (the installation of rules by a peer in some other peer) and explicit timestamps.

3.5 Efficient evaluation for a temporal logic on changing XML documents

Mikołaj Bojańczyk (University of Warsaw, PL)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Mikołaj Bojańczyk

Joint work of Bojańczyk, Mikołaj; Figueira, Diego

This talk is about a logic that describes a changing XML document. The changing XML document is modeled as a sequence of trees, over a finite alphabet. The logic can define properties such as: “every node changes its label at most twice”, or “whenever a node gets label a , then one of its descendants eventually gets label c ”. The contribution is an evaluation algorithm, which tests if a formula is true in a sequence of trees, assuming that the edit distance between consecutive trees is at most 1. The algorithm runs in time $n \log(k)$, where n is the number of trees and k is their maximal size.

3.6 Nominal Sets

Mikołaj Bojańczyk (University of Warsaw, PL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Mikołaj Bojańczyk


Joint work of Klin, Bartek; Bojańczyk, Mikołaj; Lasota, Sławomir

Nominal sets are a different kind of set theory. Nominal sets were invented by Abraham Fraenkel in the 1920’s. They were rediscovered for computer science by Gabbay and Pitts in 1999, as a way of talking about name binding in lambda terms and logical formulas. We rediscover them yet again, this time as a way of talking about data values, including data words and automata for data words.

The point in nominal sets is that there is a different notion of finite set, e.g. the set of all data words of length at most 10 is a finite set.

3.7 Feasible Verification of Expressive Business Processes

Alin Deutsch (University of California – San Diego, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Alin Deutsch

Joint work of Deutsch, Alin; Damaggio, Elio; Vianu, Victor

We revisit the static verification problem for data centric business processes, specified in a variant of IBM’s “business artifact” model. Artifacts are records of variables that correspond to business-relevant objects and are updated by a set of services equipped with pre-and postconditions, that implement business process tasks. The verification problem consists in statically checking whether all runs of an artifact system satisfy desirable properties expressed in a first order extension of linear-time temporal logic.


In previous work we identified the class of guarded artifact systems and properties, for which verification is decidable. However, the 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 an ICDT 2011 paper, 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, but the upper bound is non-elementary.

This talk reports on a new technique developed since, leading to more palatable upper bound (EXPSpace). The technique makes practical implementation feasible, and a preliminary experimental evaluation of our prototype verifier yields encouraging results.

3.8 Satisfiability for XPath

Diego Figueira (University of Edinburgh, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Diego Figueira

XPath is a node selecting language for XML documents. Although its satisfiability problem is in general undecidable, there are several syntactic restrictions that make the problem decidable. I will present a survey on the existing results on the satisfiability problem for fragments of XPath in the presence of data values. I will also mention some open problems and conjectures.

3.9 Aggregation in Probabilistic Databases via Knowledge Compilation

Robert Fink (University of Oxford, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Robert Fink

Joint work of Fink, Larisa; Han, Larisa; Olteanu, Dan


Main reference R. Fink, L. Han, D. Olteanu, “Aggregation in Probabilistic Databases via Knowledge Compilation,” VLDB 2012, pp. 490–501.

This talk presents a query evaluation technique for positive relational algebra queries with aggregates on a representation system for probabilistic data based on the algebraic structures of semirings and semimodule. The core of our evaluation technique is a procedure that compiles semimodule and semiring expressions into so-called decomposition trees, for which the computation of the probability distribution can be done in polynomial time in the size of the tree and of the distributions represented by its nodes. We give syntactic characterisations of tractable queries with aggregates by exploiting the connection between query tractability and polynomial-time compilation into decomposition trees.

The technique is incorporated into the probabilistic database engine SPROUT, which is built on top of PostgreSQL. We report on extensive performance experiments with synthetic datasets and TPC-H data.

3.10 Query Enumeration with Constant Delay

Wojtek Kazana (*ENS – Cachan, FR*)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Wojtek Kazana

In many applications the output of a query may have a huge size and enumerating all the answers may already consume too many of the allowed resources. In this case it may be appropriate to first output a small subset of the answers and then, on demand, output a subsequent small numbers of answers and so on until all possible answers have been exhausted. To make this even more attractive it is preferable to be able to minimize the time necessary to output the first answers and, from a given set of answers, also minimize the time necessary to output the next set of answers – this second time interval is known as the delay. For this it might be interesting to compute adequate index structures. The ultimate goal being to obtain index structures easily computable (say in linear time in the size of the database), that allow constant delay in the enumeration process. In this case we speak of constant delay enumeration of the query that was introduced by Durand and Grandjean.

In this talk I will outline the differences between query evaluation and enumeration. Using the example of MSO queries over trees I will try to illustrate some techniques useful in obtaining constant delay enumeration algorithms.

3.11 Evolution of Knowledge Bases: DL-Lite case

Evgeny Kharlamov (*Free University Bozen-Bolzano, IT*)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Evgeny Kharlamov

Joint work of Calvanese, Diego; Kharlamov, Evgeny; Nutt, Werner; Zheleznyakov, Dmitriy

Main reference E. Kharlamov, D. Zheleznyakov, “Capturing Instance Level Ontology Evolution for DL-Lite,” in Proc. of ISWC, 2011.

URL <http://www.inf.unibz.it/~kharlamov/publications-date.html>

We study the problem of evolution for Knowledge Bases (KBs) expressed in Description Logics (DLs) of the DL-Lite family. DL-Lite is at the basis of OWL 2 QL, one of the tractable fragments of OWL 2, the recently proposed revision of the Web Ontology Language. We review known model (MBAs) and formula-based approaches (FBAs) for evolution of propositional theories. We exhibit limitations of MBAs: they intrinsically ignore the structural properties of KBs, which leads to undesired properties of KBs resulting from such an evolution, i.e., DL-Lite is not closed under all considered MBAs. We show what causes inexpressibility and exhibit a fragment of DL-Lite that is closed under a number of MBAs. We show that standard FBAs are also not appropriate for DL-Lite evolution, either due to high complexity of computation, or because the result of such an action of evolution is not expressible in DL-Lite. We propose two formula-based approaches for which evolution is expressible in DL-Lite and can be computed in polynomial time.


The talk is based on the following papers [1, 2].

References

- 1 Diego Calvanese, Evgeny Kharlamov, Werner Nutt, and Dmitriy Zheleznyakov. *Evolution of DL-Lite Knowledge Bases*. In Proc. of ISWC, 2010
- 2 Evgeny Kharlamov and Dmitriy Zheleznyakov. *Capturing Instance Level Ontology Evolution for DL-Lite*. In Proc. of ISWC, 2011

3.12 Schema Mappings and Data Examples

Phokion G. Kolaitis (University of California – Santa Cruz, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Phokion G. Kolaitis

Joint work of Alexe, Bogdan; ten Cate, Balder; Kolaitis, Phokion G.; Tan, Wang-Chiew


Schema mappings are high-level specifications that describe the relationship between two database schemas. Schema mappings are considered to be the essential building blocks in such critical data interoperability tasks as data exchange and data integration. For this reason, they have been the focus of extensive research investigations over the past several years. Since in real-life applications schema mappings can be quite complex, it is important to develop methods and tools for illustrating, explaining, and deriving schema mappings. A promising approach to this effect is to use “good” data examples that illustrate the schema mapping at hand.

In this talk, we present an overview of recent work on characterizing and deriving schema mappings via a finite set of data examples. We show that every LAV schema mapping (i.e., a schema mapping specified by a finite set of local-as-view tuple-generating dependencies) is uniquely characterized by a finite set of universal data examples with respect to the class of all LAV schema mappings. We also show that this type of result does not hold for arbitrary GAV schema mappings (i.e., schema mappings specified by a finite set of global-as-view tuple-generating dependencies). After this, we give a necessary and sufficient algorithmic condition for a GAV schema mapping to be uniquely characterizable by a finite set of universal examples with respect to the class of all GAV schema mappings. Along the way, we establish tight connections between unique characterizability of schema mappings and homomorphism dualities.

This is joint work with Bogdan Alexe (IBM Research – Almaden), Balder ten Cate (UC Santa Cruz), and Wang-Chiew Tan (UC Santa Cruz and IBM Research – Almaden).

3.13 A new look at incompleteness in relations, XML, and beyond

Leonid Libkin (University of Edinburgh, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Leonid Libkin


While incomplete information is ubiquitous in all data models - especially in applications involving data translation or integration – our understanding of it is still not completely satisfactory. For example, even such a basic notion as certain answers for XML queries was only introduced recently, and in a way seemingly rather different from relational certain answers.

Here we propose a general approach to handling incompleteness, and test its applicability in known data models such as relations and documents. The approach is based on representing degrees of incompleteness via semantics-based orderings on database objects. We use it to both obtain new results on incompleteness and to explain some previously observed phenomena. Specifically we show that certain answers for relational and XML queries are two instances of the same general concept; we describe structural properties behind the naive evaluation of queries; answer open questions on the existence of certain answers in the XML setting; and show that previously studied ordering-based approaches were only adequate

for SQL’s primitive view of nulls. We define a general setting that subsumes relations and documents to help us explain in a uniform way how to compute certain answers, and when good solutions can be found in data exchange. We also look at the complexity of common problems related to incompleteness, and generalize several results from relational and XML contexts.

3.14 The Complexity of Evaluating Path Expressions in SPARQL

Wim Martens (Universität Bayreuth, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Wim Martens

The World Wide Web Consortium (W3C) recently included property paths in the working draft for SPARQL 1.1, a query language for RDF data. Property paths give SPARQL queries the power of evaluating regular expressions over graph data. However, they differ from regular expressions in several notable aspects.

For example, they include a limited form of negation and they can use counters as syntactic sugar. Furthermore, their semantics on graphs is defined in a non-standard manner.

We formalize the W3C semantics of property paths and investigate the impact on the complexity of various query evaluation problems on graphs. More specifically, let x and y be two nodes in an edge-labeled graph and r be a regular expression. We investigate the complexities of (1) deciding whether there exists a path from x to y that matches r and (2) counting how many paths from x to y match r .

Our main results show that, compared to an alternative semantics of regular expressions on graphs, the W3C semantics causes a significant increase in the complexity of problems (1) and (2). Whereas the alternative semantics remains in polynomial time for fairly large fragments of expressions, the W3C semantics makes problems (1) and (2) intractable almost immediately.

As a side-result, we also prove that the membership problem for regular expressions with counters and negation is in polynomial time.

3.15 Collections of XML, Schemas and Queries in FoXLib

Maarten Marx (University of Amsterdam, NL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Maarten Marx

Joint work of Grijzenhout, Steven; Marx, Maarten

Main reference S. Grijzenhout, M. Marx, “The Quality of the XML web,” in Proc. of 20th ACM Int’l Conf. on Information and Knowledge Management (CIKM’11), pp. 1719–1724, 2011.

URL <http://dx.doi.org/10.1145/2063576.2063824>

We collect evidence to answer the following question: Is the quality of the XML documents found on the web sufficient to apply XML technology like XQuery, XPath and XSLT? XML collections from the web have been previously studied statistically, but no detailed information about the quality of the XML documents on the web is available to date. We address this shortcoming in this study. We gathered 180K XML documents from the web. Their quality is surprisingly good; 85.4% is well-formed and 99.5% of all specified encodings is correct. Validity needs serious attention. Only 25% of all files contain a reference to a DTD

or XSD, of which just one third is actually valid. Errors are studied in detail. Automatic error repair seems promising. Our study is well documented and easily repeatable. This paves the way for a periodic quality assessment of the XML web.

All data is publicly available at the url <http://data.politicalmashup.nl/xmlweb>.

3.16 Asking the Right Questions in Crowd Data Sourcing

Tova Milo (Tel Aviv University, IL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Tova Milo


Crowd-based data sourcing is a new and powerful data procurement paradigm that engages Web users to collectively contribute data, analyze information and share opinions. This brings to light, out of the huge, inconsistent Web ocean, an important body of knowledge that would otherwise not be attainable. Crowd-based data sourcing democratizes data-collection, cutting companies' and researchers' reliance on stagnant, overused datasets and bears great potential for revolutionizing our information world. Yet, triumph has so far been limited to only a handful of successful projects such as Wikipedia or IMDb. This comes notably from the difficulty of managing huge volumes of data and users of questionable quality and reliability. Every single initiative had to battle, almost from scratch, the same non-trivial challenges. The ad hoc solutions, even when successful, are application specific and rarely sharable.

In this talk we consider the development of solid scientific foundations for Web-scale data sourcing. We believe that such a principled approach is essential to obtain knowledge of superior quality, to realize the task more effectively and automatically, be able to reuse solutions, and thereby to accelerate the pace of practical adoption of this new technology that is revolutionizing our life.

We discuss the desired logical, algorithmic, and methodological foundations for the management of large scale crowd-sourced data and for the development of applications over such information. This encompasses formal models capturing all the diverse facets of crowd-sourced data. This also means developing the necessary reasoning capabilities for managing and controlling data sourcing, cleaning, verification, integration, sharing, querying and updating, in a dynamic Web environment.

3.17 Solutions in XML data exchange

Filip Murlak (University of Warsaw, PL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Filip Murlak


Joint work of Bojańczyk, Mikołaj; Kołodziejczyk, Leszek A.; Murlak, Filip

The task of XML data exchange is to restructure a document conforming to a source schema under a target schema according to certain mapping rules. The rules are typically expressed as source-to-target dependencies using various kinds of patterns, involving horizontal and vertical navigation, as well as data comparisons. The target schema imposes complex conditions on the structure of solutions, possibly inconsistent with the mapping rules. In consequence, for some source documents there may be no solutions. I will discuss three

computational problems: deciding if all documents of the source schema can be mapped to a document of the target schema (absolute consistency), deciding if a given document of the source schema can be mapped (solution existence), and constructing a solution for a given source document (solution building). It turns out that the complexity of absolute consistency is rather high in general, but within the polynomial hierarchy for bounded depth schemas. The combined complexity of solution existence and solution building behaves similarly, but the data complexity is very low. In fact, even for very expressive mapping rules, based on MSO definable queries, absolute consistency is decidable and data complexity of solution existence is polynomial.

3.18 Controlling distributed systems

Anca Muscholl (Université Bordeaux, FR)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Anca Muscholl

Joint work of Genest, B.; Gimbert, H.; Muscholl, Anca; Walukiewicz, I.

We consider the problem of controlling distributed automata that cooperate via shared variables (rendez-vous). The setting corresponds to the framework of Ramadge and Wonham, where certain actions (controllable ones) can be forbidden by the local controller. Although the general question is still open, we can show that the problem is decidable on acyclic architectures, albeit of non-elementary complexity.

3.19 Deciding Twig-definability of Node Selecting Tree Automata

Frank Neven (Hasselt University, BE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Frank Neven

Node selecting tree automata (NSTAs) constitute a general formalism defining unary queries over trees.


Basically, a node is selected by an NSTA when it is visited in a selecting state during an accepting run.

We consider twig patterns as an abstraction of XPath. Since the queries definable by NSTAs form a strict superset of twig-definable queries, we study the complexity of the problem to decide whether the query by a given NSTA is twig-definable. In particular, we obtain that the latter problem is EXPTIME-complete.

In addition, we show that it is also EXPTIME-complete to decide whether the query by a given NSTA is definable by a node selecting string automaton.

3.20 Factorised Representations of Query Results

Dan Olteanu (University of Oxford, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Dan Olteanu

Joint work of Olteanu, Dan; Zavodny, Jakub

Main reference D. Olteanu, J. Zavodny, “Factorised Representations of Query Results,” arXiv:1104.0867v1 [cs.DB], to appear in ICDT 2012.

URL <http://arxiv.org/abs/1104.0867v1>

We introduce a representation system for relational data based on algebraic factorisation using distributivity of product over union and commutativity of product and union.

We give two characterisations of conjunctive queries based on factorisations of their results defined by a certain class of hyperpath decompositions of the query hypergraph.

The first characterisation concerns sizes of factorised representations. For any query, we derive a size bound that is asymptotically tight within our class of factorisations.

For relations where tuples are annotated with identifiers we also characterise the queries by the readability of their results, which is the minimum over all equivalent factorisations of the maximum number of occurrences of any identifier in that factorisation. We give a dichotomy of queries based on the readability of their results for any database and define syntactically the class of queries with bounded readability.

3.21 Alignment-based Trust for Resource Finding in Semantic P2P Networks

Marie-Christine Rousset (Université de Grenoble, FR)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Marie-Christine Rousset

Joint work of Atencia, Manuel; Euzenat, Jérôme; Pirrò, Giuseppe; Rousset, Marie-Christine

Main reference M. Atencia, J. Euzenat, G. Pirrò, M.-C. Rousset, “Alignment-Based Trust for Resource Finding in Semantic P2P Networks,” in Proc. of Int’l Semantic Web Conference (ISWC’11), pp. 51–66, 2011.

URL http://dx.doi.org/10.1007/978-3-642-25073-6_4


In a semantic P2P network, peers use separate ontologies and rely on alignments between their ontologies for translating queries.

Nonetheless, alignments may be limited (unsound or incomplete) and generate flawed translations, leading to unsatisfactory answers. In this paper we present a trust mechanism that can assist peers to select those in the network that are better suited to answer their queries. The trust that a peer has towards another peer depends on a specific query and represents the probability that the latter peer will provide a satisfactory answer. In order to compute trust, we exploit both alignments and peers’ direct experience, and perform Bayesian inference. We have implemented our technique and conducted an evaluation. Experimental results showed that trust values converge as more queries are sent and answers received.

Furthermore, the use of trust improves both precision and recall.

3.22 Expressiveness and static analysis of extended conjunctive regular path queries

Nicole Schweikardt (Goethe-Universität Frankfurt am Main, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Nicole Schweikardt

Joint work of Freydenberger, Dominik; Schweikardt, Nicole

Main reference D. Freydenberger and N. Schweikardt, “Expressiveness and static analysis of extended conjunctive regular path queries,” in Proc. of the 5th Alberto Mendelzon Int’l Workshop on Foundations of Data Management (AMW’11), vol. 749 of CEUR Workshop Proceedings, CEUR-WS.org, 2011.

URL <http://ceur-ws.org/Vol-749/paper9.pdf>

We study the expressiveness and the complexity of static analysis of extended conjunctive regular path queries (ECRPQs), introduced by Barcelo et al. (PODS’10). ECRPQs are an extension of conjunctive regular path queries (CRPQs), a well-studied language for querying graph structured databases. Our first main result shows that query containment and equivalence of a CRPQ in an ECRPQ is undecidable. This settles one of the main open problems posed by Barcelo et al.


As a second main result, we prove a non-recursive succinctness gap between CRPQs and the CRPQ-expressible fragment of ECRPQs. Apart from this, we develop a tool for proving inexpressibility results for CRPQs and ECRPQs. In particular, this enables us to show that there exist queries definable by regular expressions with backreferencing, but not expressible by ECRPQs.

This is joint work with Dominik D. Freydenberger.

The material presented in this talk was published in the proceedings of the 5th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW 2011), vol. 749 of CEUR Workshop Proceedings, CEUR-WS.org, 2011.

3.23 Finding Optimal Probabilistic Generators for XML Collections

Pierre Senellart (Telecom Paris Tech, FR)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Pierre Senellart

Joint work of Abiteboul, Serge; Amsterdamer, Yael; Deutch, Daniel; Milo, Tova; Senellart, Pierre


Main reference S. Abiteboul, Y. Amsterdamer, D. Deutch, T. Milo, P. Senellart, “Finding Optimal Probabilistic Generators for XML Collections,” in Proc. ICDT, Berlin, Germany, March 2012.

URL <http://pierre.senellart.com/publications/abiteboul2012finding.pdf>

We study the problem of, given a corpus of XML documents and its schema, finding an optimal (generative) probabilistic model, where optimality here means maximizing the likelihood of the particular corpus to be generated. Focusing first on the structure of documents, we present an efficient algorithm for finding the best generative probabilistic model, in the absence of constraints. We further study the problem in the presence of integrity constraints, namely key, inclusion, and domain constraints. We study in this case two different kinds of generators. First, we consider a continuation-test generator that performs, while generating documents, tests of schema satisfiability; these tests prevent from generating a document violating the constraints but, as we will see, they are computationally expensive. We also study a restart generator that may generate an invalid document and, when this is the case, restarts and tries again. Finally, we consider the injection of data values into the structure, to obtain a full XML document. We study different approaches for generating these values.

3.24 Learning Schema Mappings

Balder Ten Cate (University of California – Santa Cruz, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Balder Ten Cate

Joint work of Kolaitis, Phokion G.; Dalmau, Victor; Cate, Ten Cate, Balder
Main reference B. Ten Cate, V. Dalmau, Ph. Kolaitis, “Learning Schema Mappings,” in Proc. of ICDT 2012, to appear.

A schema mapping is a high-level specification of the relationship between a source schema and a target schema. Recently, a line of research has emerged that aims at deriving schema mappings automatically or semi-automatically with the help of data examples, i.e., pairs consisting of a source instance and a target instance that depict, in some precise sense, the intended behavior of the schema mapping. Several different uses of data examples for deriving, refining, or illustrating a schema mapping have already been proposed and studied.


In this paper, we use the lens of computational learning theory to systematically investigate the problem of obtaining algorithmically a schema mapping from data examples. Our aim is to leverage the rich body of work on learning theory in order to develop a framework for exploring the power and the limitations of the various algorithmic methods for obtaining schema mappings from data examples. We focus on GAV schema mappings, that is, schema mappings specified by GAV (Global-As-View) constraints. GAV constraints are the most basic and the most widely supported language for specifying schema mappings.

We present an efficient algorithm for learning GAV schema mappings using Angluin’s model of exact learning with membership and equivalence queries. This is optimal, since we show that neither membership queries nor equivalence queries suffice, unless the source schema consists of unary relations only. We also obtain results concerning the learnability of schema mappings in the context of Valiant’s well known PAC (Probably-Approximately-Correct) learning model.

Finally, as a byproduct of our work, we show that there is no efficient algorithm for approximating the shortest GAV schema mapping fitting a given set of examples, unless the source schema consists of unary relations only.

3.25 Verification over linearly ordered data domains

Szymon Toruńczyk (ENS – Cachan, FR)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Szymon Toruńczyk


Joint work of Segoufin, Luc; Toruńczyk, Szymon

We work over linearly ordered data domains equipped with finitely many unary predicates and constants. We consider nondeterministic automata processing words and storing finitely many variables ranging over the domain. During a transition, these automata can compare the data values of the current configuration with those of the previous configuration using the linear order, the unary predicates and the constants.

We show that emptiness for such automata is decidable, both over finite and infinite words, under reasonable computability assumptions on the linear order. Finally, we show how our automata model can be used for verifying properties of work-flow specifications in the presence of an underlying database.

3.26 A new characterization of the acyclic conjunctive queries, and its application to structural indexing.

Stijn Vansummeren (Université Libre de Bruxelles, BE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Stijn Vansummeren

We present a new structural characterization of the expressive power of the acyclic conjunctive queries in terms of guarded simulations. The study of this fragment of first order logic is motivated by the central role it plays in query languages across a wide range of data models.

We discuss the relevance of this result as a formal basis for constructing so-called structural indexes. Structural indexes were first proposed in the context of semi-structured query languages and later successfully applied as an XML indexation mechanism for XPath-like queries. We discuss how our main result can be instantiated to the construction of structural indexes for RDF on the Semantic Web.

3.27 Two-Variable Logic, Orders and Successors

Thomas Zeume (TU Dortmund, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Thomas Zeume

Joint work of Manuel, Amaldev; Schwentick, Thomas; Zeume, Thomas

Recent results for the finite satisfiability problem for two-variable logic over structures with linear order, preorder, successor and unary relations will be discussed in this talk.

Two-variable logic with one total preorder relation, its induced successor relation, one linear order relation and some further unary relations is EXPSPACE-complete. Actually, EXPSPACE-completeness already holds for structures that do not include the induced successor relation. As a special case, the EXPSPACE upper bound applies to two-variable logic over structures with two linear orders. A further consequence is that satisfiability of two-variable logic over data words with a linear order on positions and a linear order and successor relation on the data is decidable in EXPSPACE. Furthermore, two-variable logic is decidable on structures with two linear order successors and an order corresponding to one of the successors.

Those results are complemented by the undecidability of the finite satisfiability problem for two-variable logic over structures with two total preorder relations as well as over structures with one total preorder and two linear order relations.

Participants

- Serge Abiteboul
ENS – Cachan, FR
- Tom Ameloot
Hasselt University, BE
- Emilien Antoine
INRIA Saclay – Orsay, FR
- Timos Antonopoulos
Hasselt University, BE
- Marcelo Arenas
Univ. Católica de Chile, CL
- Pablo Barcelo
Univ. of Chile – Santiago, CL
- Meghyn Bienvenu
INRIA Saclay – Orsay, FR
- Mikołaj Bojańczyk
University of Warsaw, PL
- Pierre Bourhis
University of Oxford, UK
- Claire David
Université Paris-Est –
Marne-la-Vallée, FR
- Alin Deutsch
University of California – San
Diego, US
- Diego Figueira
University of Edinburgh, UK
- Robert Fink
University of Oxford, UK
- Amélie Gheerbrant
University of Edinburgh, UK
- Giorgio Ghelli
University of Pisa, IT
- Florent Jacquemard
ENS – Cachan, FR
- Ahmet Kara
TU Dortmund, DE
- Wojtek Kazana
ENS – Cachan, FR
- Evgeny Kharlamov
Free Univ. Bozen-Bolzano, IT
- Pekka Kilpeläinen
University of Kuopio, FI
- Christoph Koch
EPFL – Lausanne, CH
- Phokion G. Kolaitis
University of California – Santa
Cruz, US
- Sławomir Lasota
University of Warsaw, PL
- Leonid Libkin
University of Edinburgh, UK
- Sebastian Maneth
Univ. of New South Wales, AU
- Wim Martens
Universität Bayreuth, DE
- Maarten Marx
University of Amsterdam, NL
- Tova Milo
Tel Aviv University, IL
- Filip Murlak
University of Warsaw, PL
- Anca Muscholl
Université Bordeaux, FR
- Frank Neven
Hasselt University, BE
- Matthias Nierwerth
TU Dortmund, GE
- Dan Olteanu
University of Oxford, UK
- Pawel Parys
University of Warsaw, PL
- Juan L. Reutter
University of Edinburgh, UK
- Marie-Christine Rousset
Université de Grenoble, FR
- Anne Schuth
University of Amsterdam, NL
- Nicole Schweikardt
Goethe-Universität Frankfurt am
Main, DE
- Thomas Schwentick
TU Dortmund, DE
- Luc Segoufin
ENS – Cachan, FR
- Helmut Seidl
TU München, DE
- Pierre Senellart
Telecom Paris Tech, FR
- Cristina Sirangelo
ENS – Cachan, FR
- Tony Tan
University of Edinburgh, UK
- Balder Ten Cate
University of California – Santa
Cruz, US
- Sophie Tison
Université de Lille I, FR
- Szymon Toruńczyk
ENS – Cachan, FR
- Jan Van den Bussche
Hasselt University, BE
- Stijn Vansummeren
Université Libre de Bruxelles, BE
- Domagoj Vrgoc
University of Edinburgh, UK
- Thomas Zeume
TU Dortmund, DE

