Activity Report 2015

Team LINKS

Linking Dynamic Data

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

RESEARCH CENTER
Lille - Nord Europe

THEME
Data and Knowledge Representation and Processing

Table of contents

Team LINKS

Creation of the Team: 2013 January 01

Keywords:

Computer Science and Digital Science:

2.1. - Programming Languages
2.1.1. - Semantics of programming languages
2.1.3. - Functional programming
2.1.6. - Concurrent programming
3. - Data and knowledge
3.1. - Data
3.1.1. - Modeling, representation
3.1.2. - Data management, quering and storage
3.1.3. - Distributed data
3.1.4. - Uncertain data
3.1.5. - Control access, privacy
3.1.6. - Query optimization
3.1.7. - Open data
3.1.8. - Big data (production, storage, transfer)
3.1.9. - Database
3.2. - Knowledge
3.2.1. - Knowledge bases
3.2.2. - Knowledge extraction, cleaning
3.2.3. - Inference
3.2.4. - Semantic Web
3.2.5. - Ontologies
7. - Fundamental Algorithmics
7.4. - Logic in Computer Science
8. - Artificial intelligence
8.1. - Knowledge
8.2. - Machine learning

Other Research Topics and Application Domains:

6.3. - Network functions
6.3.1. - Web
6.3.3. - Network services
6.3.4. - Social Networks
6.5. - Information systems
9. - Society and Knowledge
9.4.1. - Computer science
9.4.5. - Data science
9.7. - Knowledge dissemination
9.7.1. - Open access

1. Members

Research Scientists
Joachim Niehren [Team leader, Inria, Senior Researcher, HdR]
Pierre Bourhis [CNRS, Researcher]

Faculty Members
Sophie Tison [Univ. Lille I, Professor, HdR]
Iovka Boneva [Univ. Lille I, Associate Professor]
Angela Bonifati [Univ. Lille I, Professor, until Aug 2015, HdR]
Aurélien Lemay [Univ. Lille III, Associate Professor]
Slawomir Staworko [Univ. Lille III, Associate Professor, HdR]

Engineer
Guillaume Bagan [Univ. Lille I, until Aug 2015]

PhD Students
Adrien Boiret [Univ. Lille I]
Vasile-Radu Ciucanu [Univ. Lille I, until Jul 2015]
Dimitri Gallois [Univ. Lille I, from Sep 2015]
Tom Sebastian

Post-Doctoral Fellow
Vincent Hugot [Univ. Lille I]

2. Overall Objectives

2.1. Presentation

We will develop algorithms for answering logical querying on heterogeneous linked data collections in hybrid formats, distributed programming languages for managing dynamic linked data collections and workflows based on queries and mappings, and symbolic machine learning algorithms that can link datasets by infering appropriate queries and mappings.

The following three paragraphs summarize our main research objectives.

Querying Heterogeneous Linked Data We will develop new kinds of schema mappings for semi-structured datasets in hybrid formats including graph databases, RDF collections, and relational databases. These induce recursive queries on linked data collections for which we will investigate evaluation algorithms, containment problems, and concrete applications.

Managing Dynamic Linked Data In order to manage dynamic linked data collections and workflows, we will develop distributed data-centric programming languages with streams and parallelism, based on novel algorithms for incremental query answering, study the propagation of updates of dynamic data through schema mappings, and investigate static analysis methods for linked data workflows.

Linking Data Graphs Finally, we will develop symbolic machine learning algorithms, for inferring queries and mappings between linked data collections in various graphs formats from annotated examples.

3. Research Program

3.1. Background

The main objective of LINKS is to develop methods for querying and managing linked data collections. Even though open linked data is the most prominent example, we will focus on hybrid linked data collections, which are collections of semi-structured datasets in hybrid formats: graph-based, RDF, relational, and NOSQL. The elements of these datasets may be linked, either by pointers or by additional relations between the elements of the different datasets, for instance the "same-as" or "member-of" relations as in RDF.

The advantage of traditional data models is that there exist powerful querying methods and technologies that one might want to preserve. In particular, they come with powerful schemas that constraint the possible manners in which knowledge is represented to a finite number of patterns. The exhaustiveness of these patterns is essential for writing of queries that cover all possible cases. Pattern violations are excluded by schema validation. In contrast, RDF schema languages such as RDFS can only enrich the relations of a dataset by new relations, which also helps for query writing, but which cannot constraint the number of possible patterns, so that they do not come with any reasonable notion of schema validation.

The main weakness of traditional formats, however, is that they do not scale to large data collections as stored on the Web, while the RDF data models scales well to very big collections such as linked open data. Therefore, our objective is to study mixed data collections, some of which may be in RDF format, in which we can lift the advantages of smaller datasets in traditional formats to much larger linked data collections. Such data collections are typically distributed over the internet, that some data sources have rigid query facilities that cannot be easily adapted or extended.

The main assumption that we impose in order to enable the logical approach, is that the given linked data collection must be correct in most dimensions. This means that all datasets are well-formed with respect to their available constraints and schemas, and clean with respect to the data values in most of the components of the relations in the datasets. One of the challenges is to integrate good quality RDF datasets into this setting, another is to clean the incorrect data in those dimensions that are less proper. It remains to be investigated in how far these assumptions can be maintained in realistic applications, and how much they can be weakened otherwise.

For querying linked data collections, the main problems are to resolve the heterogeneity of data formats and schemas, to understand the efficiency and expressiveness of recursive queries, that can follow links repeatedly, to answer queries under constraints, and to optimize query answering algorithms based on static analysis. When linked data is dynamically created, exchanged, or updated, the problems are how to process linked data incrementally, and how to manage linked data collections that change dynamically. In any case (static and dynamic) one needs to find appropriate schema mappings for linking semi-structured datasets. We will study how to automatize parts of this search process by developing symbolic machine learning techniques for linked data collections.

3.2. Querying Heterogeneous Linked Data

Our main objective is to query collections of linked datasets. In the static setting, we consider two kinds of links: explicit links between elements of the datasets, such as equalities or pointers, and logical links between relations of different datasets such as schema mappings. In the dynamic setting, we permit a third kind of links that point to "intentional" relations computable from a description, such as the application of a Web service or the application of a schema mapping.

We believe that collections of linked datasets are usually too big to ensure a global knowledge of all datasets. Therefore, schema mappings and constraints should remain between pairs of datasets. Our main goal is to be able to pose a query on a collection of datasets, while accounting for the possible recursive effects of schema mappings. For illustration, consider a ring of datasets D_1, D_2, D_3 linked by schema mappings M_1, M_2, M_3 that tell us how to complete a database D_i by new elements from the next database in the cycle.

The mappings M_i induce three intentional datasets I_1, I_2, and I_3, such that I_i contains all elements from D_i and all elements implied by M_i from the next intentional dataset in the ring:

$$I_1 = D_1 \cup M_1(I_2), \quad I_2 = D_2 \cup M_2(I_3), \quad I_3 = D_3 \cup M_3(I_1)$$

Clearly, the global information collected by the intentional datasets depends recursively on all three original datasets D_i. Queries to the global information can now be specified as standard queries to the intentional databases I_i. However, we will never materialize the intentional databases I_i. Instead, we can rewrite queries on one of the intentional datasets I_i to recursive queries on the union of the original datasets D_1, D_2, and D_3 with their links and relations. Therefore, a query answering algorithm is needed for recursive queries, that chases the "links" between the D_i in order to compute the part of I_i needed for the purpose of query answering.

This illustrates that we must account for the graph data models when dealing with linked data collections whose elements are linked, and that query languages for such graphs must provide recursion in order to chase links. Therefore, we will have to study graph databases with recursive queries, such as RDF graphs with SPARQL queries, but also other classes of graph databases and queries.

We study schemas and mappings between datasets with different kinds of data models and the complexity of evaluating recursive queries over graphs. In order to use schema mapping for efficiently querying the different datasets, we need to optimize the queries by taking into account the mappings. Therefore, we will study static analysis of schema mappings and recursive queries. Finally, we develop concrete applications in which our fundamental techniques can be applied.

3.3. Managing Dynamic Linked Data

With the quick growth of the information technology on the Web, more and more Web data gets created dynamically every day, for instance by smartphones, industrial machines, users of social networks, and all kinds of sensors. Therefore, large amounts of dynamic data need to be exchanged and managed by various data-centric web services, such as online shops, online newspapers, and social networks.

Dynamic data is often created by the application of some kind of service on the Web. This kind of data is intentional in the same spirit as the intentional data specified by the application of a schema mapping, or the application of some query to the hidden Web. Therefore, we will consider a third kind of links in the dynamic setting, that map to intentional data specified by whatever kind of function application. Such a function can be defined in data-centric programming languages, in the style of Active XML, XSLT, and NoSQL languages.

The dynamicity of data adds a further dimension to the challenges for linked data collections that we described before, while all the difficulties remain valid. One of the new aspects is that intentional data may be produced incrementally, as for instance when exchanged over data streams. Therefore, one needs incremental algorithms able to evaluate queries on incomplete linked data collections, that are extended or updated incrementally. Note that incremental data may be produced without end, such as a Twitter stream, so that one cannot wait for its completion. Instead, one needs to query and manage dynamic data with as low latency as possible. Furthermore, all static analysis problems are to be re-investigated in the presence of dynamic data.

Another aspect of dynamic data is distribution over the Web, and thus parallel processing as in the cloud. This raises the typical problems coming with data distribution: huge data sources cannot be moved without very high costs, while data must be replicated for providing efficient parallel access. This makes it difficult, if not impossible, to update replicated data consistently. Therefore, the consistency assumption has been removed by NoSQL databases for instance, while parallel algorithmic is limited to naive parallelization (i.e. map/reduce) where only few data needs to be exchanged.

We will investigate incremental query evaluation for distributed data-centered programming languages for linked data collections, dynamic updates as needed for linked data management, and static analysis for linked data workflows.

3.4. Linking Graphs

When datasets from independent sources are not linked with existing schema mappings, we would like to investigate symbolic machine learning solutions for inferring such mappings in order to define meaningful links between data from separate sources. This problem can be studied for various kinds of linked data collections. Before presenting the precise objectives, we will illustrate our approach on the example of linking data in two independent graphs: an address book of a research institute containing detailed personnel information and a (global) bibliographic database containing information on papers and their authors.

We remind that a schema allows to identify a collection of types each grouping objects from the same semantic class e.g., the collection of all persons in the address book and the collection of all authors in the bibliography database. As a schema is often lacking or underspecified in graph data models, we intend to investigate inference methods based on structural similarity of graph fragments used to describe objects from the same class in a given document e.g., in the bibliographic database every author has a name and a number of affiliations, while a paper has a title and a number of authors. Furthermore, our inference methods will attempt to identify, for every type, a set of possible keys, where by key we understand a collection of attributes of an object that uniquely identifies such an object in its semantic class. For instance, for a person in the address book two examples of a key are the name of the person and the office phone number of that person.

In the next step, we plan to investigate employing existing entity linkage solutions to identify pairs of types from different databases whose instances should be linked using compatible keys. For instance, persons in the address book should be linked with authors in the bibliographical database using the name as the compatible key. Linking the same objects (represented in different ways) in two databases can be viewed as an instance of a mapping between the two databases. Such mapping is, however, discriminatory because it typically maps objects from a specific subset of objects of given types. For instance, the mapping implied by linking persons in the address book with authors in the bibliographic database involves in fact researchers, a subgroup of personnel of the research institute, and authors affiliated with the research institute. Naturally, a subset of objects of a given type, or a subtype, can be viewed as a result of a query on the set of all objects, which on very basic level illustrates how learning data mappings can be reduced to learning queries.

While basic mappings link objects of the same type, more general mappings define how the same type of information is represented in two different databases. For instance, the email address and the postal address of an individual may be represented in one way in the address book and in another way in the bibliographic databases, and naturally, the query asking for the email address and the postal address of a person identified by a given name will differ from one database to the other. While queries used in the context of linking objects of compatible types are essentially unary, queries used in the context of linking information are n-ary and we plan to approach inference of general database mappings by investigating and employing algorithms for inference of n-ary queries.

An important goal in this research is elaborating a formal definition of *learnability* (feasibility of inference) of a given class of concepts (schemas of queries). We plan to following the example of Gold (1967), which requires not only the existence of an efficient algorithm that infers concepts consistent with the given input but the ability to infer every concept from the given class with a sufficiently informative input. Naturally, learnability depends on two parameters. The first parameter is the class of concepts i.e., a class of schema and a class of queries, from which the goal concept is to be inferred. The second parameter is the type of input that an inference algorithm is given. This can be a set of examples of a concept e.g., instances of RDF databases for which we wish to construct a schema or a selection of nodes that a goal query is to select. Alternatively, a more general interactive scenario can be used where the learning algorithm inquires the user about the goal concept e.g., by asking to indicate whether a given node is to be selected or not (as membership queries of Angluin (1987)). In general, the richer the input is, the richer class of concepts can be handled, however, the richer class of queries is to be handled, the higher computational cost is to be expected. The primary task is to find a good compromise and identify classes of concepts that are of high practical value, allow efficient inference with possibly simple type of input.

The main open problem for graph-shaped data studied by Links are how to infer queries, schemas, and schema-mappings for graph-structured data.

4. Application Domains

4.1. Linked Data Integration

There are many contexts in which integrating linked data is interesting. We advocate here one possible scenario, namely that of integrating business linked data to feed what is called Business Intelligence. The latter consists of a set of theories and methodologies that transform raw data into meaningful and useful information for business purposes (from Wikipedia). In the past decade, most of the enterprise data was proprietary, thus residing within the enterprise repository, along with the knowledge derived from that data. Today's' enterprises and businessmen need to face the problem of information explosion, due to the Internet's ability to rapidly convey large amounts of information throughout the world via end-user applications and tools. Although linked data collections exist by bridging the gap between enterprise data and external resources, they are not sufficient to support the various tasks of Business Intelligence. To make a concrete example, concepts in an enterprise repository need to be matched with concepts in Wikipedia and this can be done via pointers or equalities. However, more complex logical statements (i.e. mappings) need to be conceived to map a portion of a local database to a portion of an RDF graph, such as a subgraph in Wikipedia or in a social network, e.g. LinkedIn. Such mappings would then enrich the amount of knowledge shared within the enterprise and let more complex queries be evaluated. As an example, businessmen with the aid of business intelligence tools need to make complex sentimental analysis on the potential clients and for such a reason, such tools must be able to pose complex queries, that exploit the previous logical mappings to guide their analysis. Moreover, the external resources may be rapidly evolving thus leading to revisit the current state of business intelligence within the enterprise.

4.2. Data Cleaning

The second example of application of our proposal concerns scientists who want to quickly inspect relevant literature and datasets. In such a case, local knowledge that comes from a local repository of publications belonging to a research institute (e.g. HAL) need to be integrated with other Web-based repositories, such as DBLP, Google Scholar, ResearchGate and even Wikipedia. Indeed, the local repository may be incomplete or contain semantic ambiguities, such as mistaken or missing conference venues, mistaken long names for the publication venues and journals, missing explanation of research keywords, and opaque keywords. We envision a publication management system that exploits both links between database elements, namely pointers to external resources and logical links. The latter can be complex relationships between local portions of data and remote resources, encoded as schema mappings. There are different tasks that such a scenario could entail such as (i) cleaning the errors with links to correct data e.g. via mappings from HAL to DBLP for the publications errors, and via mappings from HAL to Wikipedia for opaque keywords, (ii) thoroughly enrich the list of publications of a given research institute, and (iii) support complex queries on the corrected data combined with logical mappings.

4.3. Real Time Complex Event Processing

Complex event processing serves for monitoring nested word streams in real time. Complex event streams are gaining popularity with social networks such as with Facebook and Twitter, and thus should be supported by distributed databases on the Web. Since this is not yet the case, there remains much space for future industrial transfer related to Links' second axis on dynamic linked data.

5. Highlights of the Year

5.1. Highlights of the Year

SheX

SHEX SCHEMAS FOR RDF GRAPHS IN COOPERATION WITH THE W3C

I. Boneva and S. Staworko present the RDF schema language SheX [22] in cooperation with members of the W3C. The usual open world approach of RDF is schemaless in the alphabets of RDF data are left open, so that data from different sources and with different alphabets can be unified. This raises serious problems for query writing and thus linked data integration, since the same query may become invalid when the alphabet changes. A SheX schema allows express constraints on the alphabets, node labels and edge labels of RDF graphs, so that databases queries become safe with respect to future changes, without that the alphabets need to be closed. This work is highly relevant for the future on data integration for RDF data based on schema mappings.

IJCAI

REASONABLE HIGHLY EXPRESSIVE QUERY LANGUAGES

In his IJCAI paper [17] P. Bourhis develops a highly expressive Web query language of the Datalog family, for which static analysis problems such as query containment remain decidable. The relevance of this result is explained to non-experts in a popularization article: http://www.cnrs.fr/ins2i/spip.php?article1465

5.1.1. Awards

This paper obtained the honorable mention of IJCAI .

IJCAI-highlight

LEARNING JOIN QUERIES FROM EXAMPLES

Ciucanu, A. Boneva, and S. Staworko published an article at ACM TODS [7], where they show how to learn join queries for relational databases from examples. The learning algorithm they provide is shown to satisfy Gold's learning model. Previously this model got applied only to inference of automata rather than logical queries. Furthermore, this is the first query learning algorithm that relies on equalities of data values rather than on the structure of metadata.

BEST PAPER AWARD:

[17]
P. BOURHIS, M. KRÖTZSCH, S. RUDOLPH. *Reasonable Highly Expressive Query Languages*, in "IJCAI", Buenos Aires, Argentina, July 2015, IJCAI-2015 Honorable Mention [*DOI :* 10.1007/978-3-662-47666-6_5], https://hal.inria.fr/hal-01211282

6. New Software and Platforms

6.1. QuiX Tool suite

KEYWORDS: XML - JSon - Xproc - XSLT - Schematron - Xquery - NoSQL
SCIENTIFIC DESCRIPTION

The QuiX-Tool Suite provides tools to process XML streams and documents. The QuiX-Tool Suite is based on early algorithms: query answers are delivered as soon as possible and in all practical cases at the earliest time point. The QuiX-Tool Suite provides an implementation of the main XML standart over streams. XPath, XSLT, XQuery and XProc are W3C standarts while Schematron is an ISO one. The QuiX-Tool suite is developed in the Inria transfer project QuiXProc in cooperation with Innovimax. It includes among the others existing tools such as FXP and QuixPath, along with new tools, namely X-Fun. Both, a free and a professional version are available. The ownership of QuiX-Tool Suite is shared between Inria and Innovimax. The main application of QuiX-Tool Suite is its usage in QuiXProc, an professional implementation of the W3C pipeline language XProc owned by Innovimax.

The QuiXPath language is a large fragment of XPath with full support for the XML data model. The QuiXPath library provides a compiler from QuiXPath to FXP, which is a library for querying XML streams with a fragment of temporal logic.

The X-Fun language is a functional language for defining transformations between XML data trees, while providing shredding instructions. X-Fun can be understood as an extension of Frisch's XStream language with output shredding, while pattern matching is replaced by tree navigation with XPath expressions. The QuiX-Tool suite includes QuiXSLT, which is a compiler from XSLT into a fragment of X-Fun, which can be considered as the core of XSLT. It also provides QuiXSchematron, which is a compiler from Schematron to X-Fun, and QuiXQuery, which is a compiler from XQuery to X-Fun.
FUNCTIONAL DESCRIPTION

QuiX Tool suite reads and processes large XML files without loading the entire file in main memory. Instead of building a tree representation of the XML document, QuiXProc manages data as streams (sequence of opening and closing tags).

- Participants: Joachim Niehren and Tom Sebastian
- Partner: Innovimax
- Contact: Joachim Niehren
- URL: https://project.inria.fr/quix-tool-suite/

6.2. SmartHal

FUNCTIONAL DESCRIPTION

SmartHal is a better tool for querying the HAL bibliography database, while is based on Haltool queries. The idea is that a Haltool query returns an XML document that can be queried further. In order to do so, SmartHal provides a new query language. Its queries are conjunctions of Haltool queries (for a list of laboratories or authors) with expressive Boolean queries by which answers of Haltool queries can be refined. These Boolean refinement queries are automatically translated to XQuery and executed by Saxon. A java application for extraction from the command line is available.

- Participants: Guillaume Bagan and Joachim Nierhen
- Contact: Joachim Niehren
- URL: http://smarthal.lille.inria.fr/

6.3. X-FUN

KEYWORDS: XML - Transformation - Functional programming - Compilers - Programming language
FUNCTIONAL DESCRIPTION

X-FUN is a core language for implementing various XML, standards in a uniform manner. X-Fun is a higher-order functional programming language for transforming data trees based on node selection queries.

- Participants: Pavel Labath and Joachim Niehren
- Contact: Joachim Niehren

7. New Results

7.1. Querying Heterogeneous Linked Data

7.1.1. Recursive queries

P. Bourhis published a paper at IJCAI [17] in cooperation with the University of Dresden in Germany. There he developed a highly expressive Web query language of the Datalog family, for which static analysis problems such as query containment remain decidable.

In cooperation with Links' associated team in Oxford, P. Bourhis obtained an article at ACM TODS [5], where he studies the access of hidden data by recursive queries.

V. Hugot, A. Boiret, and J. Niehren study monadic second-order logic for unordered trees with data constraints on siblings. This language can be used to define recursive queries and schemas on unordered data trees [13]. They study restrictions of the logics, for which the usual static analysis problems become decidable, and study the complexity of the decidable cases. This work was done in cooperation with Paris 7.

7.1.2. Schemas

I. Boneva and S. Staworko contribute at ICDT the RDF schema language SheX [22], which they developed in cooperation with members of the W3C. The usual open world approach of RDF is schemaless in that the alphabets of RDF data are left open, so that data from different sources and with different alphabets can be unified. This raises serious problems for query writing and thus for linked data integration, since a query may become invalid when the alphabet changes. A SheX schema allows to express constraints on the alphabets, node labels and edge labels of RDF graphs, so that database queries become safe with respect to future changes without closing the alphabet. In a previous work the studied the case of XML data trees instead of RDF graphs [6].

A. Lemay and J. Niehren propose sublinear algorithms in the style of probabilistic property testing for validating XML data trees with respect to DTD [20].

P. Bourhis studies streaming bounded repair with respect to schema violations [8]. This work is done in a cooperation with the University of Bordeaux and the University of Santiago in Chile.

7.1.3. Provenance

P. Bourhis obtained an ICALP paper [11] in cooperation with Télécom ParisTech. They show how to propagate provenance information for monadic second-order logics on trees or tree like structures with polynomial data complexity. In their provenance framework, they can show how to generalize various aggregation tasks for monadic second-order logics, that were known to be solvable with polynomal data complexity before.

In a cooperation with Tel Aviv, P. Bourhis obtained a ACM CIKM paper [18], where they show how to approximately summarize data provenance.

7.1.4. Data integration

In a cooperation with the University of Toronto, R. Ciucanu obtained a paper at PVLDB [4] on how to gain control over data integration evaluations. I. Boneva, A. Bonifati and R. Ciucaniu presented a paper on graph data exchange with target constraints [14] in the GraphQ workshop, and proved that query answering is intractable in this context.

7.2. Managing Dynamic Linked Data

7.2.1. Complex event processing

T. Sebastian, J. Niehren and D. Debarbieux propose early nested word automata for evaluating navigational XPath queries on XML streams [9]. They show how to approximate earliest query answering for such queries in a highly efficient manner and with very good precision in practice, while exact earliest query answering is known to untractable for XPath. This work was done in an industrial cooparation with Innovimax from Paris and in cooperation with the University of Bordeaux. In a follow-up work [21] they show that the `XPath` streaming algorithm for early nested word automata can be speed up considerably, when combining it with projection algorithms for nested word automata that they developed.

J. Niehren developed X-Fun [19] a uniform programming language for implementing XML standards, and showed how to implement XSLT, XProc, and XSLT in this manner. This work, that is fully implemented, was done in cooperation with the University of Bratislava.

7.2.2. Data-centered workflows

P. Bourhis presents highly expressive query languages as needed for data-centric workflows in the context of Active XML [3] in cooperation with the Dahu project from Inria Saclay.

J. Niehren presents a general framework for the reasoning with observational program semantics [10] in a cooperation with the Universities of Frankfurt and Saarbrücken in Germany.

7.3. Linking Data Graphs

S. Staworko obtained his HDR for his work on symbolic inference methods for databases [2]. R. Ciucanu obtained his PhD for his work on cross-model query inference [1] supervised bu A. Bonifati.

7.3.1. Learning path queries

A. Lemay, R. Ciucanu, and A. Bonifat have a paper and a demo at EDBT showing how to learn simple path queries on graph databases based on automata techniques [16], [15], [25], [24]. This is a very interesting starting point for using automata inference techniques in the context of graph databases.

S. Staworko obtained a paper at ICDT where he shows how to infer XML Twig queries from examples [23]. This work is done in cooperation with the University of Wrazlaw.

7.3.2. Learning join queries

R. Ciucanu, A. Boneva, and S. Staworko published an ACM TODS article [7] showing how to learn join queries for relational databases from examples. This is the first query learning algorithm satisfying Gold's learning model, that relies on equalities of data values rather than on structural information.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Grants with Industry

Innovimax is founding the PhD thesis of Tom SEBASTIAN (2011-15). The thesis is supervised by J.NIEHREN in cooperation with M.ZERGAOUI the head of the INNOVIMAX company. The software development in this context is supported by T. SEBASTIAN.

9. Partnerships and Cooperations

9.1. Regional Initiatives

Links participates in the CPER DATA (2015-19)

9.2. National Initiatives

ANR Aggreg (2014-19): Aggregation Queries.

- Particpants: J. Niehren [correspondent], P. Bourhis, A. Lemay, A. Boiret
- The coordinator is J. Niehren and the parteners are the University Paris 7 (A. Durand) including members of the Inria project DAHU (L. Ségoufin), the University of Marseille (N. Creignou) and University of Caen (E. Grandjean).
- Objective: the main goal of the Aggreg project is to develop efficient algorithms and to study the complexity of answering aggregate queries for databases and data streams of various kinds.

ANR Colis (2015-20): Correctness of Linux Scripts.

- Participants: J. Niehren [correspondent], A. Lemay, S. Tison, A. Boiret, V. Hugot.
- The coordinator is R. Treinen from the University of Paris 7 and the other partner is the Tocata project of Inria Saclay (C. Marché).

- Objective: This project aims at verifying the correctness of transformations on data trees defined by shell scripts for Linux software installation. The data trees here are the instance of the file system which are changed by installation scripts.

ANR DataCert (2015-20):

- Participants: I. Boneva [correspondent], A. Bonifati, S. Tison.
- Parteners: The coordinator is E. Contejean from the University of Paris Sud and the other partner is the University of Lyon.
- Objective: the main goals of the Datacert project are to provide deep specification in Coq of algorithms for data integration and exchange and of algorithms for enforcing security policies, as well as to design data integration methods for data models beyond the relational data model.

9.2.1. Competitivity Cluster Picom

FUI Hermes (2012-15): The future of shopping

- We participate in the Hermes project of the Pôle de compétitivité PICOM, a regional research cluster on the industy of commerce.
- Participants: I. Boneva [correspondent], A. Bonifati, J. Niehren
- Objective: Here we work on filtering publicity offers by newspaper arriving on complex event streams in real time.
- Parteners: Norsys, Auchan, etc

9.3. International Initiatives

9.3.1. Inria Associate Teams not involved in an Inria International Labs

Associated Team "Integrating Linked Data" with the Datbase group of the University of Oxford (2013-15).

9.3.2. Inria International Partners

9.3.2.1. Declared Inria International Partners

AMSud project "Foundations of Graph Databases" (2016-17)

Partners: Santiago de Chili (C. Riveros), Buenos Airos (S. Figuera), Bordeaux (G. Puppis).

9.4. International Research Visitors

9.4.1. Visits of International Scientists

George Fletcher, Eindhoven University of Technology, Belgium, Apr 2015

Martin Musicante, Federal University of Rio Grande do Norte, Bresil, Sep 2014- Oct 2015.

9.4.1.1. Internships

M. Linardi, University of Trento. On Web Data Integration, from Feb 2015 until Sep 2015.

9.4.2. Visits to International Teams

9.4.2.1. Research stays abroad

Slawek Staworko, University of Edinburgh, 2014-16.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific events selection

10.1.1.1. Member of conference program committees

J. Niehren was member of the program committees of LPAR (International Conference on Logic Programming and Automatic Reasonning) 2016, LATA (International Conference on Language and Automata Theory and Applications) 2015, and WPTE (Rewriting Techniques forProgram Transformations and Evaluation) 2015.

S. Tison is member of the program committes of FSCD (First International Conference on Formal Structures for Computation and Deduction) 2016, and CIAA (International Conference Implementation and Application of Automata) 2015.

S. Staworko is member of the program committes of PODS (ACM Symposium on Principles of Database Systems) 2016 and BICOD (British International Conference on Databases) 2015.

I. Boneva was member of program committee of EDBT (International Conference on Extending Database Technology) 2016 Vision Track.

P. Bourhis was member of program committee of RuleML (International Web Rule Symposium) 2015) and BDA (Journées Bases de données avancés) 2015

A. Bonifati was member of the program committees of VLDB (Very large Database Conference) 2015, ICDT (International Conference on Database Theory) 2015 and ICDE (International Conference of Data Engineering) 2015.

10.1.2. Journal

10.1.2.1. Member of the editorial boards

S. Tison is in the editorial committee of RAIRO-ITA (Theoretical Informatics and Applications).

J. Niehren is in the editoral board of Fundamenta Informaticae.

10.1.2.2. Reviewer - Reviewing activities

To many to be enumerated

10.1.3. Leadership within the scientific community

S. Tison is member of the “Comité National de la Recherche Scientifique (CoNRS)” (Section 6).

I. Boneva is a member of the Data Shapes Working Group of the W3C which the mission is to produce a language for defining structural constraints on RDF graphs. http://www.w3.org/2014/data-shapes/charter

10.1.4. Research administration

S. Tison has been an elected member of the Administrative committee of University of Lille 1 until October 2015. She is an elected member of the academic council of "ComUE Lille Nord de France " since November 2015.

S. Tison is a vice president of the University of Lille 1 since October 2015, where she is responsible for industrial partnerships, innovation, and valorisation.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

License: A. Bonifati, Introduction to Databases, 54h, L3, Université Lille 1, France

Master (Mocad): P. Bourhis Information extraction, 18h, M2, Université Lille 1, France

Master (Mocad): A. Bonifati Information extraction, 9h, M2, Université Lille 1, France

Master (Mocad): J. Niehren Information extraction, 9h, M2, Université Lille 1, France

Master : S. Tison Advanced algorithms and complexity, M1, 57h, Université Lille 1, France

Master : S. Tison, Fouille de données, M1, 30h, Université Lille 1, France

Master: A. Lemay, XML Technologies, 16h, M2, Université Lille 3, France

DUT : I. Boneva, 100h, Université Lillle 1, France

A. Bonifati was the co-responsable of the Master MOCAD of the University of Lille1.

A. Lemay is pedagogical responsable for Computer Science and numeric correspondant for UFR LEA Lille 3.

10.2.2. Supervision

PhD defended : R.CIUCANU, cross-model queries and schema mappings: evaluation and learning. Since Sept. 2012. Supervised by Bonifati.

PhD in progress : T. SEBASTIAN, Streaming algorithms for XPath. Since January 2011. Supervised by Niehren.

PhD in progress: A. BOIRET, Top-down tree transformations with look-ahead: foundations and learning. Since Sept. 2011. Supervised by Niehren and Lemay.

PhD in progress: D. Gallois started his PhD project on Recursive Queries under the supervison of Bourhis and Tison.

10.2.3. Juries

S. Tison was reviewer for the HDR of Sylvain Salvati at the University of Bordeaux, director of the HDR of Slawek Staworko at the University of Lille 1, and member of the HdR commitee of Pierre-Alain Reynier from the University of Marseille. She was also a member of the PhD committee fo Nadime Francis at the ENS Cachan.

10.2.4. Selection Committies

S. Tison was member of the selection committies for 2 professorships at University of Paris-Diderot, 1 assistant professorship at the Université Artois, and 2 assistant professorships at the University of Lille 3.

J. Niehren was member of the section committy for a professorship at the University of Bordeaux 1.

I. Boneva was member of the section committy for a assistant professorship at the University of Lens.

A. Lemay was member of the section committy for 2 assistant professorship at the University of Lille 3.

10.3. Popularization

Popularization on associate team Lille - Oxford: Popularization on associate team Lille - Oxford par Inria à http://www.inria.fr/centre/lille/actualites/lille-et-oxford-une-double-expertise-au-service-de-la-donnee.

Popularization on awarded ICJAI paper of P. Bourhis on "Reasonable Highly Expressive Query Languages". Une avancée en équivalence de requêtes pour interroger les données sur le web distinguée à IJCAI http://www.cnrs.fr/ins2i/spip.php?article1465

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] R. CIUCANU. *Cross-Model Queries and Schemas: Complexity and Learning*, Université Lille 1 - Sciences et Technologies, July 2015, https://hal.inria.fr/tel-01182649

[2] S. STAWORKO. *Symbolic Inference Methods for Databases*, Université de Lille, December 2015, Habilitation à diriger des recherches, https://hal.inria.fr/tel-01254965

Articles in International Peer-Reviewed Journals

[3] S. ABITEBOUL, P. BOURHIS, V. VIANU. *Highly Expressive Query Languages for Unordered Data Trees*, in "Theory of Computing Systems", 2015, 30 p. , https://hal.inria.fr/hal-01167068

[4] P. C. AROCENA, R. CIUCANU, B. GLAVIC, R. J. MILLER. *Gain Control over your Integration Evaluations*, in "Proceedings of the VLDB Endowment (PVLDB)", August 2015, vol. 8, n° 12, pp. 1960-1963, https://hal.inria.fr/hal-01187990

[5] M. BENEDIKT, P. BOURHIS, C. LEY. *Analysis of Schemas with Access Restrictions*, in "ACM Transactions on Database Systems", 2015, vol. 40, n° 1, 46 p. [*DOI :* 10.1145/2699500], https://hal.inria.fr/hal-01211288

[6] I. BONEVA, R. CIUCANU, S. STAWORKO. *Schemas for Unordered XML on a DIME*, in "Theory of Computing Systems", August 2015, vol. 57, n° 2, pp. 337–376 [*DOI :* 10.1007/s00224-014-9593-1], https://hal.inria.fr/hal-01076329

[7] A. BONIFATI, R. CIUCANU, S. STAWORKO. *Learning Join Queries from User Examples*, in "ACM Transactions on Database Systems", August 2015, vol. 40, n° 4, pp. 24:1–24:38, https://hal.inria.fr/hal-01187986

[8] P. BOURHIS, G. PUPPIS, C. RIVEROS. *Which XML Schemas are Streaming Bounded Repairable?*, in "Theory of Computing Systems", 2015 [*DOI :* 10.1007/s00224-015-9611-y], https://hal.inria.fr/hal-01211290

[9] D. DEBARBIEUX, O. GAUWIN, J. NIEHREN, T. SEBASTIAN, M. ZERGAOUI. *Early Nested Word Automata for XPath Query Answering on XML Streams*, in "Theoretical Computer Science", March 2015, n° 578, pp. 100-127, https://hal.inria.fr/hal-00966625

[10] M. SCHMIDT-SCHAUSS, D. SABEL, J. NIEHREN, J. SCHWINGHAMMER. *Observational Program Calculi and the Correctness of Translations*, in "Journal of Theoretical Computer Science (TCS)", April 2015, vol. 577, pp. 98–124 [*DOI :* 10.1016/j.tcs.2015.02.027], https://hal.inria.fr/hal-00824349

International Conferences with Proceedings

[11] A. AMARILLI, P. BOURHIS, P. SENELLART. *Provenance Circuits for Trees and Treelike Instances*, in "ICALP 2015", Kyoto, Japan, June 2015, pp. 56-68, https://hal-institut-mines-telecom.archives-ouvertes.fr/hal-01178399

[12] A. BOIRET. *Normal Form on Linear Tree-to-word Transducers*, in "10th International Conference on Language and Automata Theory and Applications", Prague, Czech Republic, J. JANOUŠEK, C. MARTÍN-VIDE (editors), March 2016, https://hal.inria.fr/hal-01218030

[13] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN. *Logics for Unordered Trees with Data Constraints on Siblings*, in "LATA : 9th International Conference on Language and Automata Theory and Applications", Nice, France, March 2015, pp. 175-187, https://hal.inria.fr/hal-01088761

[14] I. BONEVA, A. BONIFATI, R. CIUCANU. *Graph Data Exchange with Target Constraints*, in "EDBT/ICDT Workshops - Querying Graph Structured Data (GraphQ)", Bruxelles, Belgium, March 2015, pp. 171-176, https://hal.inria.fr/hal-01095838

[15] A. BONIFATI, R. CIUCANU, A. LEMAY. *Interactive Path Query Specification on Graph Databases*, in "18th International Conference on Extending Database Technology (EDBT)", Bruxelles, Belgium, March 2015, System Demo [*DOI :* 10.5441/002/DBT.2015.44], https://hal.inria.fr/hal-01097771

[16] A. BONIFATI, R. CIUCANU, A. LEMAY. *Learning Path Queries on Graph Databases*, in "18th International Conference on Extending Database Technology (EDBT)", Bruxelles, Belgium, March 2015 [*DOI :* 10.5441/002/EDBT.2015.11], https://hal.inria.fr/hal-01068055

[17] *Best Paper*
P. BOURHIS, M. KRÖTZSCH, S. RUDOLPH. *Reasonable Highly Expressive Query Languages*, in "IJCAI", Buenos Aires, Argentina, July 2015, IJCAI-2015 Honorable Mention [*DOI :* 10.1007/978-3-662-47666-6_5], https://hal.inria.fr/hal-01211282.

[18] A. ELEANOR, P. BOURHIS, S. B. DAVIDSON, D. DEUTCH, T. MILO. *Approximated Summarization of Data Provenance*, in "CIKM", Melbourn, Australia, October 2015, https://hal.inria.fr/hal-01211286

[19] P. LABATH, J. NIEHREN. *A Uniform Programming Language for Implementing XML Standards*, in "41st SOFSEM: International Conference on Current Trends in Theory and Practice of Computer Science", Pec pod Sněžkou, Czech Republic, Lecture Notes in Computer Science, Springer, January 2015, https://hal.inria.fr/hal-00954692

[20] A. M. NDIONE, A. LEMAY, J. NIEHREN. *Sublinear DTD Validity*, in "9th International Conference on. Language and Automata Theory and Applications", Nice, France, March 2015, https://hal.inria.fr/hal-00803696

[21] T. SEBASTIAN, J. NIEHREN. *Projection for Nested Word Automata Speeds up XPath Evaluation on XML Streams*, in "International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)", Harrachov, Czech Republic, January 2016, https://hal.inria.fr/hal-01182529

[22] S. STAWORKO, I. BONEVA, J. E. LABRA GAYO, S. HYM, E. G. PRUD'HOMMEAUX, H. SOLBRIG. *Complexity and Expressiveness of ShEx for RDF*, in "18th International Conference on Database Theory (ICDT 2015)", Brussels, Belgium, M. ARENAS, M. UGARTE (editors), 18th International Conference on Database Theory (ICDT 2015), March 2015 [*DOI :* 10.4230/LIPICS.ICDT.2015.195], https://hal.archives-ouvertes.fr/hal-01218552

[23] S. STAWORKO, P. WIECZOREK. *Characterizing XML Twig Queries with Examples*, in "International Conference on Database Theory", Brussels, Belgium, International Conference on Database Theory, March 2015, https://hal.inria.fr/hal-01205417

Conferences without Proceedings

[24] A. BONIFATI, R. CIUCANU, A. LEMAY. *Interactive Path Query Specification on Graph Databases*, in "31ème Conférence sur la Gestion de Données - Principes, Technologies et Applications - BDA 2015", Île de Porquerolles, France, September 2015, https://hal.inria.fr/hal-01187975

[25] A. BONIFATI, R. CIUCANU, A. LEMAY. *Learning Path Queries on Graph Databases*, in "BDA 2015 - 31ème Conférence sur la Gestion de Données - Principes, Technologies et Applications", Île de Porquerolles, France, September 2015, https://hal.inria.fr/hal-01187966

Other Publications

[26] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN. *Automata for Unordered Trees*, July 2015, Long version of GandALF'14 paper, https://hal.inria.fr/hal-01179493

[27] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN. *Logics for Unordered Trees with Data Constraints*, July 2015, 40 p. , Long version of LATA'15 paper, https://hal.inria.fr/hal-01176763

[28] I. BONEVA, J. E. LABRA GAYO, E. G. PRUD'HOMMEAUX, S. STAWORKO. *Shape Expressions Schemas*, October 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01218566

[29] G. LAURENCE, A. LEMAY, J. NIEHREN, S. STAWORKO, M. TOMMASI. *Sequential Tree-to-Word Transducers: Normalization, Minimization, and Learning*, August 2015, Long version of Lata 14 paper, https://hal.archives-ouvertes.fr/hal-01186993

[30] A. M. NDIONE, A. LEMAY, J. NIEHREN. *Sublinear DTD Validity*, August 2015, Long version of a Lata 15 paper, https://hal.inria.fr/hal-01187016