IN PARTNERSHIP WITH:
CNRS

Université Charles de Gaulle (Lille 3)

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2016

Project-Team LINKS

Linking Dynamic Data

IN COLLABORATION WITH: Centre de Recherche en Informatique, Signal et Automatique de Lille

RESEARCH CENTER
Lille - Nord Europe

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team LINKS

Creation of the Team: 2013 January 01, updated into Project-Team: 2016 June 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
2.4. - Verification, reliability, certification
2.4.1. - Analysis
2.4.2. - Model-checking
2.4.3. - Proofs
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
3.2.1. - Knowledge bases
3.2.2. - Knowledge extraction, cleaning
3.2.3. - Inference
3.2.4. - Semantic Web
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:

6.1. - Software industry
6.3.1. - Web
6.3.4. - Social Networks
6.5. - Information systems
9.4.1. - Computer science
9.4.5. - Data science
9.8. - Privacy

1. Members

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

Faculty Members
Iovka Boneva [Univ. Lille I, Associate Professor]
Aurélien Lemay [Univ. Lille III, Associate Professor]
Sylvain Salvati [Univ. Lille I, Professor, HDR]
Slawomir Staworko [Univ. Lille III, Associate Professor, HDR]
Sophie Tison [Univ. Lille I, Professor, HDR]

PhD Students
Adrien Boiret [Univ. Lille I]
Dimitri Gallois [Univ. Lille I]
José Martin Lozano [Univ. Lille I, from Oct 2016]
Momar Sakho [Inria, from Feb 2016]
Tom Sebastian [Univ. Lille I]

Post-Doctoral Fellows
Vincent Hugot [Univ. Lille I, Research Scientist]
Nicolas Bacquey [Inria, from Sep 2016]

Visiting Scientist
Domagoj Vrgoc [from Aug 2016 until Sep 2016]

Administrative Assistant
Aurore Hermant [Inria]

2. Overall Objectives

2.1. Overall Objectives

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 inferring appropriate queries and mappings.

2.2. Presentation

The following three paragraphs summarise 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 parallelisation (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

Certain Query Answering as Access Control

P. Bourhis [24] presented at **LICS** — the top conference in logic in computer science — a general framework for querying databases with visible and invisible relations. This work was done in cooperation with Oxford, Santa Cruez, and Bordeaux. It generalizes in a uniform manner the problems of certain query answering and access control for relational databases. Invisible relations are subject to the open world assumption possibly under constraints, while visible relations are subject to the closed world assumption. Bourhis then shows that the problem of answering Boolean conjunctive queries in this framework is decidable, and studies the complexity of various versions of this problem. It turns out that the complexity increases compared to the problem of certain query answering, given that the closed world assumption is adopted for the added visible relations.

Five ANR Projects

Two new ANR projects were accepted this year: *Delta* and *Headwork*. This makes Links a partner of 5 ANR projects in 2016.

PhD Defense of A. Boiret

The defense of the PhD thesis of A. Boiret [11] on "Normalization and Learning of Transducers on Trees and Words" under the supervision of J. Niehren and A. Lemay was highly appreciated by the reviewers. In particular, he illustrated very clearly how to learn top-down tree transformations subject to regular schema restriction [31], [33], [34]. Furthermore, he solve a problem open for more than 20 years on how to learn rational functions, i.e. word transducers with regular lookahead.

6. New Software and Platforms

6.1. ShEx Validator

KEYWORDS: RDF Data management - RDF - Shape Expression
FUNCTIONAL DESCRIPTION

Shape Expression schemas is a formalism for defining constraints on RDF graphs. This software allows to check whether a graph satisfies a Shape Expressions schema.

- Participants: Iovka Boneva
- Contact: Iovka Boneva
- URL: https://gforge.inria.fr/projects/shex-impl

6.2. gMark

KEYWORDS: graph benchmark - Graph Database - Graph Query
FUNCTIONAL DESCRIPTION

gMark allow the generation of graph databases and an associated set of query from a schema of the graph. gMark is based on the following principles: great flexibility in the schema definition, ability to generate big size graphs, ability to generate recursive queries and queries with a desired selectivity .

- Participants: Aurélien Lemay
- Contact: Aurélien Lemay
- URL: https://github.com/graphMark/gmark

6.3. QuiXPath

KEYWORDS: XML Streams - XPath 3.0 Queries - Aggregation - Data Joins
FUNCTIONAL DESCRIPTION

QuiXPath is a streaming implementation that covers most of XPath 3.0. It was developed during the PhD thesis of T. Sebastian funded by our industrial transfer partner Innovimax.

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

6.4. 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. Provenance

The computation of the provenance of a query answer is a classical problem in database theory. It consists in aggregating the impact of tuples of a database to a query answer. This allows to give an explanation of the query answers, that can help to judge their reliability. The computation of the provenance of a query answer is thus an aggregation problem as studied by the ANR project *Aggreg* .

P. Bourhis [20] showed at **PODS** — the top conference on database theory — that the lineage of MSO queries on treelike database instances is tractable, but not on other instances. This work was in cooperation with Telecom ParisTech and ENS Paris. As a first application, he can show that MSO query evaluation on probabilistic databases is tractable for tree like database instances, but not otherwise.

P. Bourhis applied in cooperation with Tel Aviv, provenance problems to recommendation systems. This allows to explain the end result by summarising with similar data without changing significantly results obtained in general by aggregation on the data. The corresponding tool was demonstrated at **EDBT** [32].

7.1.2. Certain Query Answering and Access Control

The problem of certain query answering consists in finding which are the certain answer of a query in a database with incomplete data, and a set of constraints representing available the knowledge on the incomplete data.

P. Bourhis [24] presented at **LICS** — the top conference in logic in computer science — a general framework for querying databases with visible and invisible relations. This work was done in cooperation with Oxford, Santa Cruez, and Bordeaux. His framework is motivated by the problem of access control for relational databases, i.e. of data leakage in relational views, but generalizes at the same time the problem of certain query answering. Invisible relations are subject to the open world assumption possibly under constraints as usual in certain query answering, while visible relations are subject to the closed world assumption. Bourhis then show that it is decidable, whether a conjunctive has an answer in this framework, when given the visible relation, the constraints, and the query as inputs. He also studies the complexity of this problem. It turns out the complexity increases from polynomial to doubly exponential, compared to certain query answering, since adding visible relations subject to the closed world assumption.

P. Bourhis studied at **IJCAI** [19] certain query answering with some transitive closure constraints, which allow to define a constraints with recursion. This work was done in collaboration with Oxford and Telecom ParisTech.

The problem of ontological query containment consists in establishing whether the certain answers of two queries subject to an ontology are included in each other. P. Bourhis [26] studied at **KR** this problem for several closely related formalisms: monadic disjunctive Datalog (MDDLog), MMSNP (a logical generalization of constraint satisfaction problems) and ontology-mediated queries (OMQs). This work was done in cooperation with Bremen.

7.1.3. Recursive Queries

At **LICS** [21] again, P. Bourhis showed in collaboration with Oxford how to lift a major restriction on decidable fixpoint logics that can define recursive queries (such as C2RPQs), specifically on guarded logic. This allows to improve significantly expressiveness of decidable fixpoint logics.

A. Lemay contributed at **TKDE** [14] the *gMark* benchmark, a tool to generate large size graph database and an associated set of queries. This work was done in cooperation with Eindhoven and previous members of Links that are now in Lyon and Clérmont-Ferrant. The tool was also demonstrated at **VLDB** [13]. Its main interest is a great flexibility (the generation of the graph can be done from a simple schema, but can also incorporate elaborate a parameters), an ability to generate recursive queries, and the possibility to generate large sets of queries of a desired selectivity. This benchmark allowed for instance to highlight difficulties for the existing query engines to deal with recursive queries of high selectivity.

7.1.4. Data Integration

P. Bourhis and S. Staworko in cooperation with Bordeaux and Oxford presented at **TODS** [17] their work on bounded repairability for regular tree languages, which is a study on whether a tree document (typically XML) can be repaired to fit a given target tree language within a bounded amount of tree editing operations. The article studies the complexity of different classes of tree languages such as non-recursive DTDs, recursive DTDs, or languages by arbitrary bottom-up tree automaton.

J.M. Lozano started his PhD project under the supervision of I. Boneva and S. Staworko. His topic subscribes the ANR project *Datacert* on data integration and certification.

7.1.5. Schema Validation

A. Boiret, V. Hugot and J. Niehren studied schemas for JSON documents in **Information and Computation** [15]. This work was done in collaboration with Paris 7. A JSON document is an unordered data trees, so schemas for such documents are best seen as automata for unordered data trees. The paper generalizes several previous formalisms for automata on unordered trees in a uniform framework. Whether the equivalence of two schemas can be tested in P-time is studied for various instances of the framework.

This work subscribes to the ANR project *Colis* where unranked data trees are used as models of linux file systems. In this context, N. Bacquey started his postdoc on the verification of linux installation scripts.

7.2. Managing Dynamic Linked Data

7.2.1. Complex Event Processing

Complex event processing can be seen as the problem is to answer queries on data graphs, for graphs that arrive on streams. These queries may contain aggregates, so this work subscribes to the ANR project *Aggreg*.

In his PhD thesis, T. Sebastian [12] developed with his supervisor J. Niehren streaming algorithms covering all of XPath 3.0 queries on `XML` streams. For this, they proposed a higher-order query language λXP, showed how to give a formal semantics of all of XPath 3.0 by compilation to λXP, and then how to evaluate λXP queries on `XML` streams. These algorithms were implemented in the QuiXPath tool.

At **SOFSEM**, they proposed a new technique to speed up the evaluation of navigational XPath queries on XML streams based on document projection. The idea is to skip those parts of the stream that are irrelevant for the query. This speeds up the evaluation of navigation XPath queries by a factor of 4 in usual Xpath benchmarks.

M. Sakho started his PhD project on hyperstreaming query answering algorithms for graphs under the supervision of J. Niehren and I. Boneva. Part of this work will be continued with out visitor D. Vrgoc from Santiago di Chili.

7.2.2. Data Centric Workflows

Data-centric workflows are complex programs that can query and update a database. The usage of data-centric workflows for crowd sourcing is the topic of the ANR Project *HeadWork*.

In collaboration with ENS Cachan and San Diego, P. Bourhis presented at **ICDT** [18] techniques on collaborative access control in a distributed query and data exchange language (Webdamlog). The goal of this work was to provide a semantic to data exchange rules defined by Webdamlog. It also allowed to prove that it is possible to formally verify whether there are data leakages.

P. Bourhis with Tel Aviv defined at **ICDE** [25] a notion of provenance for data-centric workflows, and proved that it can be used to explain the provenance of fact in the final instance of an execution. This provenance is used to answer three main questions: *why* does a specific tuple appear in the answer of a query, *what if* the initial database is changed (Revision problem), and *how to* change the query to obtain a missing tuple.

7.3. Linking Data Graphs

7.3.1. Learning Transformations

We consider the problem to learn queries and query-based transformations on semi-structured data from examples.

A. Boiret obtained his PhD for his work on the "Normalization and Learning of Transducers on Trees and Words" under the supervision of J. Niehren and A. Lemay. In this year, he showed how to learn top-down tree transformations with regular schema restrictions [31], [33], [34]. At **LATA** [22], he deepened a result of a previous PhD student of Links on learning sequential tree-to-word transducers (with output concatenation), by showing who to find normal forms for less restrictive linear tree-to-word transducers. At **DLT** [23], he could show in cooperation with Munich, that the equivalence problem of this class of transducers is in polynomial time, even though their normal forms may be of exponential size.

In the context of learning RDF graph transformations, S. Staworko presented a cooperation with Edinburg at **VLDB** [27]. Using bisimulation technique, he aims at aligning datas of two RDF Graphs that takes into account blank velues, changes in ontology and small differences in data values and in the structure of the graph. the alignement of graphs is an important first step for the inference of transformations.

7.3.2. Learning Join Queries

S. Staworko published in **TODS** an article [16] on learning join queries from user examples in collaboration with Universities of Lyon and Clermont-Ferrand that present techniques that allow the automatic construction of a join query through interaction with a user that simply labels sets of tuples to indicate whether the tuple is in the target query or not.

8. Partnerships and Cooperations

8.1. Regional Initiatives

Links participates in the CPER DATA (2015-19)

8.2. National Initiatives

ANR Aggreg (2014-19): Aggregated 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], S. Tison, J. Lozano.
- Partners: 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.

ANR Headwork (2016-21):

- Participants: P. Bourhis [correspondant], J. Niehren, M. Sakho.
- Scientific partners: The coordinateur is D. Gross-Amblard from the Druid Team (Rennes 1). Other partners include the Dahu team (Inria Saclay) and Sumo (Inria Bretagne)
- Industrial partners: Spipoll, and Foulefactory.
- Objective: The main object is to develop data-centric workflows for programming crowd sourcing systems in flexible declarative manner. The problem of crowd sourcing systems is to fill a database with knowledge gathered by thousands or more human participants. A particular focus is to be put on the aspects of data uncertainty and for the representation of user expertise.

ANR Delta (2016-21):

- Participants: P. Bourhis [correspondent], D. Gallois.
- Partners: The coordinator is M. Zeitoun from LaBRI, other partners are LIF (Marseille) and IRIF (Paris-Diderot).
- Objective: Delta is focused on the study of logic, transducers and automata. In particular, it aims at extending classical framework to handle input/output, quantities and data.

8.3. International Initiatives

8.3.1. Inria International Partners

8.3.1.1. Declared Inria International Partners

AMSud project “Foundations of Graph Databases” (2015-16)

Partners: Chili (C. Riveros), Buenos Aires (Figueira), Bordeaux (G. Puppis).

8.4. International Research Visitors

8.4.1. Visits of International Scientists

Domagoj Vrgoc, DCC PUC Chile, From Aug 2016 Until Sep 2016

8.4.2. Visits to International Teams

8.4.2.1. Research Stays Abroad

Slawek Staworko, University of Edinburgh, 2014-16.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Selection

9.1.1.1. Member of the Conference Program Committees

J. Niehren was member of the program committees of LPAR (International Conference on Logic Programming and Automatic Reasoning) 2016.

S. Tison is member of the program committees of FSCD (First International Conference on Formal Structures for Computation and Deduction) 2016.

S. Staworko is member of the program committes of PODS (ACM Symposium on Principles of Database Systems) 2016.

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 Provenance week 2016, IJCAI (International Joint Conference on Artificial Intelligence) 2016.

9.1.2. Journal

9.1.2.1. 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.

9.1.2.2. Reviewer - Reviewing Activities

Too many to be enumerated.

9.1.3. Leadership within the Scientific Community

S. Tison has been member of the “Comité National de la Recherche Scientifique (CoNRS)” (Section 6) until June 2016.

9.1.4. Scientific Expertise

P. Bourhis has expertised a project in the ’Recherche Formation Innovation Atlanstic’ program of Pays de la Loire Region.

S. Tison has been a member of the scientific board of the company See-d

9.1.5. Research Administration

S. Tison 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.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Bachelor : S. Tison, Discrete Mathematics, 36h, Université Lille 1, France

Bachelor : S. Salvati, Computer Science Info, L1, 96h, Université de Lille 1, France

Bachelor : S. Salvati, Automata and Languages, L3, 36h, Université de Lille 1, France

Bachelor : S. Salvati, Algorithmic and Operational Research, L3, 36h, Université de Lille 1, France

Master 2 (MOCAD): P. Bourhis and J. Niehren Information extraction, 20h30, M2, Université Lille 1, France

Master 2 (MOCAD) : I. Boneva and P. Bourhis, Algorithms for Database, 21h, M1, Université Lille 1, France

Master 1 : S. Tison, Advanced algorithms and complexity, M1, 54h, Université Lille 1, France

Master 1 : S. Tison, Business Intelligence, M1, 28h, Université Lille 1, France

Master 1 : A. Lemay, XML Technologies, 16h, M2, Université Lille 3, France

Master 1 : S. Staworko is co-head of the master ’Web Analyst’ in Université Lille 3, France

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

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

9.2.2. Supervision

PhD defended : T. Sebastian, Streaming algorithms for XPath. 2011-2016. Supervised by Niehren.

PhD defended: A. Boiret. Normalization and Learning of Transducers on Trees and Words. 2011-2016. Supervised by Niehren and Lemay.

PhD in progress: D. Gallois. Since 2015. Recursive Queries. Supervised by Bourhis and Tison.

PhD in progress: M. Sakho. Hyperstreaming Query answering on graphs. Since 2016. Supervised by Niehren and Boneva.

PhD in progress: J.M. Lozano. On data integration for mixed database formats. Supervised by Boneva and Staworko.

9.2.3. Juries

P. Bourhis was a member of PhD committee for Antoine Amarilli (ParisTech)

S. Tison was a member of the PhD committees for Yoann Dufresne and A. Boiret and reviewer for the PhD of Carles Creus (Universitat Politecnica de Catalunya).

A. Lemay was a member of the PhD committee for A. Boiret

J. Niehren was a member of PhD committees for A. Boiret, T. Sebastian and Louis Gallaraga (ParisTech).

9.2.4. Internships

A. Durey, University of Lille I. On Implemetation of a Shex validator, from Jun 2016 until Sep 2016. Supervison I. Boneva.

R. Li, Ecole Centrale de Lille, from May 2016 until Aug 2016. On aggregate queries for data mining. Supervisors J. Niehren and P. Bourhis.

9.2.5. Selection Committies

I. Boneva was member of the selection committee for an assistant professorship at the University of Lille I.

S. Tison was member of the selection committee for a professorship at the University of Rouen.

9.3. Popularization

I. Boneva has supervised second year students from DUT while they conducted activities on introduction to programming to 9-10 years old students in Villeneuve d'Ascq.

9.4. Standardization

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. Bibliography

Major publications by the team in recent years

[1] A. AMARILLI, P. BOURHIS, P. SENELLART. *Tractable Lineages on Treelike Instances: Limits and Extensions*, in "PODS (Principles of Database Systems)", San Francisco, United States, June 2016, pp. 355-370, https://hal-institut-mines-telecom.archives-ouvertes.fr/hal-01336514

[2] M. BENEDIKT, P. BOURHIS, M. VANDEN BOOM. *A Step Up in Expressiveness of Decidable Fixpoint Logics*, in "Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science", New York, United States, July 2016, https://hal.inria.fr/hal-01413890

[3] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN. *Automata for Unordered Trees*, in "Information and Computation", July 2016 [*DOI :* 10.1016/J.IC.2016.07.012], https://hal.inria.fr/hal-01179493

[4] A. BONIFATI, R. CIUCANU, S. STAWORKO. *Learning Join Queries from User Examples*, in "ACM Transactions on Database Systems", February 2016, vol. 40, n^{o} 4, pp. 24:1–24:38, https://hal.inria.fr/hal-01187986

[5] P. BOURHIS, M. BENEDIKT, B. TEN CATE, G. PUPPIS. *Querying Visible and Invisible Information*, in "LICS 2016 - 31st Annual ACM/IEEE Symposium on Logic in Computer Science", New York City, United States, July 2016, pp. 297-306 [*DOI :* 10.1145/2933575.2935306], https://hal.archives-ouvertes.fr/hal-01411118

[6] 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

[7] P. BOURHIS, C. LUTZ. *Containment in Monadic Disjunctive Datalog, MMSNP, and Expressive Description Logics*, in "Principles of Knowledge Representation and Reasoning", Cape Town, South Africa, April 2016, https://hal.inria.fr/hal-01413887

[8] P. BUNEMAN, S. STAWORKO. *RDF Graph Alignment with Bisimulation*, in "VLDB 2016 - 42nd International Conference on Very Large Databases", New Dehli, India, Proceedings of the VLDB Endowment, September 2016, vol. 9, n° 12, pp. 1149 - 1160 [*DOI :* 10.14778/2994509.2994531], https://hal.inria.fr/hal-01417156

[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] 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

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] A. BOIRET. *Normalization and Learning of Transducers on Trees and Words*, Université de Lille, November 2016, https://tel.archives-ouvertes.fr/tel-01396543

[12] T. SEBASTIAN. *Evaluation of XPath Queries on XML Streams with Networks of Early Nested Word Automata*, Universite Lille 1, June 2016, https://hal.inria.fr/tel-01342511

Articles in International Peer-Reviewed Journals

[13] G. BAGAN, A. BONIFATI, R. CIUCANU, G. FLETCHER, A. LEMAY, N. ADVOKAAT. *Generating Flexible Workloads for Graph Databases*, in "Proceedings of the VLDB Endowment (PVLDB)", June 2016, https://hal.inria.fr/hal-01330111

[14] G. BAGAN, A. BONIFATI, R. CIUCANU, G. FLETCHER, A. LEMAY, N. ADVOKAAT. *gMark: Schema-Driven Generation of Graphs and Queries*, in "IEEE Transactions on Knowledge and Data Engineering", November 2016, https://hal.inria.fr/hal-01402575

[15] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN. *Automata for Unordered Trees*, in "Information and Computation", July 2016 [*DOI :* 10.1016/J.IC.2016.07.012], https://hal.inria.fr/hal-01179493

[16] A. BONIFATI, R. CIUCANU, S. STAWORKO. *Learning Join Queries from User Examples*, in "ACM Transactions on Database Systems", February 2016, vol. 40, n° 4, pp. 24:1–24:38, https://hal.inria.fr/hal-01187986

[17] P. BOURHIS, C. RIVEROS, S. STAWORKO, G. PUPPIS. *Bounded Repairability for Regular Tree Languages*, in "ACM Transactions on Database Systems", June 2016, vol. 41, n° 3, pp. 1-45 [*DOI :* 10.1145/2898995], https://hal.archives-ouvertes.fr/hal-01411116

International Conferences with Proceedings

[18] S. ABITEBOUL, P. BOURHIS, V. VIANU. *A formal study of collaborative access control in distributed datalog*, in "ICDT 2016 - 19th International Conference on Database Theory", Bordeaux, France, W. MARTENS, T. ZEUME (editors), March 2016, https://hal.inria.fr/hal-01290497

[19] A. AMARILLI, M. BENEDIKT, P. BOURHIS, M. VANDEN BOOM. *Query Answering with Transitive and Linear-Ordered Data*, in "Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016", New York, United States, July 2016, https://hal.inria.fr/hal-01413881

[20] A. AMARILLI, P. BOURHIS, P. SENELLART. *Tractable Lineages on Treelike Instances: Limits and Extensions*, in "PODS (Principles of Database Systems)", San Francisco, United States, June 2016, pp. 355-370, https://hal-institut-mines-telecom.archives-ouvertes.fr/hal-01336514

[21] M. BENEDIKT, P. BOURHIS, M. VANDEN BOOM. *A Step Up in Expressiveness of Decidable Fixpoint Logics*, in "Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science", New York, United States, July 2016, https://hal.inria.fr/hal-01413890

[22] 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

[23] A. BOIRET, R. PALENTA. *Deciding Equivalence of Linear Tree-to-Word Transducers in Polynomial Time*, in "Developments in Language Theory", Montreal, Canada, Lecture Notes in Computer Science, Springer, July 2016, vol. 9840, pp. 355-367, https://hal.archives-ouvertes.fr/hal-01429110

[24] P. BOURHIS, M. BENEDIKT, B. TEN CATE, G. PUPPIS. *Querying Visible and Invisible Information*, in "LICS 2016 - 31st Annual ACM/IEEE Symposium on Logic in Computer Science", New York City, United States, July 2016, pp. 297-306 [*DOI :* 10.1145/2933575.2935306], https://hal.archives-ouvertes.fr/hal-01411118

[25] P. BOURHIS, D. DEUTCH, Y. MOSKOVITCH:. *Analyzing data-centric applications: Why, what-if, and how-to.* , in "32nd IEEE International Conference on Data Engineering, ICDE 2016", Helsinki, Finland, May 2016, https://hal.inria.fr/hal-01413879

[26] P. BOURHIS, C. LUTZ. *Containment in Monadic Disjunctive Datalog, MMSNP, and Expressive Description Logics*, in "Principles of Knowledge Representation and Reasoning", Cape Town, South Africa, April 2016, https://hal.inria.fr/hal-01413887

[27] P. BUNEMAN, S. STAWORKO. *RDF Graph Alignment with Bisimulation*, in "VLDB 2016 - 42nd International Conference on Very Large Databases", New Dehli, India, Proceedings of the VLDB Endowment, September 2016, vol. 9, n° 12, pp. 1149 - 1160 [*DOI :* 10.14778/2994509.2994531], https://hal.inria.fr/hal-01417156

[28] 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

Conferences without Proceedings

[29] G. BAGAN, A. BONIFATI, R. CIUCANU, G. FLETCHER, A. LEMAY, N. ADVOKAAT. *Génération de Requêtes pour les Bases de Données Orientées Graphes*, in "32ème Conférence sur la Gestion de Données - Principes, Technologies et Applications - BDA 2016", Futuroscope, Poitiers, France, November 2016, https://hal.inria.fr/hal-01402582

[30] G. BAGAN, A. BONIFATI, R. CIUCANU, G. FLETCHER, A. LEMAY, N. ADVOKAAT. *gMark : Génération de Graphes et de Requêtes Dirigée par le Schéma*, in "32ème Conférence sur la Gestion de Données - Principes, Technologies et Applications - BDA 2016", Futuroscope, Poitiers, France, November 2016, https://hal.inria.fr/hal-01402580

[31] A. BOIRET, A. LEMAY, J. NIEHREN. *Learning Top-Down Tree Transducers with Regular Domain Inspection*, in "International Conference on Grammatical Inference 2016", Delft, Netherlands, October 2016, https://hal.inria.fr/hal-01357186

Other Publications

[32] E. AINY, P. BOURHIS, S. B. DAVIDSON, D. DEUTCH, T. MILO. *PROX: Approximated Summarization of Data Provenance*, March 2016, International Conference on Extending Database Technology, Poster - Démonstration, https://hal.inria.fr/hal-01420452

[33] A. BOIRET, A. LEMAY, J. NIEHREN. *A Learning Algorithm for Top-Down Tree Transducers*, August 2016, working paper or preprint, https://hal.inria.fr/hal-01357627

[34] A. BOIRET, A. LEMAY, J. NIEHREN. *Learning Top-Down Tree Transformations with Regular Inspection*, August 2016, working paper or preprint, https://hal.inria.fr/hal-01357631

[35] I. BONEVA. *Comparative expressiveness of ShEx and SHACL (Early working draft)*, March 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01288285