IN PARTNERSHIP WITH:
CNRS

Université Charles de Gaulle (Lille 3)

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2017

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:

A2.1. - Programming Languages
A2.1.1. - Semantics of programming languages
A2.1.3. - Functional programming
A2.1.6. - Concurrent programming
A2.4. - Verification, reliability, certification
A2.4.1. - Analysis
A2.4.2. - Model-checking
A2.4.3. - Proofs
A3.1. - Data
A3.1.1. - Modeling, representation
A3.1.2. - Data management, quering and storage
A3.1.3. - Distributed data
A3.1.4. - Uncertain data
A3.1.5. - Control access, privacy
A3.1.6. - Query optimization
A3.1.7. - Open data
A3.1.8. - Big data (production, storage, transfer)
A3.1.9. - Database
A3.2.1. - Knowledge bases
A3.2.2. - Knowledge extraction, cleaning
A3.2.3. - Inference
A3.2.4. - Semantic Web
A4.7. - Access control
A4.8. - Privacy-enhancing technologies
A7. - Theory of computation
A7.2. - Logic in Computer Science
A9.1. - Knowledge
A9.2. - Machine learning
A9.7. - AI algorithmics

Other Research Topics and Application Domains:

B6.1. - Software industry
B6.3.1. - Web
B6.3.4. - Social Networks
B6.5. - Information systems
B9.4.1. - Computer science
B9.4.5. - Data science
B9.8. - Privacy

1. Personnel

Research Scientists

Joachim Niehren [Team leader, Inria, Senior Researcher, HDR]
Pierre Bourhis [CNRS, Researcher, until Sep 2017]

Faculty Members

Iovka Boneva [Université Lille 1, Associate Professor]
Aurélien Lemay [Université Lille 3, Associate Professor]
Sylvain Salvati [Université Lille 1, Professor, HDR]
Slawomir Staworko [Université Lille 3, Associate Professor, HDR]
Sophie Tison [Université Lille 1, Professor, HDR]
Florent Capelli [Université Lille 3, Associate Professor, from Sep 2017]
Charles Paperman [Université Lille 3, Associate Professor, from Sep 2017]

Post-Doctoral Fellows

Vincent Hugot [Université de Lille 1, Postdoc, until Aug 2017]
Nicolas Bacquey [Inria]
Adrien Boiret [Université Lille 1, until Aug 2017]

PhD Students

Dimitri Gallois [Université Lille 1]
Paul Gallot [Inria, from Oct 2017]
Jose-Martin Lozano [Université Lille 1]
Momar Sakho [Inria]

Administrative Assistants

Nathalie Bonte [Inria, from Oct 2017]
Aurore Hermant [Inria, until Juli 2017]

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. Book with the W3C on schemas validation for the semantic Web

I. Boneva et al. published a book [25] on Validating RDF Data based on the schema language ShEx. This book may have a important impact on the semantic Web community, given that one of her co-authors works for the W3C (E. Prud'hommeaux) where ShEx is considered for standardization.

5.2. Two associate professors recruited

F. Capelli and C. Paperman were hired as Associate Professors for LINKS by the University of Lille 3, so we are currently working on their integration.

5.3. Papers at PODS, LICS, 3 x ICALP, STACS, 2 x IJCAI

This year we obtained exceptional publications in all main theory conferences concerning databases, logic, and artificial intelligence.

5.4. ICALP best paper award

Pierre Bourhis' paper with Oxford (Benedikt and Vanden Boom) in Track B of ICALP'17 won the best paper award!

BEST PAPER AWARD:

[17]
M. BENEDIKT, P. BOURHIS, M. V. BOOM. *Characterizing Definability in Decidable Fixpoint Logics* , in "ICALP 2017 - 44th International Colloquium on Automata, Languages, and Programming", Varsovie, Poland, I. CHATZIGIANNAKIS, P. INDYK, F. KUHN, A. MUSCHOLL (editors), July 2017, vol. 107, 14 p. [*DOI :* 10.4230/LIPICS.ICALP.2017.107], https://hal.inria.fr/hal-01639015

6. New Software and Platforms

6.1. ShEx validator

Validation of Shape Expression schemas
KEYWORDS: Data management - RDF
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.

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

6.2. gMark

gMark: schema-driven graph and query generation
KEYWORDS: Semantic Web - Data base
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 - ability to generate queries with a desired selectivity

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

6.3. SmartHal

KEYWORD: Bibliography
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. On top of this, we have build a tool for producing the citation lists for the evaluation report of the LIFL, which can be easily adapter to other Labs.

- Contact: Joachim Niehren
- URL: http://smarthal.lille.inria.fr/

6.4. QuiXPath

KEYWORDS: XML - NoSQL - Data stream
SCIENTIFIC DESCRIPTION: The QuiXPath tools supports a very large fragment of XPath 3.0. The QuiXPath library provides a compiler from QuiXPath to FXP, which is a library for querying XML streams with a fragment of temporal logic.
FUNCTIONAL DESCRIPTION: QuiXPath is a streaming implementation of XPath 3.0. It can query large XML files without loading the entire file in main memory, while selecting nodes as early as possible.

- Contact: Joachim Niehren
- URL: https://project.inria.fr/quix-tool-suite/

6.5. X-FUN

KEYWORDS: Programming language - Compilers - Functional programming - Transformation - XML
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: Joachim Niehren and Pavel Labath
- Contact: Joachim Niehren

7. New Results

7.1. Querying Heterogeneous Linked Data

7.1.1. Aggregates

Aggregation refers to the computation of aggregates in databases, that is, the computation of a function of the answer of a query, such as counting the number of answers, finding the optimal one for a given objective function or enumerating all of them with a small delay between two distinct answers. The goal of aggregation is typically to compute such aggregates without explicitly generating the whole set of answers. We study aggregation problem within the ANR project *Aggreg* coordinated by Niehren.

At ***ICALP*** Bourhis (with Amarilli, Jachiet and Mengel) [13] developped a new algorithm to efficiently enumerates the solutions of certain type of circuits. They apply their result to give new proofs previous results on efficient enumeration for queries defined by tree automata or FO queries over structures with bounded tree width by using these circuits as aggregates to represent the set of all solutions of a query and then enumerating them.

Again at ***ICALP*** [15] Bacquey in an collaboration with Caen and Marseille (Grandjean and Olive) prove that linear time complexity on cellular automata is exactly characterized by inductive first-order Horn formulas. The method of proof also implies the following result: the enumeration of the ground atoms that are consequences of any inductive first-order Horn formula on a given structure can be performed in linear time (in the cardinality of the domain of the structure) by a cellular automaton (of appropriate dimension).

7.1.2. Provenance

Provenance is a type of aggregates that aims at exhibiting the contributions 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. Provenance is studied within the ANR project *Aggreg*.

In a paper at ICDT [14], Bourhis (with Amarilli, Monet and Senellart) studies the combined complexity for computing circuit representation of the provenance, which were used to efficiently evaluate aggregations tasks. In particular, they exhibit a recursive language of queries capturing path queries that compute a compact representation of the provenance.

7.1.3. Recursive Queries

At ***PODS*** [21], P. Bourhis proposed a formalisation of JSON documents, query languages and schema. This work is a collaboration with Chile. After having defined a clean theoretical framework to study JSON documents, Bourhis and his co-authors study the decidability and complexity of navigational query answering for different languages, relating each of them with existing implementations. Finally, they extend the documents with recursion together with a suitable querying language and study the complexity of query evaluating and query answering in this case.

At ***ICALP*** [17], P. Bourhis studied in a collaboration with Oxford the problem of definability in decidable fixpoint logic. Bourhis and his co-authors gives new characterisation of formulas that can be expressed in decidable logic with fixpoint. One of their main result is an effective characterisation of the formulas of the guarded negation fragment with fixpoint that can be expressed in the guarded fragment with fixpoint. Their techniques are then extended to effectively characterise the first order formulas that can be defined in the guarded fragment.

A. Lemay contributed at ICDE [16] 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. 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. Tison presented at ***IJCAI*** [18] — the top conference in Artificial Intelligence — a new ontology mediated query answering system (OMQA) for JSON document. This work is a collaboration with researchers from the University of Montpellier. The strength of their contribution lies in the fact that their ontology is very expressive and yet gives a tractable query answering system. Moreover, they establish a non-trivial connection between their query answering system and term rewriting, allowing them to pinpoint the exact complexity of query answering and to evaluate it directly over KV-stores.

Also a ***IJCAI*** [20], P. Bourhis studied guarded ontology languages that are compatible with cross product. This work was done in cooperation with Edinburgh and Vienna. Cross product is a useful modelling tool that allow to connect every element of one relation to every element of another relation. However, in this paper, Bourhis and his co-authors show that its introduction into guarded ontology – even when it is limited to two relations – quickly leads to the undecidability of query evaluation and query answering. However, they isolate fragments where one can add cross products without losing the decidability of these problems by either restricting the queries or the ontology.

7.1.5. Schema Validation

I. Boneva presented at ISWC [19] her work on ShEx 2.0 (Shape Expression Language 2.0), a language to describe the vocabulary and the structure of an RDF graph. This work is a collaboration with Oviedo and MIT. The language is based on the notion of shapes, a typing system supporting algebraic operations, recursive references to other shapes or Boolean combination. In the paper, Boneva and her co-authors give efficient algorithms to test if an RDF graph satisfies a shapes schema together with implementation guidelines. Her research on the topic has also led to the publication of a book [25] on the validation of RDF data, containing among other things her contribution to ShEx.

JSON documents are basically unordered data trees. Schemas for unordered data trees can thus be defined by appropriate notions of tree automata for unordered trees, as studied in a systematic manner by Boiret, Hugot, and Niehren [11] in cooperation with Treinen from Paris 7. Alternatively, schemas can be defined by closed logic formulas in the logics proposed by the same authors in [12] .They showed that logics for unordered data trees with equality tests of data values of siblings nodes remain decidable, and thus the equivalence problems of the corresponding tree automata. In contrast, the problem becomes undecidable when comparing cousins for equality of data values.

7.2. Managing Dynamic Linked Data

7.2.1. Complex Event Processing

In his PhD project [24], Sakho supervised by Niehren and Boneva proposes studies the complexity of answering automata queries on hyperstreams. A hyperstream is collections of streams that are connected with each other. The motivation for hyperstreams is to avoid blocking when composing of several stream processes. They show that the problem of deciding whether a tuple can is certain for selection on a hyperstream by query defined by finite automaton is PSPACE-complete.

7.2.2. Transformations

Symbolic tree transducers define transformations of data trees Adrien Boiret, Vincent Hugot and Joachim Niehren could show at at DLT [23] that the equivalence problem of symbolic top-down tree transducers can be reduced to that of standard top-down tree transducers. Thereby, the existing equivalence testers can be lifted to the symbolic tree transducers, yielding the algorithms needed for verification tasks in the ANR projec CoLiS. An implementation of such an algorithm by Nicolas Bacquey is on the way.

P. Gallot and S. Salvati presented their work on 1-register streaming string transducers at ***STACS*** [22]. This work is a collaboration with University of Bordeaux. Streaming String Transducers have recently gained a growing interest since they can be used to model transformations on data streams. In this work, P. Gallot, S. Salvati and their co-authors prove that 1-register streaming string transducers can be decomposed as a finite union of functional transducers. An immediate corollary of this result is that the equivalence of such transducers is decidable, which means that we can check if two given transducers represent the same transformation on data streams.

8. Partnerships and Cooperations

8.1. 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, N. Bacquey, P. Gallot, S. Salvati.
- 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: J. Niehren, P. Bourhis [correspondent], S. Salvati, N. Bacquey, 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.

ANR Bravas (2017-22):

- Participants: S. Salvati [correspondent]
- Scientific Partners: The coordinator is Jérôme Leroux from LaBRI, University of Bordeaux. The other partner is LSV, ENS Cachan.
- Objective: The goal of the BraVAS project is to develop a new and powerful approach to decide the reachability problems for Vector Addition Systems (VAS) extensions and to analyze their complexity. The ambition here is to crack with a single hammer (ideals over well-orders) several long-lasting open problems that have all been identified as a barrier in different areas, but that are in fact closely related when seen as reachability.

8.2. European Initiatives

Edinburgh-Links exchange projet funded by the University of Lille. The coordiator is Slawek Staworko.

Lille-Oxford cooperation project funded by the University of Lille. Links' contact is Pierre Bourhis.

8.3. International Initiatives

8.3.1. Inria International Partners

Niehren and Bourhis continue to cooperate with Domagoy Vrgoc from the University of Satiago di Chile, also after the end of the AMSud project.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. Member of the Conference Program Committees

I. Boneva was member of the program committees of Alberto Mendelson Workshop 2017.

P. Bourhis was member of the program committees of IC (Ingénierie des Connaissances) 2017.

P. Bourhis was the program chair of the demonstration track of BDA (Gestion de Données – Principes, Technologies et Applications) 2017.

J. Niehren is the chair of the WPTE 2018 workshop collocated with FLOCS in Oxfored

J. Niehren was a member of the program committee of BDA 2017

J. Niehren was a member of the programm committee of WPTE 2017.

S. Staworko was member of the program committees of ISWC (International Semantic Web Conference) 2017.

S. Tison was member of the program committees of FSCD (First International Conference on Formal Structure for Computation and Deduction) 2017.

S. Tison was member of the program committees of Highlight 2017.

9.1.2. Journal

9.1.2.1. Member of the Editorial Boards

J. Niehren is editor of Fundamenta Informaticae.

S. Salvati is managing editor of JoLLI (Journal for Logic, Language and Information).

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

9.1.3. Research Administration

I. Boneva was an elected member of CRIStAL laboratory concil until the December 1st, 2017.

J. Niehren is member of the Board of the Inria Lille's Board of the Comité des Equipe-Projects.

J. Niehren is head of the Inria team Links.

S. Salvati is secretary of the The Association for Logic, Language and Information (http://www.folli.info).

S. Salvati is vice-head of the Inria team Links.

S. Tison is president of CITC EuraRFID since June 2017.

S. Tison is a member of coordination committee I-Site ULNE since April 2017.

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

I. Boneva gives a Master 1 semester on Algorithms for databases.

F. Capelli organized a research school on Knowledge Compilation at ENS de Lyon in December 2017.

A. Lemay is pedagogical responsible for Computer Science and Numeric correspondent for UFR LEA, Lille 3.

J. Niehren was teaching the course "Foundations of Data and Knowledge Bases" at the University of Cape Coast, Ghana, as part of the PhD summer school of the Academie without Borders.

J. Niehren was teaching the course "Foundations of Databases" as part of the masters 2 Mocad on Information Extraction at the University of Lille 1.

S. Salvati is pedagogical responsible of Master Miage FA, Lille 1.

S. Salvati organised the reasearch label for Computer Science Bachelor, Lille 1.

S. Staworko is pedagogical responsible of the Web Analyst Master, Lille 3.

S. Staworko has created an online lecture on R as part of the partnership between Lille 3 University and Dakar University, Senegal.

S. Tison is pedagogical responsible of first year ACT Master, Lille 1.

S. Tison has organized Catalyst Contest 2017 in Lille.

9.2.2. Supervision

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.

PhD in progress: P. Gallot. On safety of data transformations. Started on October 2017. Supervised by Lemay and Salvati.

Projet de Fin d'Étude (PFE): N. Crosetti. On weighted dependency aggregation. Supervised by Capelli, Niehren and Ramon (Team MAGNET).

9.2.3. Juries

S. Salvati was the reviewer of El Makki VOUNDY's thesis *Langages ε-sûrs et caractérisations des langages d'ordres supérieurs* defended on November 15, 2017, University of Marseille.

S. Salvati was a member of the jury of Félix Baschenis' thesis *Minimisation de ressources pour les transductions régulières sur les mots* defended on December 5, 2017, University of Bordeaux.

9.2.4. Selection committies

P. Bourhis was a member of the selection committee for two assistant professorships at the University of Lille III.

A. Lemay was a member of the selection committee for two assistant professorships at the University of Lille III.

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

S. Tison was a member of the selection committee for an assistant professorship at the University of Marne-la-Vallée.

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

S. Tison is a member of the jury of "Agrégation de mathématiques, option D".

9.3. Popularization

Joachim Niehren contributed an article in the december issue of "Inria by Lille" titeled "Interroger les bases de données d'une manière plus intélligentes"

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. V. BOOM. *Characterizing Definability in Decidable Fixpoint Logics* , in "ICALP 2017 - 44th International Colloquium on Automata, Languages, and Programming", Varsovie, Poland, I. CHATZIGIANNAKIS, P. INDYK, F. KUHN, A. MUSCHOLL (editors), July 2017, vol. 107, 14 p. [*DOI :* 10.4230/LIPICS.ICALP.2017.107], https://hal.inria.fr/hal-01639015

[3] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN. *Automata for Unordered Trees*, in "Information and Computation", April 2017, vol. 253, pp. 304-335 [*DOI :* 10.1016/J.IC.2016.07.012], https://hal.inria.fr/hal-01179493

[4] I. BONEVA, J. G. LABRA GAYO, E. G. PRUD 'HOMMEAUX. *Semantics and Validation of Shapes Schemas for RDF*, in "ISWC2017 - 16th International semantic web conference", Vienna, Austria, October 2017, https://hal.archives-ouvertes.fr/hal-01590350

[5] 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

[6] 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

[7] 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

[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^{o} 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^{o} 578, pp. 100-127, https://hal.inria.fr/hal-00966625

[10] V. HUGOT, A. BOIRET, J. NIEHREN. *Equivalence of Symbolic Tree Transducers*, in "DLT 2017 - Developments in Language Theory", Liege, Belgium, August 2017, vol. 105, 12 p. [*DOI :* 10.1007/978-3-642-29709-0_32], https://hal.inria.fr/hal-01517919

Publications of the year

Articles in International Peer-Reviewed Journals

[11] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN. *Automata for Unordered Trees*, in "Information and Computation", April 2017, vol. 253, pp. 304-335 [*DOI :* 10.1016/J.IC.2016.07.012], https://hal.inria.fr/hal-01179493

[12] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN. *Logics for Unordered Trees with Data Constraints*, in "Journal of Computer and System Sciences (JCSS)", February 2017, 40 p. , https://hal.inria.fr/hal-01176763

International Conferences with Proceedings

[13] A. AMARILLI, P. BOURHIS, L. JACHIET, S. MENGEL. *A Circuit-Based Approach to Efficient Enumeration* , in "ICALP 2017 - 44th International Colloquium on Automata, Languages, and Programming", Varsovie, Poland, I. CHATZIGIANNAKIS, P. INDYK, A. MUSCHOLL (editors), July 2017, pp. 1-15 [*DOI :* 10.4230/LIPICS.ICALP.2017.111], https://hal.inria.fr/hal-01639179

[14] A. AMARILLI, P. BOURHIS, M. MONET, P. SENELLART. *Combined Tractability of Query Evaluation via Tree Automata and Cycluits*, in "ICDT 2017 - International Conference on Database Theory", Venice, Italy, March 2017 [*DOI :* 10.4230/LIPICS.ICDT.2017.6], https://hal.inria.fr/hal-01439294

[15] N. BACQUEY, E. GRANDJEAN, F. OLIVE. *Definability by Horn formulas and linear time on cellular automata*, in "ICALP 2017 - 44th International Colloquium on Automata, Languages and Programming", Warsaw, Poland, I. CHATZIGIANNAKIS, P. INDYK, F. KUHN, A. MUSCHOLL (editors), Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, July 2017, vol. 80, pp. 1-14 [*DOI :* 10.4230/LIPICS.ICALP.2017.99], https://hal.archives-ouvertes.fr/hal-01494246

[16] G. BAGAN, A. BONIFATI, R. CIUCANU, G. FLETCHER, A. LEMAY, N. ADVOKAAT. *gMark: Schema-Driven Generation of Graphs and Queries*, in "Data Engineering (ICDE), 2017 IEEE 33rd International Conference on", San Diego, United States, April 2017 [*DOI :* 10.1109/ICDE.2017.38], https://hal.inria.fr/hal-01591706

[17] *Best Paper*
M. BENEDIKT, P. BOURHIS, M. V. BOOM. *Characterizing Definability in Decidable Fixpoint Logics* , in "ICALP 2017 - 44th International Colloquium on Automata, Languages, and Programming", Varsovie, Poland, I. CHATZIGIANNAKIS, P. INDYK, F. KUHN, A. MUSCHOLL (editors), July 2017, vol. 107, 14 p. [*DOI :* 10.4230/LIPICS.ICALP.2017.107], https://hal.inria.fr/hal-01639015.

[18] M. BIENVENU, P. BOURHIS, M.-L. MUGNIER, S. TISON, F. ULLIANA. *Ontology-Mediated Query Answering for Key-Value Stores*, in "IJCAI: International Joint Conference on Artificial Intelligence", Melbourne, Australia, August 2017, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01632090

[19] I. BONEVA, J. G. LABRA GAYO, E. G. PRUD 'HOMMEAUX. *Semantics and Validation of Shapes Schemas for RDF*, in "ISWC2017 - 16th International semantic web conference", Vienna, Austria, October 2017, https://hal.archives-ouvertes.fr/hal-01590350

[20] P. BOURHIS, M. MORAK, A. PIERIS. *Making Cross Products and Guarded Ontology Languages Compatible*, in "IJCAI 2017 - Twenty-Sixth International Joint Conference on Artificial Intelligence", Melbourne, Australia, August 2017, pp. 880-886 [*DOI :* 10.24963/IJCAI.2017/122], https://hal.inria.fr/hal-01638346

[21] P. BOURHIS, J. L. REUTTER, F. SUÁREZ, D. VRGOČ. *JSON: Data model, Query languages and Schema specification*, in "PODS 2017 - Proceedings of the Thirty-Sixth ACM SIGMOD-SIGACT-

SIGART Symposium on Principles of Database Systems", Chicago, United States, May 2017 [*DOI :* 10.1145/3034786.3056120], https://hal.inria.fr/hal-01639182

[22] P. GALLOT, A. MUSCHOLL, G. PUPPIS, S. SALVATI. *On the decomposition of finite-valued streaming string transducers*, in "34th International Symposium on Theoretical Aspects of Computer Science (STACS)", Hannover, Germany, March 2017 [*DOI :* 10.4230/LIPICS], https://hal.archives-ouvertes.fr/hal-01431250

[23] V. HUGOT, A. BOIRET, J. NIEHREN. *Equivalence of Symbolic Tree Transducers*, in "DLT 2017 - Developments in Language Theory", Liege, Belgium, August 2017, vol. 105, 12 p. [*DOI :* 10.1007/978-3-642-29709-0_32], https://hal.inria.fr/hal-01517919

Conferences without Proceedings

[24] M. SAKHO, I. BONEVA, J. NIEHREN. *Complexity of Certain Query Answering on Hyperstreams*, in "BDA 2017 - 33ème conférence sur la « Gestion de Données — Principes, Technologies et Applications »", Nancy, France, November 2017, n^{o} 10, https://hal.archives-ouvertes.fr/hal-01609498

Scientific Books (or Scientific Book chapters)

[25] J. E. L. GAYO, E. PRUD'HOMMEAUX, I. BONEVA, D. KONTOKOSTAS. *Validating RDF Data*, Morgan & Claypool, September 2017, vol. 7, n^{o} 1, pp. 1 - 328 [*DOI :* 10.2200/S00786ED1V01Y201707WBE016], https://hal.archives-ouvertes.fr/hal-01667426

Other Publications

[26] A. BOIRET, A. LEMAY, J. NIEHREN. *A Learning Algorithm for Top-Down Tree Transducers*, July 2017, working paper or preprint, https://hal.inria.fr/hal-01357627