IN PARTNERSHIP WITH:
CNRS

Université de Lille

Activity Report 2019

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
A2.1.4. - Functional programming
A2.1.6. - Concurrent programming
A2.4. - Formal method for 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
A9.8. - Reasoning

Other Research Topics and Application Domains:

B6.1. - Software industry
B6.3.1. - Web
B6.3.4. - Social Networks
B6.5. - Information systems
B9.5.1. - Computer science
B9.5.6. - Data science
B9.10. - Privacy

1. Team, Visitors, External Collaborators

Research Scientist
Joachim Niehren [Team leader, Inria, Senior Researcher, HDR]

Faculty Members
Iovka Boneva [Université de Lille, Associate Professor]
Florent Capelli [Université de Lille, Associate Professor]
Aurélien Lemay [Université de Lille, Associate Professor, HDR]
Charles Paperman [Université de Lille, Associate Professor]
Sylvain Salvati [Université de Lille, Professor, HDR]
Slawomir Staworko [Université de Lille, Associate Professor, HDR]
Sophie Tison [Université de Lille, Professor, HDR]

Post-Doctoral Fellow
Bruno Guillon [Inria, until Aug 2019]

PhD Students
Nicolas Crosetti [Inria]
Lily Gallois [Université de Lille]
Paul Gallot [Inria]
Jose Martin Lozano [Université de Lille]
Momar Sakho [Inria]

Technical staff
Antonio Al Serhali [Inria, Engineer, from Nov 2019]
Jeremie Dusart [Inria, Engineer]

Interns and Apprentices
Fanny Canivet [Ecole Nationale Supérieure de Cognitique, from Jun 2019 until Jul 2019]
Daniel Fiogbe [Université de Lille, from Apr 2019 until Jun 2019]
Gireg Lezoraine [Université de Lille, from Jun 2019 until Jul 2019]
Quentin Mittelman [Ecole Nationale Supérieure de Cognitique, from Jun 2019 until Aug 2019]

Administrative Assistant
Nathalie Bonte [Inria]

Visiting Scientists
Rustam Azimov [JetBrains, Saint-Petersburg State University (Russia), Apr 2019]
Semyon Grigorev [JetBrains, Saint-Petersburg State University (Russia), Apr 2019]

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 summarize our main research objectives.

Querying Heterogeneous Linked Data We 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, where data sources may have rigid query facilities that cannot be easily adapted or extended.

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

For querying linked data collections, the main problems are to resolve the heterogeneity of data formats and schemas, to understand the efficiency and expressiveness of recursive queries, that can follow links repeatedly, to answer queries under constraints, and to optimize query answering algorithms based on static analysis. When linked data is dynamically created, exchanged, or updated, the problems are how to process linked data incrementally, and how to manage linked data collections that change dynamically. In any case (static and

dynamic) one needs to find appropriate schema mappings for linking semi-structured datasets. We will study how to automatize parts of this search process by developing symbolic machine learning techniques for linked data collections.

3.2. Querying Heterogeneous Linked Data

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

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

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

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

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

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

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

3.3. Managing Dynamic Linked Data

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

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

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

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

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

3.4. Linking Graphs

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

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

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

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

we plan to approach inference of general database mappings by investigating and employing algorithms for inference of n-ary queries.

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

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

4. Application Domains

4.1. Linked Data Integration

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

4.2. Data Cleaning

The second example of application of our proposal concerns scientists who want to quickly inspect relevant literature and datasets. In such a case, local knowledge that comes from a local repository of publications belonging to a research institute (e.g. HAL) need to be integrated with other Web-based repositories, such as DBLP, Google Scholar, ResearchGate and even Wikipedia. Indeed, the local repository may be incomplete or contain semantic ambiguities, such as mistaken or missing conference venues, mistaken long names for the publication venues and journals, missing explanation of research keywords, and opaque keywords.

We envision a publication management system that exploits both links between database elements, namely pointers to external resources and logical links. The latter can be complex relationships between local portions of data and remote resources, encoded as schema mappings. There are different tasks that such a scenario could entail such as (i) cleaning the errors with links to correct data e.g. via mappings from HAL to DBLP for the publications errors, and via mappings from HAL to Wikipedia for opaque keywords, (ii) thoroughly enrich the list of publications of a given research institute, and (iii) support complex queries on the corrected data combined with logical mappings.

4.3. Real Time Complex Event Processing

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

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Data Integration and Schema Validation

The ShEx language for defining RDF schemas was proposed and developed earlier by the Links team in cooperation with the W3C. S. Staworko et al. now studied the containment problem for ShEx schemas for RDF documents. They showed at *PODS* [10] – the best database theory conference – that the problem is decidable, but co-NEXP-hard. This is a joint work with P. Wieczorek from the University of Wroclaw, Poland.

5.1.2. Aggregates

Florent Capelli et al. showed at *STACS* [7] – a top conferences in theoretical computer science – a new knowledge compilation procedure for quantified Boolean formulas allowing to decide the satisfiability quantified Boolean formulas with bounded tree width in polynomial time. This can be applied in particular to first-order database queries with quantifiers. This is joined work with S. Mengel from the CNRS in Lens.

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.

RELEASE FUNCTIONAL DESCRIPTION: ShExJava now uses the Commons RDF API and so support RDF4J, Jena, JSON-LD-Java, OWL API and Apache Clerezza. It can parse ShEx schema in the ShEcC, ShEJ, ShExR formats and can serialize a schema in ShExJ.

To validate data against a ShExSchema using ShExJava, you have two different algorithms: - the refine algorithm: compute once and for all the typing for the whole graph - the recursive algorithm: compute only the typing required to answer a validate(node,ShapeLabel) call and forget the results.

- Contact: Iovka Boneva
- URL: http://shexjava.lille.inria.fr/

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

6.6. ShapeDesigner

KEYWORDS: Validation - Data Exploration - Verification

FUNCTIONAL DESCRIPTION: ShapeDesigner allows construct a ShEx or SHACL schema for an existing dataset. It combines algorithms to analyse the data and automatically extract shape constraints, and to edit and validate shape schemas.

- Contact: Jeremie Dusart
- URL: https://gitlab.inria.fr/jdusart/shexjapp

7. New Results

7.1. Querying Heterogeneous Linked Data

7.1.1. Data Integration and Schema Validation

Data integration requires knowledge about the structure of the various data. Such a structure is usually described by schemas. While for relational databases, schemas are hard-coded, this is not the case for many other formats. In XML for instance, several schema formalisms exists, such as DTD, XML Schema or Schematron. The Links Project-Team investigate the problem of defining schemas and use them to data, in particular for RDF and JSON Formats.

With P. Wieczorek of the University of Wroclaw, Poland, S. Staworko et al. have studied the containment problem of ShEx schemas for RDF documents in *PODS* [10].

Also, J. Dusart develops under the supervision of I. Boneva and S. Staworko the software *ShEx Validator* so as to foster the practical usage of ShEx. It is also worth noting that ShEx is now being adopted by several institutions such as *WikiData*.

7.1.2. Aggregates

Aggregation refers to computations that are alien to mere logical data manipulation (e.g. such as in relational algebra). Typically, aggregation means counting the number of answers, or performing other kinds of statistics. We have a slightly larger understanding as we may also include enumerating all answers with a *small delay*. Aggregation algorithms are generally subtle as they in most cases avoid the explicit generation of the whole set of answers. We study aggregation problems within the ANR project *Aggreg* coordinated by Niehren.

In the same spirit, Capelli et al. (in a joint work with Mengel from the CNRS in Lens) showed at *STACS* [7] a new knowledge compilation procedure which allows a polynomial algorithm to test the satisfiability quantified Boolean formulas with bounded tree width. In *Theory of Computing Systems*, [25], Capelli also gave a taxonomy of results according to various restrictions of tree-width of graphs.

Also, in *Theory of Computing Systems*, [25], Capelli gave a taxonomy of results according to various restrictions of tree-width of graphs.

Finally, in an article in *JCSS* [14], F. Capelli (with Bergougnoux and Kanté from Bordeaux and Clermont-Ferrand) propose an algorithm for counting the number of transversals (i.e. subset of nodes intersecting all hyperedges) in some hypergraphs.

7.1.3. Certain Query Answering

When data is incomplete, logical constraints and knowledge about its intended structure help to infer the answers of queries. This inference problem is known as *certain query answering*.

L. Gallois and S. Tison [6] presented in *IJCAI* - one of the main conferences of Artificial Intelligence. L. Gallois and S. Tison study boundedness of the chase procedure in the context of positive existential rules, providing decidability results for several classes and outlining the complexity of the problem. This work is done in collaboration with P. Bourhis and Graphik team-project. These results also belong to the PhD thesis of L. Gallois [11] supervised by S. Tison and P. Bourhis.

7.2. Managing Dynamic Linked Data

7.2.1. Complex Event Processing

Complex event processing requires to answer queries on streams of complex events, i.e., nested words or equivalently linearizations of data trees, but also to produce dynamically evolving data structures as output.

In an article published in *LATA* [17], I. Boneva, J. Niehren and M. Sakho studied certain query answering for hyperstreams - which are collections of connected streams - with *complex events* (i.e. that correspond to tree patterns). They showed that the problem is EXP-complete in general, and obtained PTIME algorithms when restricted to *linear* tree patterns (possibly with compression) and to deterministic tree automata.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Grants with Industry

Strapdata C. Paperman is actively collaborating with the Strapdata company on efficient distributed graph database using an Apache novel technology to query distributed graph *Gremlin* that could benefit of the main product of Strapdata: Elassandra as a *database backend*.

9. Partnerships and Cooperations

9.1. Regional Initiatives

- Links is member of the CPER Data (2016-19)
- Lozano's PhD project (2016-19) is co-funded by the Region Nord-Pas de Calais
- Sakho's PhD project is co-funded by the Region Nord-Pas de Calais
- Gallot's PhD project (2017-20) is co-funded by the Region Nord-Pas de Calais
- Crosetti's PhD project (2018-21) is co-funded by the Region Haut de France. This is joined work with J. Ramon from the Inria project Magnet

9.2. National Initiatives

ANR Aggreg (2014-19): Aggregated Queries.
Participants: Joachim Niehren [correspondent], Aurélien Lemay, Adrien Boiret [University of Mons, Belgium], Florent Capelli.

- The coordinator is J. Niehren and the partners are the Université Paris 7 (A. Durand) including members of the Inria project DAHU (L. Ségoufin), the Université de Marseille (N. Creignou) and Université de 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: Joachim Niehren [correspondent], Aurélien Lemay, Sophie Tison, Adrien Boiret [University of Mons, Belgium], Vincent Hugot [INSA Centre-Val de Loire], Nicolas Bacquey [Twig], Paul Gallot, Sylvain Salvati.

- The coordinator is R. Treinen from the Université 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: Iovka Boneva [correspondent], Sophie Tison, Jose Martin Lozano.

- Partners: The coordinator is E. Contejean from the Université Paris-Sud and the other partner is the Université de 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: Joachim Niehren [correspondent], Momar Sakho, Nicolas Crosetti, Florent Capelli.

- 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: Joachim Niehren [correspondent], Sylvain Salvati, Aurélien Lemay, Nicolas Bacquey [Twig], Lily 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):
Participant: Sylvain Salvati [correspondent].

- Scientific Partners: The coordinator is Jérôme Leroux from LaBRI, Université de 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.

9.3. European Initiatives

Oxford, UK: An exchange project with the computer science lab of the University of Oxford is funded by the Université de Lille via the CRIStAL Lab. Links' members produced many common publications over the years with Oxford. Links' contact is C. Paperman.

Wroclaw, Poland: S. Staworko has regular exchange with the University of Wroclaw. This has led to a publication at *PODS* [10] together with P. Wieczorek.

Saint-Petersburg, Russia: S. Salvati and J. Niehren started a cooperation with the Saint-Petersburg State University, via a month-long visit by R. Azimov and S. Grigorev.

Oviedo, Spain: I. Boneva has an active cooperation with the University of Oviedo.

9.4. International Initiatives

9.4.1. Informal International Partners

Santiago de Chile, Chile: S. Staworko and I. Boneva have a collaboration with C. Riveros from the Pontifical Catholic University of Chile since 2018.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events: Selection

10.1.1.1. Chair of Conference Program Committees

J. Niehren was co-chair of the programm committee of WPTE 2019.

10.1.1.2. Member of the Conference Program Committees

- J. Niehren: member of the Program Committee of LATA 2019
- S. Tison: member of the program committee of FSCD (International Conference on Formal Structure for Computation and Deduction) 2019.
- S. Staworko: member of the program committee of EDBT 2020: International Conference on Extending Database Technology, Copenhagen, Denmark, March 30–April 2, 2020; Demonstration Track
- F. Capelli: member of the program committee of IJCAI 2019

10.1.2. Journal

10.1.2.1. 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)

10.1.3. Invited Talks

- S. Staworko has been invited in GT Automata, Logic, Games and Algebra (ALGA) to give the talk *Shape Expressions Schemas for RDF: Semantics, Complexity, and Inference* (Oct 2019, Paris)
- I. Boneva and J. Dusart have been invited to Ghent University, Belgium Wikidata and Wikibase Workshop: developing a Wikibase instance (3 - 5 July 2019) as expert on ShapeDesigner
- F. Capelli has been invited in the seminar of LIMD (Université de Savoie) (13/06/2019)
- F. Capelli has been invited in the Dagstuhl Seminar "Deduction Beyond Satisfiability" (10/09/2019)
- F. Capelli has been invited in the seminar of LACL (Université Paris-Est Creteil) (18/11/2019)
- F. Capelli has been invited in "Journée du GT ALGA" (Paris, 11/10/201)9

10.1.4. Scientific Expertise

- S. Tison: member of the coordinating committee of I-Site Université Lille Nord Europe, about innovation and relationship with social economical world
- S. Tison: head of CITC-Eurarfid until June 2019
- J. Niehren: member of the board of the comittee of project-teams of Inria Lille

10.1.5. Research Administration

- S. Salvati: elected alternate member of Inria Evaluation Commission (2019-2023)
- F. Capelli: co-organizer of *Groupe de Travail* of CNRS IMIA (Informatique Mathématique Intelligence Artificielle)
- F. Capelli: organiser of a AI DAY for the GDR IM
- F. Capelli: organiser of an international workshop on knowledge compilation (project SACRe Kocoon - http://kocoon.gforge.inria.fr/)

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

I. Boneva teaches computer science in DUT Informatique of Université de Lille

F. Capelli teaches computer science in UFR LEA of Université de Lille for around 200h per year (Licence and Master). He is also responsible of remediation of Licence 1 in its UFR.

A. Lemay teaches computer science in UFR LEA of Université de Lille for around 200h per year (Licence and Master). He is also responsible for computer science and numeric correspondent for its UFR.

J. Niehren gives two lessons for the 2nd year students of the Master MOCAD (Université de Lille): one on databases (20.5h) and one on information extraction (21h).

C. Paperman teaches computer science for a total of around 200h per year. He gives lessons in UFR MISASH (Université de Lille), in Licence and Master. He also gives a database lesson of 25h in Master MOCAD (Université de Lille).

S. Salvati teaches computer science for a total of around 230h per year in computer science departement of Université de Lille. That includes Introduction to Computer Science (L1, 50h), Logic (L3, 50h), Algorithmic and operational research (L3, 36h), Functional Programming (L3, 35h), Research Option (L3, 10h), Semantic Web (M2, 30h), Advanced Databases (M1, 20h). He is *directeur d'étude* of Master MIAGE FA. He is a member of conseil de departement in Computer Science department of Université de Lille and of the ad-hoc commission of Doctoral School that studies PhD applications in computer science.

S. Staworko teaches computer science for a total of around 200h in UFR MIME (Université de Lille). He is co-head of the master web-analyst at Université de Lille.

S. Tison teaches computer science for a total of around 120h at Université de Lille. That includes a course on Advanced algorithms and complexity (54h, M1) and Business Intelligence (36h, M1).

S. Tison is member of the selection Board for "Agrégation" in Mathematics, more specifically in charge of the option "Computer Science".

10.2.2. Supervision

PhD: L. Gallois, Recursive Queries, defended on December 19, 2019, supervised by P. Bourhis (Team SPIRALS) and S. Tison

PhD in progress: N. Crosetti, Privacy Risks of Aggregates in Data Centric-Workflows, supervised by F. Capelli, J. Niehren, J. Ramon (Team MAGNET) and S. Tison

PhD in progress: P. Gallot, On safety of data transformations, since October 2017, supervised by A. Lemay and S. Salvati

PhD in progress: J.M. Lozano, On data integration for mixed database formats, supervised by I. Boneva and S. Staworko

PhD in progress: M. Sakho, Hyperstreaming Query answering on graphs, since 2016, supervised by J. Niehren and I. Boneva

10.2.3. Juries

10.2.3.1. PhDs committees

- J. Niehren was a reviewer of the PhD thesis of *Antony Lick* at ENS Cachan 2019
- S. Tison was member of the jury for *Xinzhe Wu*, Université de Lille, March 2019
- S. Tison was member of the jury for *Stathis Delivorias*, Université de Montpellier, September 2019 (reviewer)
- S. Tison was member of the jury for *Nicolas Bloyet*, Université Bretagne Sud, December 2019 (reviewer)

10.2.3.2. HDR committees

- S. Tison was member of the jury for *Arnaud Carayol*, December 2019, Université Paris Est-Marne la Vallée (reviewer)
- S. Salvati was member of the jury for *Colin Riba*, ENS Lyon, December 2019

10.3. Popularization

10.3.1. Education

RIC Days S. Salvati organizes the *Recherche Innovation et Créativité days*. It presents research professions to student (primarily Master students)

10.3.2. Interventions

Introduction to programming 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

TFJM^2 In 2019, L. Gallois participated in the organization of this event of the *Tournoi Français des jeunes mathématiciennes et mathématiciens*, a national contest of mathematics for high-school students

10.3.3. Internal action

- Inria by Lille: J. Niehren contributed an article in the december issue of "Inria by Lille" titled "Interroger les bases de données d'une manière plus intelligente"

11. Bibliography

Major publications by the team in recent years

[1] A. AMARILLI, C. PAPERMAN. *Topological Sorting with Regular Constraints*, in "45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)", Prague, Czech Republic, July 2018, https://hal.archives-ouvertes.fr/hal-01950909

[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. , ICALP-2017 Best paper award of Track B [*DOI :* 10.4230/LIPICS.ICALP.2017.107], https://hal.inria.fr/hal-01639015

[3] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN. *Logics for Unordered Trees with Data Constraints*, in "Journal of Computer and System Sciences", December 2018, 40 p. , https://hal.inria.fr/hal-01176763

[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. 1-38, https://hal.inria.fr/hal-01187986

[6] P. BOURHIS, M. LECLÈRE, M.-L. MUGNIER, S. TISON, F. ULLIANA, L. GALLOIS. *Oblivious and Semi-Oblivious Boundedness for Existential Rules*, in "IJCAI 2019 - International Joint Conference on Artificial Intelligence", Macao, China, August 2019, https://hal-lirmm.ccsd.cnrs.fr/lirmm-02148142

[7] F. CAPELLI, S. MENGEL. *Tractable QBF by Knowledge Compilation*, in "36th International Symposium on Theoretical Aspects of Computer Science (STACS 2019)", Berlin, Germany, March 2019, https://arxiv.org/abs/1807.04263 , https://hal.archives-ouvertes.fr/hal-01836402

[8] 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

[9] 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

[10] S. STAWORKO, P. WIECZOREK. *Containment of Shape Expression Schemas for RDF*, in "SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)", Amsterdam, Netherlands, June 2019, https://hal.inria.fr/hal-01959143

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] L. GALLOIS. *Dialog between chase approach and string rewriting system approach*, Université de Lille, December 2019, https://tel.archives-ouvertes.fr/tel-02445754

Articles in International Peer-Reviewed Journals

[12] A. AMARILLI, F. CAPELLI, M. MONET, P. SENELLART. *Connecting Knowledge Compilation Classes and Width Parameters*, in "Theory of Computing Systems", June 2019, https://arxiv.org/abs/1811.02944 [*DOI :* 10.1007/s00224-019-09930-2], https://hal.inria.fr/hal-02163749

[13] M. BENEDIKT, P. BOURHIS, M. VANDEN BOOM. *Definability and Interpolation within Decidable Fixpoint Logics*, in "Logical Methods in Computer Science", 2019, https://hal.archives-ouvertes.fr/hal-02392433

[14] B. BERGOUGNOUX, F. CAPELLI, M. M. KANTÉ. *Counting Minimal Transversals of β-Acyclic Hypergraphs*, in "Journal of Computer and System Sciences", May 2019, https://arxiv.org/abs/1808.05017 [*DOI :* 10.1016/J.JCSS.2018.10.002], https://hal.inria.fr/hal-01923090

[15] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN. *Logics for Unordered Trees with Data Constraints*, in "Journal of Computer and System Sciences", January 2019, 40 p. [*DOI :* 10.1016/J.JCSS.2018.11.004], https://hal.inria.fr/hal-01176763

International Conferences with Proceedings

[16] E. ALLART, C. VERSARI, J. NIEHREN. *Computing Difference Abstractions of Metabolic Networks Under Kinetic Constraints*, in "CMSB 2019 - 17th International Conference on Computational Methods in Systems Biology", Trieste, Italy, Lecture Notes in Computer Science, Springer, September 2019, vol. 11773, pp. 266-285 [*DOI :* 10.1007/978-3-030-31304-3_14], https://hal.archives-ouvertes.fr/hal-02302463

[17] I. BONEVA, J. NIEHREN, M. SAKHO. *Regular Matching and Inclusion on Compressed Tree Patterns with Context Variables*, in "LATA 2019 - 13th International Conference on Language and Automata Theory and Applications", Saint Petersburg, Russia, January 2019, https://hal.inria.fr/hal-01811835

[18] P. BOURHIS, M. LECLÈRE, M.-L. MUGNIER, S. TISON, F. ULLIANA, L. GALLOIS. *Oblivious and Semi-Oblivious Boundedness for Existential Rules*, in "IJCAI 2019 - International Joint Conference on Artificial Intelligence", Macao, China, August 2019, https://hal-lirmm.ccsd.cnrs.fr/lirmm-02148142

[19] F. CAPELLI. *Knowledge Compilation Languages as Proof Systems*, in "Theory and Applications of Satisfiability Testing – SAT 2019", Lisbon, Portugal, July 2019, pp. 90-99, https://arxiv.org/abs/1903.04039 [*DOI :* 10.1007/978-3-030-24258-9_6], https://hal.inria.fr/hal-02445523

[20] F. CAPELLI, S. MENGEL. *Tractable QBF by Knowledge Compilation*, in "36th International Symposium on Theoretical Aspects of Computer Science (STACS 2019)", Berlin, Germany, March 2019, https://arxiv.org/abs/1807.04263 , https://hal.archives-ouvertes.fr/hal-01836402

[21] S. STAWORKO, P. WIECZOREK. *Containment of Shape Expression Schemas for RDF*, in "PODS 2019 - 38th ACM SIGMOD-SIGACT-SIGAI Symposium on PRINCIPLES OF DATABASE SYSTEMS", Amsterdam, Netherlands, ACM Press, June 2019, pp. 303-319 [*DOI :* 10.1145/3294052.3319687], https://hal.inria.fr/hal-01959143

Other Publications

[22] I. BONEVA, J. DUSART, D. FERNÁNDEZ ALVAREZ, J. E. L. GAYO. *Shape Designer for ShEx and SHACL Constraints*, October 2019, ISWC 2019 - 18th International Semantic Web Conference, Poster, https://hal.archives-ouvertes.fr/hal-02268667

[23] I. BONEVA, J. DUSART, D. FERNÁNDEZ ALVAREZ, J. E. LABRA GAYO. *Semi Automatic Construction of ShEx and SHACL Schemas*, July 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02193275

[24] I. BONEVA, J. NIEHREN, M. SAKHO. *Approximating Certain Query Answers on Nested Hyperstreams*, April 2019, working paper or preprint, https://hal.inria.fr/hal-02092276

[25] F. CAPELLI. *Knowledge compilation languages as proof systems*, June 2019, https://arxiv.org/abs/1903.04039 - working paper or preprint, https://hal.inria.fr/hal-02163761

[26] F. CAPELLI, N. CROSETTI, J. NIEHREN, J. RAMON. *Dependency Weighted Aggregation on Factorized Databases*, January 2019, https://arxiv.org/abs/1901.03633 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01981553