IN PARTNERSHIP WITH:
CNRS

Université Charles de Gaulle (Lille 3)

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2018

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 Fellows
Nicolas Bacquey [Inria, until Aug 2018]
Bruno Guillon [Inria, from Nov 2018]

PhD Students
Nicolas Crosetti [Inria, from Oct 2018]
Dimitri Gallois [Université de Lille]
Paul Gallot [Inria]
Jose Martin Lozano [Université de Lille]
Momar Sakho [Inria]

Technical staff
Jeremie Dusart [Inria]

Administrative Assistant
Nathalie Bonte [Inria]

Visiting Scientists
Rustam Azimov [Saint Petersburg University, from Aug 2018 until Nov 2018]
Herminio Garcia Gonzalez [Oviedo University, from Sep 2018 until Dec 2018]

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

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

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

3. Research Program

3.1. Background

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

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

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

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

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

3.2. Querying Heterogeneous Linked Data

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

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

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

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

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

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

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

3.3. Managing Dynamic Linked Data

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

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

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

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

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

3.4. Linking Graphs

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

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

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

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

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

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

4. Application Domains

4.1. Linked Data Integration

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

4.2. Data Cleaning

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

4.3. Real Time Complex Event Processing

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

5. Highlights of the Year

5.1. Highlights of the Year

Containment for RDF Schemas

The ShEx language for defining RDF schemas was proposed and developed earlier by the Links team in cooperation with the W3C. Slawek Staworko et al. now studied the containment problem for ShEx schemas for RDF documents. They showed at **PODS** [7] – the best database theory conference – that the problem is decidable, but co-NEXP-hard. This is joined work with P. Wieczorek from the University of Wrazlaw.

Foundations of AI: Knowledge Compilation

Florent Capelli et al. showed at **STACS** [15] – 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.

Foundations of AI: Constrained Topologial Sort

Charles Paperman et al. showed at **ICALP** [8] – a top conferences in theoretical computer science – how to compute efficiently topological sorts of graphs under regular constraints. The problem was initially introduced in the context of preferential query answer for uncertain databases, where on usually wants to sort the query answers by some preferences, that are known only partially. It becomes then crucial to look for total orders on the answer set satisfying regular constraints that specify the preferences. Finding such an order for regular constraints was know to be infeasible in general. In this article, a class of regular constraints is identified for which this problem becomes tractable. A (partial) decidable dichotomy theorem is proven drawing the frontier between the kind of constraints which are feasible from those which are not. This is joined work with A. Amarilli from Telecom Paristech.

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

7. New Results

7.1. Querying Heterogeneous Linked Data

7.1.1. Data Integration

The PhD project of Lozano on relational to RDF data integration is progressing under the direction of Boneva, and Staworko. At **AMW** [9] they studied the *relational to RDF data exchange problem*. They focus in particular on a preliminary analysis of the consistency problem for relational to RDF data exchange with target ShEx schema.

7.1.2. Schema Validation

Shape Expression Language 2.0 (ShEx) is a language to describe the vocabulary and the structure of an RDF graph. It is base on the notion of shapes, a typing system supporting algebraic operations, recursive references to other shapes or Boolean combination.

In their **PODS** paper [7], Staworko studied the *containment problem* for ShEx (in cooperation with Wieczorek from Wrazlaw). Containement is a classical subject for schema-related issue in database theory. The authors proved that it is decidable for ShEx-schema, but with a untractable complexity (co-NEXP-hard). They also carefuly craft restriction of ShEx schema to design tractable-but-still-signifiant fragments.

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.

The topic of the PhD project of M. Sakho supervised by Niehren and Boneva is to generalize algorithms for querying streams to hyperstreams. These are collections of linked streams as naturally produced as intermediate results of complex events processing. Hyperstreams are incomplete descriptions of relational structures, so they can queried similarly to incomplete databases, for which the notion of a certain query answer is most appropriate.

In a paper published at **RP** [13], they studied certain query answering for hyperstreams with simple events. Such hyperstreams can be identified with compressed string patterns. They proved that the certain query answering for regular queries on compressed string patterns is PSPACE-complete, independently of whether the finite automata defining the regular queries are assumed deterministic or not, and independently of whether compression is permitted or not. They also showed that the problem is in PTIME when restricted to *linear* string patterns (possibly with compression) and to deterministic finite automata.

In a paper published at **LATA** [6], the studied certain query answering on hyperstreams of complex events. Such hyperstreams can be modeled by compressed tree pattern with context variables. They showed that certain query answering for regular queries on compressed tree pattern with context variables is EXP-complete, independently of whether the tree automata defining the regular queries are assumed deterministic or not, and independently of whether compression is permitted or not. They also showed that the problem is in PTIME when restricted to *linear* tree patterns (possibly with compression) and to deterministic tree automata.

7.2.2. Transformations

In his PhD project – belonging to the ANR Colis– Gallot with his supervisors Salvati and Lemay presented higher order tree transducers which extend macro tree transducers. Moreover they obtained nice properties such as the closure of the transducers under composition. Algorithms to compute such compositions are proposed. Those algorithms perform partial evaluation and are guided by semantic interpretations over finite domains.

Another virtue of higher-order transducers is that their *linear* syntactic restriction make them equivalent to logically defined MSO transductions. One of the composition algorithm proposed preserves the linearity. Furthermore, we have also showed that we can decrease the order of linear transducer (i.e. the complexity of the functions it handles) when this one is larger than 4.

These results are unpublished paper for now.

7.3. Foundations of AI

Various problems of databases and knowledge bases are closely related to foundational problems in artificial intelligence, since they are rooted in logic or graph theory.

7.3.1. Knowledge Compilation

Many problems in Artificial Intelligence boil down to the exploration of the solution set (called the models) of logical formulas. Such an exploration can be finding one model of the formula, counting the number of models or enumerating them all. However, even for simple quantifier-free formulas, those explorations are known be untractable (NP-hard).

Knowledge compilation encompasses methods that aim to change the representation of the set of models in order to get tractable algorithms for (some of) those tasks. A big computational cost is paid during the compilation time but then replying to queries become tractable on the new representation. More generally, the core of Knowledge compilation is the study of the trade-off between the size of the representation and the easiness of queries. This subject is of interest for both Artificial Intelligence and Database communities.

At **STACS** [15], Capelli, in cooperation with Mengel from CRIL (Lens), studied knowledge compilation techiques for quantified Boolean formulas. Deciding the existence of models for such formulas is known to climb arbitrarly high the polynomial time hierarchy. The authors provide an efficient compilation procedure for formulas having a *bounded tree-width* generalizing results from SAT solving.

7.3.2. Aggregation and Enumeration for Graphs

Aggregation and enumeration are not relevant for answer sets of database queries but equally for any kinds of sets, most typically defined by combinatoric problems on graphs.

In a paper published at **ICALP** [8], Paperman proposed (in cooperation with Amarilli from Telecom Paristech) to study the problem of finding so called *topological sort* satisfying constraints provided by regular expressions. Searching topological sort happens typically in situations where an order is *uncertain*. For instance, in relational database where users provides a partial preference order, or in concurrent and distributed programming where some tasks can be executed in an arbitrary order. A classical task in *preferential query answering* is to find a topological sort satisfying some global constrained. Typically, to find a total order satisfying all (or most) of the customers. The paper provides and proves sufficient conditions on the *shape of the constraints* to make the problem tractable (P-time) as well as sufficient condition to make the problem NP-hard. They also prove a complete dichotomy for an adapted and well chosen version of the constrained topological sort problem.

In an article in **JCSS** [2], Capelli (with Bergougnoux and Kanté from Bordeaux and Clérmont-Ferrand) propose an algorithm for counting the number of *transversal* in some *hypergraphs*. Here, a hypergraph is a collection of sets – called *hyperedges* over a *ground set* and a traversal is a subset intersecting all hyperedges. In full generality, counting the number of minimal traversals in a hypergraph is a hard problem: it is known to be $\#P$-complete. They proved that under the assumptions of β-acyclicity, it is possible to count all the minimal traversals can be done in polynomial times.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Grants with Industry

Posos. A. Lemay is directing an internship of a master student (*projet de fin d'étude*) in cooperation with the POSOS company from Amiens. The goal of this collaboration is to work on efficient schema for a large pharmaceutical Knowledge Base.

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.

- Gillot'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: J. Niehren [correspondent], P. Bourhis, A. Lemay, A. Boiret, F. Capelli.
- The coordinator is J. Niehren and the partners 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: J. Niehren, M. Sakho, N. Crosetti, F. 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: J. Niehren, S. Salvati, A. Lemay, 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.

9.3. European Initiatives

Oxford: A exchange project with the computer science lab of the University of Oxford is funded by the University of Lille via the Cristal Lab. Links' member produced many common publications over the years with Oxford. Links' contact is Paperman.

Wrazlaw: Staworko has regular exchange with the University of Wrazlaw. This has led to a publication at **PODS** [7] together with P. Wieczorek.

Saint Petersburg: Salvati and Niehren started a cooperation with the University of Saint Petersburg, via a 3 months visit of R. Azimov in 2018.

Oviedo: Boneva started a cooperation with the University of Oviedo, via a 3 months visit of H. Garcia Gonzalez in 2018.

9.4. International Initiatives

9.4.1. Informal International Partners

Santiago de Chile: S. Staworko started a collaboration with C. Riveros from the Pontificia Universidad Catolica de Chile in 2018.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

Several researchers has visited us:

- Filip Mazowiecki, a researcher from Warsaw University and currently in post-doctorate in Bordeaux to work with Charles Paperman.
- Rustam Azimov, a Russian PhD students from Saint Petersburg State University, to collaborate with Sylvain Salvati and Joachim Niehren.
- Michaël Cadilhac, a researcher from Oxford University to work with Charles Paperman.
- Cristian Riveros, an Assistant Professor at the Department of Computer Science at the Pontificia Universidad Catolica de Chile.
- Henning Fernau, Professor at Universität Trier and Andreas Maletti, Professor at Universität Leipzig, visited us during the HDR defense of Aurelien Lemay.

9.5.1.1. Internships

- Nicolas Crosetti started an internship supervised by Florent Capelli, Joachim Niehren and Jan Ramon. His internship has evolved into the preparation of a PhD thesis.
- Chen Huan, from Centrale Lille, has done an internship under the supervision of Sylvain Salvati and Joachim Niehren.

9.5.2. Visits to International Teams

- Charles Paperman visited Michaël Cadilhac form the verification team of the University of Oxford.
- Joachim Niehren got invited by Hilal Zaid for a visit at the American University of Palestine in August 2018.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

- F. Capelli: workshop organisation of *Graph and Constraints* (27/08) within the conference Constraint Programming (CP) 2018, Lille.
- F. Capelli: organisation of annual meeting of GT ALGA (Groupe de Travail Automata, Logic, Games, Algebra of CNRS) the 15th and 16th of October at Lille.

10.1.2. Scientific Events Selection

10.1.2.1. Chair of Conference Program Committees

- J. Niehren was is was chair of the Program Committee of WPTE 2018.
- J. Niehren was is was co-chair of the Program Committee of WPTE 2019.

10.1.2.2. Member of the Conference Program Committees

- F. Capelli: member of Program Committee of International Joint Conference on Artificial Intelligence (IJCAI) 2018.
- F. Capelli: member of Program Committee of workshop Quantified Boolean Formulas (QBF) within FLoC conference (Federated Logic Conference).
- F. Capelli: member of Program Committee of workshop Graph and Constraints, 2018.
- S. Tison: member of Program Committee of RAIRO ITA, 2018.
- J. Niehren: member of the Program Committee of LATA 2019.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

J. Niehren is editor of Fundamenta Informaticæ.

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).

10.1.4. Invited Talks

- B. Guillon gave invited talks at Mid-term Meeting of ANR Delta in Bordeaux in December 2018.
- A. Lemay gave invited talks at Mid-term Meeting of ANR Delta in Bordeaux in December 2018.
- Joachim Niehren gave an invited talk at the American University of Palestine in August 2018.
- F. Capelli get invited to talk at seminars of CRIL at Lens, LACL in Créteil and VALDA in Paris.

10.1.5. Scientific Expertise

- S. Tison: member of coordinator of i-SIte ULNE, about innovation and relationship with social economical world.
- S. Tison: Head of CITC-Eurarfid.
- J. Niehren is member of the board of the committee of project-teams of Inria Lille.

10.1.6. Research Administration

- F. Capelli: Co-organizer of *Groupe de Travail* of CNRS IMIA (Informatique Mathématique Intelligence Artificielle)

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

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

C. Paperman is pedagogical responsible for Master MIASHS "parcours" WebAnalyste.

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

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

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

S. Salvati organized the research label for Computer Science Bachelor, Lille.

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

10.2.2. Supervision

HdR : Aurélien Lemay, *Machine Learning Techniques for Semistructured Data*, Université de Lille, Fri 16th Nov.

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

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.

10.2.3. Juries

S. Tison: Vice-Présidente du jury Agrégation de Mathématiques (co-pillote option D- Informatique).

S. Tison: Jury de Thèse Lucien Mousin.

S. Tison: PhD Committee of Narjes Jomaa.

F. Capelli: PhD Committee of Mikaël Monet.

J. Niehren: Habilitation Committee of Aurélien Lemay.

10.3. Popularization

10.3.1. Internal action

General Assembly of Inria Lille Niehren presented Links work on data and knowledge bases on the Dynamic Semantic Crosswords demonstration during a general assembly of Inria Lille in July 2018.

10.3.2. Creation of media or tools for science outreach

Dynamic Semantic Crosswords Bacquey's demonstration system on dynamic semantic crosswords is presented in the new showroom of Inria Lille in the new building Place. The demo generates dynamically crosswords while streaming Twitter feeds, depending on a semantic topic specified by the user. The specification can be given by a list of hashtags, and in the future by a XPath 3.0 query, that can be executed on streams by using Links QuiXPath tool. This illustrates the work on complex event processing by Niehren and his students during the last years.

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] A. LEMAY. *Machine Learning Techniques for Semistructured Data*, Université de Lille, November 2018, Habilitation à diriger des recherches, https://tel.archives-ouvertes.fr/tel-01929944

Articles in International Peer-Reviewed Journals

[2] B. BERGOUGNOUX, F. CAPELLI, M. M. KANTÉ. *Counting Minimal Transversals of β-Acyclic Hypergraphs*, in "Journal of Computer and System Sciences", November 2018, https://arxiv.org/abs/1808.05017 [*DOI :* 10.1016/J.JCSS.2018.10.002], https://hal.inria.fr/hal-01923090

[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] F. CAPELLI, Y. STROZECKI. *Incremental delay enumeration: Space and time*, in "Discrete Applied Mathematics", August 2018, https://arxiv.org/abs/1703.01928 [*DOI :* 10.1016/J.DAM.2018.06.038], https://hal.inria.fr/hal-01923091

[5] L. DAVIAUD, C. PAPERMAN. *Classes of languages generated by the Kleene star of a word*, in "Information and Computation", October 2018, vol. 262, n^{o} Part 1, pp. 90-109, https://hal.archives-ouvertes.fr/hal-01943493

International Conferences with Proceedings

[6] M. SAKHO, I. BONEVA, J. NIEHREN. *Regular Matching and Inclusion on Compressed Tree Patterns with Context Variables*, in "13th International Conference on Language and Automata Theory and Applications (LATA)", Saint Petersburg, Russia, Springer, March 2019, https://hal.inria.fr/hal-01926011

[7] 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

Conferences without Proceedings

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

[9] I. BONEVA, J. LOZANO, S. STAWORKO. *Relational to RDF Data Exchange in Presence of a Shape Expression Schema*, in "AMW 2018 - 12th Alberto Mendelzon International Workshop on Foundations of Data Management", Cali, Colombia, May 2018, pp. 1-16, https://hal.archives-ouvertes.fr/hal-01775199

[10] I. BONEVA, J. NIEHREN, M. SAKHO. *Certain Query Answering on Compressed String Patterns: From Streams to Hyperstreams*, in "RP 2018 - 12th International Conference on Reachability Problems", Marseille, France, September 2018, https://hal.archives-ouvertes.fr/hal-01609498

Research Reports

[11] S. SALVATI. *On is an n-MCFL*, Université de Lille, Inria, CRIStAL CNRS, April 2018, https://hal.archives-ouvertes.fr/hal-01771670

Other Publications

[12] I. BONEVA, J. NIEHREN, M. SAKHO. *Approximating Certain Query Answering on Hyperstreams*, June 2018, Technical report, https://hal.inria.fr/hal-01811835

[13] I. BONEVA, J. NIEHREN, M. SAKHO. *Certain Query Answering on Compressed String Patterns: From Streams to Hyperstreams (long version)*, July 2018, working paper or preprint, https://hal.inria.fr/hal-01846016

[14] 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

[15] F. CAPELLI, S. MENGEL. *Knowledge Compilation, Width and Quantification*, July 2018, https://arxiv.org/abs/1807.04263 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01836402

[16] F. CAPELLI, Y. STROZECKI. *Enumerating models of DNF faster: breaking the dependency on the formula size*, October 2018, working paper or preprint, https://hal.inria.fr/hal-01891483