Evaluation of Inria Theme Data and Knowledge Representation and Processing

Project-team LINKS

03/10/2019

Project-team title: LINKS

Scientific leader: Joachim Niehren

Research centers: INRIA Lille - Nord Europe

Contents

1 Personnel

Current composition of the project-team:

Research scientists and faculty members:

- Joachim Niehren [Team leader, Inria, Senior Researcher, HDR]
- Sylvain Salvati [Team co-head, Université de Lille, Professor, HDR]
- 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]
- Sławomir Staworko [Université de Lille, Associate Professor, HDR]
- Sophie Tison [Université de Lille, Professor, HDR]

Engineers:

- Jeremie Dusart [Inria]

Post-docs:

- Bruno Guillon [Inria]

Ph.D. students:

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

Administrative assistant:

- Nathalie Bonte [Inria]

Personnel at the start of the evaluation period (7/08/2015)

	INRIA	CNRS	University	Other	**Total**
DR (1) / Professors	1		1		3
CR (2) / Assistant professors		1	3		4
ARP and SRP (3)					
Permanent engineers (4)					
Temporary engineers (5)					
Post-docs			1		1
PhD Students			3		3
Total	1	1	9		11

(1) “Senior Research Scientist (Directeur de Recherche)”
(2) “Junior Research Scientist (Chargé de Recherche)”
(3) “Inria Advanced Research Position” and “Inria Starting Research Position”
(4) “Civil servant (CNRS, INRIA, ...)”
(5) “Associated with a contract (Ingénieur Expert, Ingénieur ADT, . . .)”

Personnel at the time of the evaluation (03/10/2019)

	INRIA	CNRS	University	Other	**Total**
DR / Professors	1		2		3
CR / Assistant professors			5		5
ARP and SRP					
Permanent engineers					
Temporary engineers	1				1
Post-docs					
PhD Students	3		2		5
Total	5	0	9		14

Changes in the scientific staff

DR / Professors / ARP CR / Assistant Professors / SRP	INRIA	CNRS	University	Other	**Total**
Arrivals			2		
Departures		1			

Comments: Departure of Pierre Bourhis who joined the INRIA Spirals Project-team. Arrival of S. Salvati in 2016, C. Paperman and F. Capelli in 2017.

Current position of former project-team members

- Pierre Bourhis, CR CNRS 2013-2017 , currently CR CNRS at SPIRALS Project-Team
- Adrien Boiret, PhD student 2011-2016, currently Post-doc at Université libre de Bruxelles
- Tom Sebastian, PhD student 2010-2016, currently R&D manager at Transaction Factory
- Vincent Hugot, post-doc 2013-2017, currently associate professor at Université de Franche-Comté
- Nicolas Bacquey, post-doc 2016-2018, currently Engineer at Twig

2 Research goals and results

2.1 Keywords

Computer Science and Digital Science:

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

Other Research Topics and Application Domains:

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

2.2 Context and overall goals of the project

We first start by restating here the project, as it has been defined at its creation in 2012.

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 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. Open data collections are typically distributed over the internet, are built with a large number of data models, and data sources often have rigid query facilities that cannot be easily adapted or extended. Exploiting these valuable collections is a main challenge for the team. 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.

An important assumption of our work is that the linked data collections are correct in most dimensions. This hypothesis allows us to base our work on logical methods. It intuitively 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 one is to clean the incorrect data in those dimensions that are subject to errors. We shall investigate how far these assumptions can be maintained in realistic applications, and how they can be weakened otherwise.

Querying linked data collections implies coping with **heterogeneity**: heterogeneity of data formats, heterogeneity of schemas... It also requires to handle recursive queries that can follow links repeatedly, to answer queries under constraints, and to optimize query answering algorithms based on static analysis. When we further consider that linked data is *dynamically* created, exchanged, or updated, other difficulties come into the picture related to incremental processing and management of evolving distributed data. In both cases (static and dynamic) relying on appropriate *schema mappings* for linking semi-structured datasets can be of great help. As such mappings may not be explicitly given, we study how to automatically build them using *symbolic machine learning* techniques.

The following three paragraphs summarize our main research objectives in the period 2012-2019.

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 investigate evaluation algorithms, containment problems, and concrete applications.

Managing Dynamic Linked Data In order to manage dynamic linked data collections and workflows, we 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 develop symbolic machine learning algorithms, for inferring queries and mappings between linked data collections in various graphs formats from annotated examples.

2.3 Research results

2.3.1 Research axis: Querying Heterogeneous Linked Data

Personnel

Permanent Iovka Boneva, Pierre Bourhis, Florent Capelli, Aurélien Lemay, Joachim Niehren, Charles Paperman, Slawek Staworko and Sophie Tison.

Nonpermanent Nicolas Crosetti, Lily Gallois, Jose Martin Lozano, Jeremie Dusart, Bruno Guillon.

Project-team positioning

Logical queries are in a core interest of PODS community and thus largely studied by database theory groups internationally (Oxford, Tel Aviv, Edinburgh, University of California, etc etc). Heterogenous linked data in RDF format is a core topic of the International Semantic Web Conference (ICWS), which also has a very broad international community.

The Inria projects in these fields all belong to the theme of the present evaluation. Inria's Valda team (P. Senellart, with ENS Paris) shares the interest on aggregate queries, but with a different motivation originating from probabilistic databases. This lead to a series of common publications. Inria's GraphIK team (M.L. Mugnier with LIRMM) shares some comon interests on ontology based database queries and on extensions of Datalog motivated by data integration. This has led to 2 common publications. Inria's Cedar team (I. Manolescu with LIX) has common interests on graph databases, and Inria's mOeX team (J. Euzenat in Grenoble) on the semantic Web. With the latter two projects there were no cooperations in the evaluation period.

Scientific achievements

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.

Hightligh

Boneva and Staworko proposed the Shape Expression Language ShEx for defining RDF schemas in cooperation with people form the MIT and the W3C in 2015, where ShEx was considered for standardization since then. Boneva presented Shex 2.0 at the International Semantic Web Conference (ISWC) [33]. She also published a book with a ShEx tutorial [51] with the people from the W3C and developed the ShEx Validator tool for practical usage with Dusart. It is worth noting that ShEx got adopted by so important institutions as WikiData. On the theoretical side, Staworko et. al studied the containment problem for ShEx schemas for RDF documents and published their results at PODS, the top-most database theory conference [48]. They showed that the containment problem is decidable, but co-NEXP-hard. This was a joined work with the University of Wrazlaw.

JSON documents being basically unordered data trees, schemas for JSON can be defined by notions of tree automata for unordered trees. This has been studied in a systematic manner by A. Boiret, V. Hugot, and J. Niehren in cooperation with R. Treinen from Paris 7, published in Information and Computation [8]. Alternatively, schemas can be defined by closed logic formulas in the logics proposed by the same authors in [9] in LATA. This work subscribes to the ANR Project *Colis* where unranked data trees are used as linux scripts.

When data does not comply with its schema, it needs repair. Bourhis and Staworko in cooperation with Bordeaux and Oxford presented at TODS [11] their work on bounded repairability for regular tree languages, which is a study on whether a tree document (typically XML) can be repaired to fit a given target tree language within a bounded amount of tree editing operations.

Finally, J. Lozano is finish his PhD project under the supervision of Boneva and Staworko. His topic subscribes the ANR project *Datacert* on data integration from relational databases to RDF graphs, with a first publication in [34].

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.

Hightligh

Capelli (in cooperation with CNRS in Lens) showed at STACS [44] a new knowledge compilation procedure from quantified boolean formulas to a certain class of logical circuits. It implies that the satisfiability quantified boolean formulas with bounded tree width can be decided in polynomial time. This compilation procedure can be applied in particular to first-order database queries with quantifiers. In cooperation with Telecom Paristech, the CNRS in Lens, and Grenoble, Bourhis [21] presented a ICALP a new algorithm to efficiently enumerate the solutions of such logical circuits. Beside of the application go logical queries with quantifiers, this result unifies over several previous algorithms for efficient enumeration.

A variety of other results were obtained that are relevant to aggreation issues. In Theory of Computing Systems, [64], Capelli gave a taxonomy of results according to various restrictions of tree-width of graphs. They give a better understanding of various data structures that can be used to build aggregates. At ICALP [25] Bacquey in an collaboration with Caen and Marseille proved a complexity result on cellular automata that has incidence on enumeration solutions of a first-order Horn formula. Also, in ICALP [24], C. Paperman proposed (in cooperation with Telecom Paristech) to study the problem of finding *topological sort* satisfying constraints provided by regular expressions. In an JCSS article [7], F. Capelli (with Bordeaux and Clermont-Ferrand) propose an algorithm for counting the number of transversals (i.e. subset of nodes intersecting all hyperedges) in some hypergraphs.

Finally, N. Crosetti started a PhD project 2018 under the supervision of Capelli, Tison, Niehren, and Ramon (from Inria's Magnet team). The question is how to solve aggregation problems for computing statistics from answer sets of database queries, as needed in machine learning [62].

Provenance Provenance is a type of aggregates that reflects the contributions of data to the answer of a query in a database. Provenance explains and allow the verification of queries answers. It is studied within the ANR project *Aggreg*. The team studied it both from a theoretical and a practical perspective.

Hightligh

In ICDT [22], Bourhis with Telecom Paristech and Inria's Valda team studied the combined complexity for computing circuit representations of the provenance. In particular, they exhibit a recursive language of queries capturing path queries that compute a compact representation of the provenance. Bourhis applied provenance techniques to recommendation systems in cooperation with Tel Aviv. This work provides a summary of data that explains of the end results. The corresponding tool was demonstrated at EDBT [53]. Last but not least, Bourhis [23] showed at PODS, in collaboration with Telecom ParisTech and ENS Paris, that the lineage of MSO queries on tree-like database instances is tractable, but not on any other instances. In SAT [64], F. Capelli uses knowledge compilation technics over bounded-tree width graphs to include not only provenance information, but also information that allows one to check the validity of results derived from the aggregate.

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.

Hightligh

P. Bourhis [37] presented at LICS—the top conference in logic in computer science—a general framework for querying databases with visible and invisible relations. This work was done in cooperation with Oxford, Santa Cruez, and Bordeaux. It generalizes in a uniform manner the problems of certain query answering and access control for relational databases. Invisible relations are subject to the open world assumption possibly under constraints, while visible relations are subject to the closed world assumption. Bourhis then shows that the problem of answering Boolean conjunctive queries in this framework is decidable, and studies the complexity of various versions of this problem. It turns out that the complexity increases compared to the problem of certain query answering, given that the closed world assumption is adopted for the added visible relations.

Bourhis also studied at IJCAI [20] certain query answering with some transitive closure constraints, which use recursion to define constraints. This work was done in collaboration with Oxford and Telecom ParisTech.

Bourhis and Tison, in collaboration with Inria's GraphIK team presented at IJCAI [29] a rule-based ontology language for JSON records and its computational properties. This work provides precise exact combined complexity of query answering in this framework and tractability results for data complexity.

The problem of ontological query containment consists in establishing whether the certain answers of two queries subject to an ontology are included in each other. Bourhis [40] studied at KR this problem for several closely related formalisms: monadic disjunctive Datalog (MDDLog), MMSNP (a logical generalization of constraint satisfaction problems) and ontology-mediated queries (OMQs). This work was done in cooperation with Bremen.

Gallois should finish her PhD project this year. In collaboration with Inria's GraphIK team [39], Bourhis, Gallois and Tison study at IJCAI 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.

Recursive Queries In RDF graph, dealing with queries that involve recursion is a major issue. It is the major difference with query languages such as SQL. From the logical perspective, it is also what separates MSO logic from FO logic.

Lemay contributed at TKDE [6] and then at ICDE [26] the *gMark* benchmark, a tool to generate large size graph database and an associated set of queries. This work was done in cooperation with Eindhoven and previous members of Links who are now in Lyon and Clermont-Ferrant. The tool was also demonstrated at VLDB [5]. Its main interests are a great flexibility, an ability to generate recursive queries, and the possibility to generate large sets of queries having a desired selectivity. This benchmark allowed for instance to highlight difficulties for the existing query engines to deal with recursive queries with high selectivity.

At LICS [28] Bourhis showed in collaboration with Oxford how to lift a major restriction on decidable fixpoint logics that can define recursive queries (such as C2RPQs), specifically on guarded logic. This result significantly improves the expressiveness over known decidable fixpoint logics. As a follow-up, Bourhis studied at ICALP [27] the problem of definability in decidable fixpoint logics. Their techniques effectively characterize first order formulas that can be defined in the guarded fragment.

Recursive queries also occur on tree-shaped document. As such, Bourhis proposed a formalisation of JSON documents, query languages and schema in collaboration with Chile, at PODS [42]. They study the decidability and complexity of query answering for different navigational languages with recursive operations and relate each of them with existing implementations.

Collaborations

Oxford The team has an exchange project with the computer science lab of the university of Oxford that produced many common publication over the years. Links' contact is Paperman. There are many common publications with Bourhis.

MIT, W3C, Ovideo The definition of the ShEx Language is done in collaboration with J. Labra Gaya (University of Ovideo) and Eric Prud'hommeaux (MIT/W3C). This includes joint publications ([33] and [51]).

Wraclaw Staworko studies ShEx formallly with P. Wieczorek from the University of Wraclaw) [48].

Telecom Paristech The most important partner in Paris is A. Amarilli from Telecom Paristech. He published on aggregates ([21] with P. Bourhis and [24] with C. Paperman), or on certain query Answering ([37])

Chile, Argentina, Bordeaux The team had an AMSUD collaboration from 2015 to 2016 named 'foundation of Graph Database' with Chile, Argentia and Bordeaux. This led notably to a publication in PODS [42].

Inria Paris One other imporant partner in paris is P. Senellart from Inria's Valda team at ENS Paris leading some common publications with P. Bouhis [54, 23]

Paris 7 The study on unordered trees and JSON formalism is done with R. Treinen of Paris 7, with two joint publications ([8] and [9])

Lens F. Cappelli and P. Bourhis cooperate with S. Mengel from the CNRS in Lens knowledge compilation ([44], [21])

Inria Montpellier S. Tison and P. Bourhis has an ongoing cooperation with GraphIK (Graphs for Inferences on Knowledge, joint team between Inria, LIRMM and INRA) on Knowledge Representing and Reasoning, leading to two publications [39, 29].

Caen, Marseille, Paris7 The ANR project Aggreg coordinated by Niehren is in cooperation with A. Durand from Paris 7, E. Grandjean from Caen, and Nadja Creignou in Marseille.

External support

Two PhD project were cofunded by the ANR projects Aggreg and DataCert with the Region Haut-de-France. The third PhD project is funded by the University of Lille. The AMSUD Project 'Foundations of Graph Database' added some travel money.

Self assessment

All 4 highlights of Links in the evaluation period belong to this research axis, so this axis is the most successful of Links. The most important contributions is the ShEx language and tool for defining RDF schemas. It found much interest internationaly both from an engineering and a theoretical perspective. The results were published in the respectively top conference of the database theory community (PODS) and of the semantic web community (ICSW). Our theoretical work on aggregate queries lead to important insights on how to represent answer sets of database queries in a concise manner. Links was one of the main players internationally, who started to make this topic popular at STACS and PODS. The close relationship between aggregation and provenance also raised much interest. The ANR project Aggreg coordinated by J. Niehren help to make this possible. The theoretical work on logical queries (Certain Query Answering, Recusive Queries) was highly successful with publications at LICS and ICALP. Despite of all these successes, the full task of querying heterogenous linked data has not yet been tackeled.

2.3.2 Research axis: Managing Dynamic Linked Data

Personnel

Permanent Iovka Boneva, Pierre Bourhis, Joachim Niehren, Sylvain Salvati, Sophie Tison.
Nonpermanent Paul Gallot, Momar Sakho, Nicolas Bacquey, Vincent Hugot.

Project-team positioning

Complex event processing and data-centric workflows for service orchestration are topics of the SIGMOD community. One of the leading groups on complex event processing currently is in Chile (C. Riveros). There are no other Inria teams working on this topic to the best of our knowledge. Concerning data-centric workflows, they are studie a the U.C. San Diego (V. Vianu). We are cooperating with the french groups on the topic through the ANR project Headwork (2016-21) on concepts of crowed-sourcing platforms: The Druid group of the University in Rennes (D. Gross-Amblard), Inria's Valda team, Inria's Sumo team. Our results on programming with data subscribes to domain of program verification, which is internationally represented by the POPL community, with various teams in France and at Inria, such as for instance the Inria's Toccata team (C. Marché, Inria Saclay).

Scientific achievements

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

In his PhD thesis, T. Sebastian [3] developed with his supervisor J. Niehren streaming algorithms covering all of XPath 3.0 queries on `XML` streams, using a higher-order query language called λXP. At SOFSEM [47], they proposed a new technique to speed up by a factor of 4 the evaluation of navigational XPath queries on XML streams based on document projection. The idea is to skip those parts of the stream that are irrelevant for the query. These algorithms were implemented in the QuiXPath tool.

In 2016, M. Sakho started his PhD project on hyperstreaming query answering algorithms for graphs under the supervision of J. Niehren and I. Boneva. Hyperstreams are collections of streams connected with together. In a paper published at RP [35], they studied certain query answering for hyperstreams with *simple events* (i.e. that correspond to string patterns). They obtained PSPACE-complete algorithms for this problem in general, and they also showed that the problem is in PTIME when restricted to *linear* string patterns (possibly with compression) and to deterministic finite automata. At LATA [36], this have been extended to hyperstreams of *complex events* (corresponding 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.

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

In collaboration with ENS Cachan and San Diego, P. Bourhis presented at ICDT [19] techniques on collaborative access control in a distributed query and data exchange language (Webdamlog). The goal of this work was to provide a semantic to data exchange rules defined by Webdamlog. It also allowed to prove that it is possible to formally verify whether there are data leakages.

P. Bourhis defined with Tel Aviv at **ICDE** [38] a notion of provenance for data-centric workflows, and proved that it can be used to explain the provenance of fact in the final instance of an execution. This provenance is used to answer three main questions: *why* does a specific tuple appear in the answer of a query, *what if* the initial database is changed (Revision problem), and *how to* change the query to obtain a missing tuple.

Programmning with Data The ANR Project Colis aims to formalize Linux install scripts to perform static analysis on them, for instance to prove that install and uninstall script leave the system unchanged. The Links team focuses its study on tree transducers adapted to this problem.

First, structure of filesystems requires tree formalisms which are unranked, unordered, and based on an unbounded alphabets. Beside the unranked aspect already presented in axis 1 in link with JSON, A. Boiret, V. Hugot and J. Niehren studied symbolic tree transducers at DLT [46], which allow to deal with data trees. They proved in particular that the equivalence problem of symbolic top-down tree transducers can be reduced to that of standard top-down tree transducers, yielding the algorithms needed for verification tasks in the ANR project CoLiS.

The limitation of this approach, based on classical top-down tree transducers, is that it does not capture non-local operations on the filesystem. For this, P. Gallot with his supervisors S. Salvati and A. Lemay develop in his PhD project higher order tree transducers which extend on macro tree transducers. In particular, they are closed under composition, with practical algorithms to compute such composition. Also, syntactic restriction of linearity make them equivalent to logically defined MSO transductions ; one of the composition algorithm we proposed preserves the linearity. Furthermore, they 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 for now.

On an different problem, P. Gallot and S. Salvati presented their work on 1-register streaming string transducers at STACS [45], in collaboration with University of Bordeaux. In this work, P. Gallot, S. Salvati and their co-authors prove that 1-register streaming string transducers can be decomposed as a finite union of functional transducers, which allow to obtain decidability results for the equivalence problem.

Collaborations

Paris 7, Inria Saclay On program verification we cooerpate with Paris 7 (T. Treinen) and the Inria's Toccata team (C. Marché) through the ANR project CoLiS. This lead to some comon publications [46, 8]. some of which were reported in Section 6.1.1 on schemas for JSON documents [9].

U.C. San Diego, Inria Paris We cooperation on data-centric workflows for crowd sourcing systems [19] with the Inria's Valda team (S. Abiteboul) and the University of San Diego (V. Vianu). This work subscribes to the ANR project HeadWork.

Tel Aviv Our practival approaches on data-centric workflows for crowd sourcing systems [38] were done in cooperation with the University of Tel Aviv (Tova Milo).

Bordeaux In the context of program verification we are cooperting on transducers with the University of Bordeaux [45]. This cooperation subscribes to the ANR project Delta (A. Muscholl, G. Puppis).

External support

The research ANR projects HeadWork and CoLiS co-funded two PhD thesis with the Region Haut-de-France. The ANR projects Delta and Bravo gave some travel money.

Self assessment

Our theoretical results complex event processing are very satisfactory and published in good conferences (LATA, RP). The practical activity (SOFSEM) will be reactivated in the next period. Our work on data-centric workflows in the ANR project HeadWork produced very good publication (ICDT) too. Very good publications were obtained in our work on programming with data in the ANR project CoLiS (DLT, STACS) in which we took a high scientific risc. All three lines of research have a good potential for transfer. The transfer actions in the current period were restricted to data-centric workflows though.

2.3.3 Research axis: Linking Graphs

Personnel

Permanent Aurélien Lemay and Slawek Staworko.

Project-team positioning

The work on learning tree transformations is at the intersection between the community on grammatical inference and on transducers, which are both well-established traditonally various theory conferences. Leading players internationally are TU Munich (H. Seidl), U. Pensylvenia (R. Alur), U. Libre de Brussel (E. Filiot) beside many other. The work on query learning addresses the database community. These topics are not studied by other Inria projects

Scientific achievements

Learning Transformations The definition of a query or of a transformation may be a tedious task for a non-expert user, especially considering the diversity and the complexity of database and query formalisms used on formats such as XML, JSON or RDF. To solve this problem, we propose to use machine learning based from examples or from simple interactions with the end-user.

A. Boiret obtained his PhD for his work on the "Normalization and Learning of Transducers on Trees and Words" under the supervision of J. Niehren and A. Lemay. He showed how to learn top-down tree transformations with regular schema restrictions [30, 56, 55]. At LATA [32], he deepened a result of a previous PhD student of Links on learning sequential tree-to-word transducers (with output concatenation), by showing how to find normal forms for less restrictive linear tree-to-word transducers. At DLT [31], he could show in cooperation with Munich, that the equivalence problem of this class of transducers is in polynomial time, even though their normal forms may be of exponential size.

Also, A. Lemay presented in 2018 its HDR on learning queries and transformations for semi-structured data, which covers problems presented in those works.

Finally, in the context of learning RDF graph transformations, Staworko presented a cooperation with Edinburg at VLDB [43]. Using bisimulation technique, he aims at aligning datas of two RDF Graphs that takes into account blank values, changes in ontology and small differences in data values and in the structure of the graph. the alignement of graphs is an important first step for the inference of transformations.

Learning Join Queries S. Staworko published in TODS an article [10] on learning join queries from user examples in collaboration with Universities of Lyon and Clermont-Ferrand that present techniques that allow the automatic construction of a join query through interaction with a user that simply labels sets of tuples to indicate whether the tuple is in the target query or not.

Collaborations

Edingburg The contributions on graph alignement [43] were done in a collaboration with P. Buneman (University of Edinburg).

Lyon, Clermont-Ferrand The work on join queries [10] was done in collaboration with A. Bonifati (University of Lyon) and R. Ciucanu (University of Clermont-Ferrand), two former members of the team LINKS.

Self assessment

On transducer induction, we obtained very good publications (DLT) and (LATA), and also on query induction (TODS). But given that all PhD projects on this topic were running out, the activity on transducer learning shifted to the previous subaxis on programming with data in section 2.3.2. The activity on query induction shifted to the previous subaxis on RDF schema validation in section 2.3.1.

2.4 Evolution of research directions during the evaluation period

In the evaluation period, LINKS pursued the following tree axis:

1. Querying Heterogeneous Linked Data
2. Managing Dynamic Linked Data
3. Linking Data Graphs

Axis 1. on querying heterogeneous linked data was particularly successful, with all 4 highlight being there. The second axis on managing dynamic data was running very well. Concering Axis 3. on linking graphs the two main activities shifted over rather natually two Axis 1 and 2.

In Axis 1 we indeed did not yet have the time to tackle the full problem of querying heterogenous linked data. The missing part is graph data integration. Now that the RDF schema language ShEx is available, we can start to work on RDF schema mapping as needed for RDF data integration. Furthermore, symbolic inference techniques from the third axis are promising in that context. We therefore propose to distinguish a new Axis 3 on the graph data integration with inference and schemas.

One notorious remaining problem in rest of Axis 1 is the algorithmic problem of how to query datagraphs efficiently. The difficulties are the the datagraphs may be very large and distributed and that queries may be recursive. We want to tackle such problems with various techniques. In partiular, the usage of circuits opens new oportuntities in this context. One idea is to that pattern in datagraphs can identified with logical queries on datagraphs. Therefore, problems to find frequent graph patterns in datagraphs as needed for data analysis with machine learning and datamining techniques could possibly be be reduced to problems of circuits that represent the answers set of the logial database query.

Concerning Axis 2 on managing dynamic data, we propose to make our work evolve to the problem of monitoring datagraphs. We propose to do so based on novel functional programming languages that we want to develop. The work on querying hyperstreams will be shifted to the objective to hyperstreaming program evaluation. We do not plan to pursue the topic of data-centric workflows any further.

As a result of these evolutions, we propose to update the structure of our research agenda as follows:

1. Querying Datagraphs
2. Monitoring Datagraphs
3. Graph Data Integration

3 Knowledge dissemination

3.1 Publications

	2016	2017	2018	2019
PhD Theses	2			
H.D.R. (*)			1	
Journals	5	5	2	3
Conference proceedings (**)	12	11	3	4
Book chapters				
Books (written)		1		
Books (edited)				
Patents				
General audience papers				
Technical reports			2	

(*) HDR Habilitation à diriger des Recherches
(**) Conferences with a program committee

3.1.1 Major Journals

Here are the major journals in the field and the number of papers coauthored by members of the project-team during the evaluation period.

- Information and Computation (IC) : 2 publications
- Journal of Computer System Sciences (JCSS) : 2 publications
- Transactions on Database Systems (TODS) : 2 publications
- Theoretical Computer Science (TCS) : 1 publication

3.1.2 Major Conferences

Here are the major conferences in the field, and the number of papers coauthored by members of the project-team during the evaluation period:

- International Colloquium on Automata, Languages and Programming (ICALP): 4 publications, including one best-paper award
- International Conference on Database Theory (ICDT) : 3 publications
- International Joint Conference on Artificial Intelligence (IJCAI): 4 publications
- Symposium on Logic in Computer Science (LICS): 2 Publications
- Principle of Database Systems (PODS): 3 publications
- Symposium on Theoretical Aspects of Computer Science (STACS) : 2 publications
- Very Large Databases (VLDB) : 2 publications
- International Semantic Web Conference (ISWC): 1 publication

3.1.3 Major Publications

We list here five representative major publications of the team:

- **Querying Visible and Invisible Information** [37] by P. Bourhis, M. Benedikt, B. Ten Cate and G. Puppis, published in LICS 2016,
- **Semantics and Validation of Shapes Schemas for RDF** [33] by I. Boneva; J. G. Labra Gayo, E. G. Prud'Hommaux. In ISWC 2017 - 16th International semantic web conference, 2017.
- **Automata for Unordered Trees** [9], by A. Boiret, V. Hugot, J. Niehren and R. Treinen, published in Information and Computation 2017.

- **Knowledge Compilation, Width and Quantification** [44], by F. Capelli and S. Mengel, published in STACS 2019,
- **Containment of Shape Expression Schemas for RDF** [48], by S. Staworko and P. Wieczorek, published in PODS 2019,

3.2 Software

ShEx Validator 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.
Web site: http://shexjava.lille.inria.fr/. Self-assessment:
- Audience: A-4 (large audience, used by people outside the team).
- Software originality: SO-4 (original software implementing a fair number of original ideas).
- Software maturity: SM-2 (basic usage works, terse documentation).
- Evolution and maintenance: EM-3 (good quality middle-term maintenance).
- Software distribution and licensing: SDL-4 (public source or binary distribution on the Web).

ShapeDesigner ShapeDesigner allows to construct a ShEx or a SHACL schema for an existing graph dataset. It combines algorithms to analyse the data and automatically extract shape constraints, and to edit and validate shape schemas.
Web site: https://gitlab.inria.fr/jdusart/shexjapp. Self-assessment:
- Audience: A-4 (large audience, used by people outside the team).
- Software originality: SO-3 (original software reusing known ideas and introducing new ideas).
- Software maturity: SM-2 (basic usage works, terse documentation).
- Evolution and maintenance: EM-3 (good quality middle-term maintenance).
- Software distribution and licensing: SDL-4 (public source or binary distribution on the Web).

gMark: schema-driven graph and query generation 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
Web site: https://github.com/graphMark/gmark. Self-assessment:
- Audience: A-3 (ambitious software, usable by people outside the team).
- Software originality: SO-4 (original software implementing a fair number of original ideas).
- Software maturity: SM-3 (well-developed software, good documentation, reasonable software engineering).
- Evolution and maintenance: EM-1 (no real future plans).
- Software distribution and licensing: SDL-4 (public source or binary distribution on the Web).

SmartHal 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 CRISTAL, which can be easily adapter to other Labs.

Web site: http://smarthal.lille.inria.fr/. Self-assessment:

- Audience: A-3 (ambitious software, usable by people outside the team).
- Software originality: SO-3 (original software reusing known ideas and introducing new ideas).
- Software maturity: SM-3 (well-developed software, good documentation, reasonable software engineering).
- Evolution and maintenance: EM-2 (basic maintenance to keep the software alive).
- Software distribution and licensing: SDL-4 (Web).

QuiXPath 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. 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.

Web site: https://project.inria.fr/quix-tool-suite/. Self-assessment:

- Audience: A-3 (ambitious software, usable by people outside the team).
- Software originality: SO-4 (original software implementing a fair number of original ideas).
- Software maturity: SM-3 (well-developed software, good documentation, reasonable software engineering).
- Evolution and maintenance: EM-4 (a replacement for QuixPAth is under developement).
- Software distribution and licensing: SDL-3, SDL-4 (a part is distributed to industrial partners in a contractual setting, another is published under open source license GPL-3.0).

X-Fun 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.

Self-assessment:

- Audience: A-1 (internal prototype).
- Software originality: SO-4 (original software implementing a fair number of original ideas).
- Software maturity: SM-2 (basic usage works, terse documentation).
- Evolution and maintenance: EM-2 (basic maintenance to keep the software alive).

Networkdisk Networkdisk is a python library in developemment that aims to manipulate large graph databases. This library extends the graph management library Networkx by using technics that store and manipulate graphs on disk rather than on memory to allow to deal with larger structures.

Web site: https://gitlab.inria.fr/guillonb/networkdisk. Self-assessment:

- Audience: A-1 (internal prototype).
- Software originality: SO-3 (original software reusing known ideas and introducing new ideas).
- Software maturity: SM-1 (demos work, loose documentation).
- Software distribution and licensing: SDL-5 (external packaging and distribution, as part of a popular open source distribution or a commercially-distributed product).

3.3 Technology transfer and socio-economic impact

- I. Boneva was a member of the Data Shapes Working Group of the W3C until the end of the group in 2017. Her mission was to produce a language for defining structural constraints on RDF graphs. http://www.w3.org/2014/data-shapes/charter.

- S. Tison was vice-dean of University Lille 1 in charge of innovation and partnership (2016-2017). She chaired the IOT cluster CITC (2017-2019). She was a member of scientific council of the society See-d. She is currently in the coordination team of I-Site "Université Lille-Nord Europe", in charge of the Development of relations with the socio-economic world. S. TisonJ'ai mis certaines de mes responsabilités ... mais je ne sais si c'est pertinent. elles sont ailleurs de toutes façons.

3.4 Teaching

I. Boneva teaches computer science in DUT Informatique of University of Lille (France)

F. Capelli teaches computer science in UFR LEA of University of 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 University of Lille for around 200h per year (Licence and Master). He is also responsible for computer science and numeric correspondant for its UFR.

J. Niehren gives a lesson on database (20h30) and one one on information extraction (21h) in Master 2 MOCAD (University of Lille).

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

S. Salvati teaches computer science for a total of around 230h per year in computer science departement of University of 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 University of Lille and of the ad-hoc commission of Doctoral School that studies PhD applications in computer science.

S. Staworko S. Staworko teaches computer science of around 200h in UFR MIME (University of Lille). He is co-head of the master web-analyst in University of Lille. He was co-head from sept 2016 to sept 2017 and the head from sept 2017 to sept 2018

S. Tison teaches computer science for around 120h in University of Lille. That includes a lesson on Discrete Mathematics (36h, L1), 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".

3.5 General audience actions

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

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.

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 intélligente"

Dynamic Semantic Crosswords A Demonstration system on *dynamic semantic crosswords*, developped by N. Bacquey, 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 illustrate the work on complex event processing by J. Niehren

and his students during the last years. J. 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.

TFJM^2 In 2018, L. Gallois was the coach of the winning team of *Tournoi Français des jeunes mathématiciennes et mathématiciens*, a national contest of mathematics for high-school students. She also participated in the organization of this event in 2019.

3.6 Visibility

3.6.1 Prizes and Awards

Best Paper Award P. Bourhis' paper with Oxford (Benedikt and van den Boom) won the best paper award of ICALP'17 (Track B) [27]

3.6.2 Member of Conference Program Committees

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

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

IJCAI P. Bourhis was member of program committee of Provenance week 2016, IJCAI (International Joint Conference on Artificial Intelligence) 2016.

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

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

CP F. Capelli: workshop organisation of *Graph and Constraints* (27/08) within the conference Constraint Programming (CP) 2018, Lille.

GT ALGA F. Capelli organised the annual meeting of GT ALGA (Groupe de Travail Automata, Logic, Games, Algebra of CNRS) the 15th and 16th of October at Lille.

IJCAI F. Capelli was a member of Program Committee of International Joint Conference on Artificial Intelligence (IJCAI) 2018.

QBF F. Capelli was a member of Program Committee of workshop Quantified Boolean Formulas (QBF) 2018 within FLoC conference (Federated Logic Conference)

Graph Constraints F. Capelli: member of Program Committee of workshop Graph and Constraints, 2018.

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

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

WPTE J. Niehren was the chair of the WPTE 2018 workshop collocated with FLOCS in Oxfored. He is also a member of the program committee of WPTE 2017.

WPTE J. Niehren was chair of the program committee of WPTE 2018.

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

LATA J. Niehren: member of the Program Committee of LATA 2019

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

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

FSCD S. Tison was member of the program committees of FSCD (International Conference on Formal Structure for Computation and Deduction) 2016, 2017, 2019.

Highlight S. Tison was member of the program committees of Highlights of Logic, Game and Automata 2017.

3.6.3 Member of the Journal Editorial Boards

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

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

JOLLI S. Salvati is in the editorial board of the Journal of Logic, Language and Information (JOLLI)

3.6.4 Leadership within the Scientific Community

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

GT IMIA F. Capelli is the responsible of the French National working group GT IMIA on IA (artificial intelligence) and IM (informatique mathematique).

3.6.5 Scientific Expertise

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

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

3.6.6 Research Administration

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

IMIA F. Capelli is a Co-organizer of *Groupe de Travail* of CNRS IMIA (Informatique Mathématique Intelligence Artificielle) in 2018

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

INRIA Team J. Niehren is head of the INRIA team Links. S. Salvati is *vice-head*

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

ComUE S. Tison is an elected member of the academic council of *ComUE Lille Nord de France* since November 2015.

University of Lille S. Tison is an elected member of the executive board of *University of Lille*.

i-Site ULNE S. Tison is a coordinator member of i-Site "Université Lille-Nord Europe", about innovation and relationship with social economical world.

University of Lille1 S. Tison was a vice president of the University of Lille 1 from 2016 to 2017, where she was responsible for industrial partnerships, innovation, and valorisation.

CITC-EuraRFID S. Tison chaired CITC-Eurarfid June 2017- June 2019 (https://www.citc-eurarfid.com/)

4 Funding

National initiatives

- **ANR Aggreg** (2014-19): Aggregated Queries.
  Particpants: J. Niehren [correspondent], P. Bourhis, A. Lemay, A. Boiret
  The coordinator is J. Niehren and the parteners are the University Paris 7 (A. Durand) including members of the Inria project DAHU (L. Ségoufin), the University of Marseille (N. Creignou) and University of Caen (E. Grandjean). The total amount for the team is 165 k€.
  Objective: the main goal of the Aggreg project is to develop efficient algorithms and to study the complexity of answering aggregate queries for databases and data streams of various kinds.
- **ANR Colis** (2015-20): Correctness of Linux Scripts.
  Participants: J. Niehren [correspondent], A. Lemay, S. Tison, A. Boiret, V. Hugot.
  The coordinator is R. Treinen from the University of Paris 7 and the other partner is the Tocata project of Inria Saclay (C. Marché). The total amount for the team is 150 k€.
  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. The total amount for the team is 75 k€.
  Objective: the main goals of the Datacert project are to provide deep specification in Coq of algorithms for data integration and exchange and of algorithms for enforcing security policies, as well as to design data integration methods for data models beyond the relational data model.
- **ANR Headwork** (2016-21):
  Participants: P. Bourhis [correspondant], J. Niehren, M. Sakho.
  Scientific partners: The coordinateur is D. Gross-Amblard from the Druid Team (Rennes 1). Other partners include the Dahu team (Inria Saclay) and Sumo (Inria Bretagne). The total amount for the team is 78 k€.
  Industrial partners: Spipoll, and Foulefactory.
  Objective: The main object is to develop data-centric workflows for programming crowd sourcing systems in flexible declarative manner. The problem of crowd sourcing systems is to fill a database with knowledge gathered by thousands or more human participants. A particular focus is to be put on the aspects of data uncertainty and for the representation of user expertise.
- **ANR Delta** (2016-21):
  Participants: P. Bourhis [correspondent], L. Gallois.
  Partners: The coordinator is M. Zeitoun from LaBRI, other partners are LIF (Marseille) and IRIF (Paris-Diderot). The total amount for the team is 140 k€.
  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 J. Leroux from LaBRI, University of Bordeaux. The other partner is LSV, ENS Cachan. The total amount for the team is 9 k€.
  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.

European projects

Edinburg Links was a member of an exchange project between the Universities of Edinburgh and Lille. The coordinator is S. Staworko.

Oxford An 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 C. Paperman.

Wroclaw S. Staworko has regular exchange with the University of Wroclaw. This has led to a publication at PODS [48] together with P. Wieczorek.

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

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

Industrial contracts

We had promsing contacts with various companies but non of these did not lead to funded interactions so far.
Sopra/Steria
StrapData
Posos

Inria Project Labs, Exploratory Research Actions and Technological Development Actions

None

Associated teams and other international projects

AMSUD P. Bourhis from Links had an AMSUD collaboration from 2015 to 2016 named 'foundation of Graph Database' with Chile, Argentia and Bordeaux. This led notably to a publication in PODS [42].

5 Follow up to the previous evaluation

Many applications of RDF do not call for a schema that imposes strict structural constraints on the topology of the RDF graph. However, slowly but surely RDF becomes a format for uses that previously have been reserved for formats such as XML or JSON e.g., sharing data between applications or making publicly available fragments of data that is typically stored in relational databases. In those applications the need for schema has been clearly identified, and for instance, a schema allows to easily ascertain the structure of an RDF document, which is essential for formulating queries, and typically permits validating the RDF document against the schema, thus giving some guarantee that processing the RDF document will run smoothly and will not have unexpected and potentially dangerous side-effects. We observe an increased interest in development of graph databases as illustrated by the efforts of the Linked Data Benchmark Council (LDBC) in defining a formal model for property graphs together with a novel query language (GQL) and a suitable schema for property graphs. Our work fits within this trend and we believe that our expertise. Indeed, we view and investigate RDF as a graph database rather a format for representing semantic information. We believe that our investigations are of interest to broad academic audience, a belief backed up by our recent academic accomplishments.

Indeed the problems of data integration and data exchange for RDF can be easily reduced to their relational equivalents since an RDF can be viewed as a set of triples, and consequently, can be modeled with a single relation. Our findings show, however, that the above problems gain particular flavors for RDF in the presence of schema and cannot be easily addressed with the existing solutions for relational databases. One shortcoming of the existing results for relational databases is their focus on subclasses of FO queries, such as conjunctive queries, which are not considered suitable for querying graph databases and where elements of regular path expressions are often call for. Also, various notions may turn out to be inadequate. For instance, shape schemas (ShEx and SHACL) can be expressed using tuple generating dependencies (tgd's) and equality generating dependencies (egd's). While it is know that universal solutions, essential in computing certain query answers, can be constructed if the set of tdg's and egd's is weakly acyclic. However, recursive shape schemas yield sets of dependencies that is not necessarily weakly acyclic. Consequently, a universal solution may not exists but we prose an alternative notion of universal simulation solution that can be constructed for cyclic sets of tgd's and egd's and can be used to compute certain answers to nested regular path queries without the inverse operator.

6 Objectives for the next four years

6.1 Research axis: Querying Data Graphs

Linked data is often abstracted as datagraphs: nodes carry information and edges are labeled. Internet, the semantic web, open data, social networks and their connections, information streams such as twitter are examples of such datagraphs. An axis of Links is to propose methods and tools so as to extract information from datagraphs. We dwell in a wide spectrum of tools to construct these methods: circuits, compilation, optimization, logic, automata, machine learning. Our goal is to extend the kinds of information that can be extracted from datagraphs while improving the efficiency of existing ones. This axis is split within two themes. The first one pertains to the use of lwo level representation by means of circuits to compute efficiently complex numerical aggregates that will find natural applications in AI. The second one proposes to explore path oriented query language and more particularly their efficient evaluation by means of efficient compilation and machine learning methods so as to have manageable statistics.

Circuits for Data Analysis. Circuits are consice representations of data sets that recently found a unifying interest in various areas of artificial intelligence. A circuit may for instance represent the answer set of a database query as a dag whose operators are unions (for disjunction) and cartesian products (for conjunction). Similarly, it may also represent the set of all matches of a pattern in a graph. The structure of the circuit may give rise to efficient algorithms to process large data sets as they are typically an order of magnitude smaller that the sets the represent. Among others, their applications range from knowledge representation/compilation, counting the number of solutions of queries, efficient query answering, factorized databases.

The problem of compiling a logical query and a database to an appropriate circuits is now well explored and mature. Therefore we can rely on circuits when wanting to apply database queries to data analysis in artificial intelligence. We propose to study novel tasks on circuits coming from database queries for this purpose. In particular we plan to develop algorithms that can solve various optimization problems on data sets that are given by circuits. Such algorithms should find natural application in machine learning or data mining, in cooperation, that we plan to develp with Ramon from the Inria project Magnet.

Path Query Optimization. Graph databases are easily queried with path descriptions. Most often these paths are described by means of regular expressions. This makes path queries difficult as the use of Kleene star makes them recursive. In relational DBMS, recursion is almost never used and it is not advised to use it. Moreover graph databases emerged in contexts where the amount of data is huge, complex and rather unstructured. As a consequence many of the querying languages not only accommodate recursion but also some nesting of path queries. For the practical evaluation of such queries we plan to explore several methods that range from optimization based on statistics, compilation and machine learning.

Path queries induce naturally a particular way of traversing datagraphs, however such traversals my visit a large number of nodes. Following other traversals could limit the number of visited nodes and then speed up the query evaluation process. Using statistics computed on the graph, we shall explore optimization methods that find good traversals. Once such a traversal is found it will be compiled to a Datalog program so as to implement recursion. We will benefit here from optimization methods developed for Datalog and that will also restrict the part of the graph that is visited. Our goal is to propose algorithms that perform well on large graphs.

When the graph is too large, it may not be reasonable to try to find exact solutions. Practical uses of such graphs should be limited to simple queries. We think however that recursion should still be preserved. We propose to limit ourselves in finding path (possibly obeying certain regular

constraints). We are going to explore approximate representations, called embeddings, of the graph topology obtained with machine learning techniques. This poses several challenges. First we need to develop methods that allow us to compute them. Second, we need to find ways of using these embeddings so as to answer well to queries. And finally establish which kinds embeddings are more suited to certain datagraphs topology.

6.2 Research axis: Monitoring Data Graphs

Programming languages are connected to data processing in at least two ways: first to build applications that interact with databases via updates and queries; second complex computations are required when answering queries typically during the construction of complex aggregates. We are interested in developing programming language techniques that address these two aspects in the context of datagraphs rather than with traditional relational databases. Moreover, we shall take into account the dynamic aspects of datagraphs which shall evolve through updates. The methods we shall develop will monitor changes in datagraphs and react according to the modifications.

Functional Programming Languges for Data Graphs. The integration of programming languages with data is most mature in data trees. We will take inspiration on the ideas that have been developed in that context. Our first aim is to develop the integration of a query language that can rely on programming primitives so as to produce structured outputs. As second is to develop a orchestration language that composes various queries with structured output into a pipeline and ultimately into a service.

Concerning the query language we will base follow ideas coming from functional progamming languages and XPath 3.0. Empirical experiments and prototype developments will take stock of previous work for data trees, such as the language X-Fun and λ-XP. Mixing techniques from query optimization, program analysis, type checking we will develop highly efficient optimization technique for this language. Furthermore we will apply static analysis techniques based on automata and transducers to verify query behaviors.

Concerning the description services that use datagraphs as a backend for storing data, functional programming seems also a good candidate. However, orchestration requires the synchronization of concurrent executions of queries so as to ensure the correct behavior of services. Concurrent constructs need to be built in the language. The high level of concurrence enabled by the notion of *futures* seems an interesting candidate for this task.

Hyperstreaming. We see complex-event datagraph processes as programs that stream datagraphs: read them and output them for futher processing. As usual in such system long computations of a given process may block a whole chain of processes. To improve the responsiveness of these systems, we propose the notion of *hyperstreaming*. The idea takes foundations on automata based streaming methods that we developed during the evaluation period. Hyperstreams generalize streams in that they contain *holes* (unknown data) that can be filled later by some process. In order to avoid suspension as possible, a process monitoring a hyperstream performs clever speculative computations the result of which is selected when the hole is filled. The objectives for the next period are to develop tools for hyperstreaming query answering and to lift them to hyperstreaming program evaluation. On the conceptual side, the notion of *certain query answers* is related to hyperstreams but is better grasped by *certain program outputs*.

6.3 Research axis: Graph Data Integration

We intend to continue to develop theoretical foundations and practical tools for integration of linked data with RDF being their principal format. Because from its conception the main credo of

RDF has been "just publish your data," the problem at hand faces two important challenges: data quality and data heterogeneity.

Data Quality with Schemas and Repairing with Inference. Data impurities in RDF may be divided into two categories: *data value errors* and *structural irregularities*. Data value errors e.g., misspellings due to manual data entry, have been investigated thoroughly in the context of relational databases, with solutions ranging from dictionary methods to similarity measures (string edit distance) to rule based data modifications, and we intend to investigate adapting the existing approaches to RDF. The main challenge in handling data quality issue in RDF springs from high degree of freedom in how information is structured in RDF, the same information may be represented with a different structure within the same document. We intend to develop intelligent approaches that identify data impurities and structural violations and use interaction with expert users to identify the suitable repairing actions.

We plan to define suitable measures for quality of RDF documents. Our approaches will be based on a schema language, such as ShEx and SHACL, and we shall explore suitable variants of graph alignment and graph edit distance to capture similarity between the existing RDF document and its possible repaired versions that satisfy the schema. The central problem is identifying graph fragments that are responsible for failure to satisfy the schema, a nontrivial task given the potentially unbound and cyclic structure of a graph. Naturally, the data quality measure would yield for each offending fragment a set of possible reparations, although such sets can be very large and we intend to explore enumeration approaches to identify the most relevant ones. Additionally, we intend to define a rule based language for defining the desired reparations. Our intention is to assist an expert user in repairing a potentially large RDF document: the user is presented with offending fragments and for each makes a choice among a list of possible reparations; user choices are intelligently generalized to a set of rules that are applied to remaining offending fragments. We plan to develop practical RDF repairing systems in collaboration with industrial partners.

Integration and Graph Mappings with Schemas and Inference. The second problem pertaining to integration of RDF data sources is their heterogeneity. We intend to continue to identify and study suitable classes of mappings between RDF documents conforming to potentially different and complementary schemas. We intend to assist the user in constructing such mappings by developing rich and expressive graphical languages for mappings. Also, we wish to investigate inference of RDF mappings with the active help of an expert user. We will define interactive protocols that allow the input to be sufficiently informative to guide the inference process while avoiding the pitfalls of user input being too ambiguous and causing combinatorial explosion.

7 Bibliography of the project-team

The bibliography is limited to the evaluation period. See below for references to older publications by the team or to publications by others. The `Export_RA.bib` *file is to be generated from HAL using* `haltools.inria.fr`*. TBR.*

Doctoral Dissertations and Habilitation Theses

[1] A. BOIRET, Normalization and Learning of Transducers on Trees and Words, Theses, Université de Lille, November 2016, [hal:tel-01396543].

[2] A. LEMAY, Machine Learning Techniques for Semistructured Data, Habilitation à diriger des recherches, Université de Lille, November 2018, [hal:tel-01929944].

[3] T. SEBASTIAN, Evaluation of XPath Queries on XML Streams with Networks of Early Nested Word Automata, Theses, Universite Lille 1, June 2016, [hal:tel-01342511].

Articles in International Journals

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

[5] G. BAGAN, A. BONIFATI, R. CIUCANU, G. FLETCHER, A. LEMAY, N. ADVOKAAT, Generating Flexible Workloads for Graph Databases, *Proceedings of the VLDB Endowment (PVLDB) 9*, 13, June 2016, p. 1457–1460, [hal:hal-01330111].

[6] G. BAGAN, A. BONIFATI, R. CIUCANU, G. FLETCHER, A. LEMAY, N. ADVOKAAT, gMark: Schema-Driven Generation of Graphs and Queries, *IEEE Transactions on Knowledge and Data Engineering 29*, 4, April 2017, p. 856–869, [doi:10.1109/TKDE.2016.2633993], [hal:hal-01402575].

[7] B. BERGOUGNOUX, F. CAPELLI, M. M. KANTÉ, Counting Minimal Transversals of β-Acyclic Hypergraphs, *Journal of Computer and System Sciences*, May 2019, https://arxiv.org/abs/1808.05017, [doi:10.1016/j.jcss.2018.10.002], [hal:hal-01923090].

[8] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN, Automata for Unordered Trees, *Information and Computation 253*, April 2017, p. 304–335, [doi:10.1016/j.ic.2016.07.012], [hal:hal-01179493].

[9] A. BOIRET, V. HUGOT, J. NIEHREN, R. TREINEN, Logics for Unordered Trees with Data Constraints, *Journal of Computer and System Sciences*, January 2019, p. 40, [doi:10.1016/j.jcss.2018.11.004], [hal:hal-01176763].

[10] A. BONIFATI, R. CIUCANU, S. STAWORKO, Learning Join Queries from User Examples, *ACM Transactions on Database Systems 40*, 4, February 2016, p. 24:1–24:38, [hal:hal-01187986].

[11] P. BOURHIS, C. RIVEROS, S. STAWORKO, G. PUPPIS, Bounded Repairability for Regular Tree Languages, *ACM Transactions on Database Systems 41*, 3, June 2016, p. 1–45, [doi:10.1145/2898995], [hal:hal-01411116].

[12] F. CAPELLI, Y. STROZECKI, Incremental delay enumeration: Space and time, *Discrete Applied Mathematics*, August 2018, https://arxiv.org/abs/1703.01928, [doi:10.1016/j.dam.2018.06.038], [hal:hal-01923091].

[13] L. DAVIAUD, C. PAPERMAN, Classes of languages generated by the Kleene star of a word, *Information and Computation 262*, Part 1, October 2018, p. 90–109, [hal:hal-01943493].

[14] D. DHALI, F. COUTTE, A. A. ARIAS, S. S. AUGER, V. V. BIDNENKO, G. CHATAIGNÉ, M. LALK, J. NIEHREN, D. S. JOANA, C. VERSARI, P. JACQUES, Genetic engineering of the branched fatty acid metabolic pathway of Bacillus subtilis for the overproduction of surfactin C14 isoform, *Biotechnology Journal*, April 2017, p. 23, [doi:10.1002/biot.201600574], [hal:hal-01502183].

[15] N. FIJALKOW, C. PAPERMAN, Monadic Second-Order Logic with Arbitrary Monadic Predicates, *ACM Transaction on the Web 9*, 3, 2017, p. 39 – 56, [doi:10.1145/3091124], [hal:hal-01587624].

[16] G. MADELAINE, C. LHOUSSAINE, J. NIEHREN, E. TONELLO, Structural simplification of chemical reaction networks in partial steady states, *BioSystems 149*, November 2016, p. 34–49, This is a journal extension of a paper published at the CMSB'2015 conference, [doi:10.1016/j.biosystems.2016.08.003], [hal:hal-01350517].

[17] G. MADELAINE, E. TONELLO, C. LHOUSSAINE, J. NIEHREN, Simplification of Reaction Networks, Confluence and Elementary Modes, *Computation*, February 2017, Extended version of a conference paper at CMSB'2016, [hal:hal-01471074].

[18] J. NIEHREN, C. VERSARI, M. JOHN, F. COUTTE, P. JACQUES, Predicting Changes of Reaction Networks with Partial Kinetic Information, *BioSystems 149*, July 2016, p. 113–124, [hal:hal-01239198].

International Conferences with Proceedings

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

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

[21] A. AMARILLI, P. BOURHIS, L. JACHIET, S. MENGEL, A Circuit-Based Approach to Efficient Enumeration , *in : ICALP 2017 - 44th International Colloquium on Automata, Languages, and Programming*, I. Chatzigiannakis, P. Indyk, A. Muscholl (editors), p. 1–15, Varsovie, Poland, July 2017, [doi:10.4230/LIPIcs.ICALP.2017.111], [hal:hal-01639179].

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

[23] A. AMARILLI, P. BOURHIS, P. SENELLART, Tractable Lineages on Treelike Instances: Limits and Extensions, *in : PODS (Principles of Database Systems)*, p. 355–370, San Francisco, United States, June 2016, [hal:hal-01336514].

[24] A. AMARILLI, C. PAPERMAN, Topological Sorting with Regular Constraints, *in : ICALP 2018 - 45th International Colloquium on Automata, Languages, and Programming*, *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018*, *45*, 1, p. 115:1–115:14, Prague, Czech Republic, July 2018, [hal:hal-01950909].

[25] N. BACQUEY, E. GRANDJEAN, F. OLIVE, Definability by Horn formulas and linear time on cellular automata, *in : ICALP 2017 - 44th International Colloquium on Automata, Languages and Programming*, I. Chatzigiannakis, P. Indyk, F. Kuhn, A. Muscholl (editors), *80*, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, p. 1–14, Warsaw, Poland, July 2017, [doi:10.4230/LIPIcs.ICALP.2017.99], [hal:hal-01494246].

[26] G. BAGAN, A. BONIFATI, R. CIUCANU, G. FLETCHER, A. LEMAY, N. ADVOKAAT, gMark: Schema-Driven Generation of Graphs and Queries, *in : Data Engineering (ICDE), 2017 IEEE 33rd International Conference on*, p. 63–64, San Diego, United States, April 2017, [doi:10.1109/ICDE.2017.38], [hal:hal-01591706].

[27] M. BENEDIKT, P. BOURHIS, M. V. BOOM, Characterizing Definability in Decidable Fixpoint Logics , *in : ICALP 2017 - 44th International Colloquium on Automata, Languages, and Programming*, I. Chatzigiannakis, P. Indyk, F. Kuhn, A. Muscholl (editors), *107*, p. 14, Varsovie, Poland, July 2017, [doi:10.4230/LIPIcs.ICALP.2017.107], [hal:hal-01639015].

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

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

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

[31] A. BOIRET, R. PALENTA, Deciding Equivalence of Linear Tree-to-Word Transducers in Polynomial Time, *in : 20th International Conference on Developments in Language Theory (DLT 2016)*, *Developments in Language Theory - 20th International Conference, DLT 2016, Montréal, Canada, July 25-28, 2016, Proceedings*, *9840*, Springer, p. 355–367, Montreal, Canada, July 2016, [doi:10.1007/978-3-662-53132-7_29], [hal:hal-01429110].

[32] A. BOIRET, Normal Form on Linear Tree-to-word Transducers, *in : 10th International Conference on Language and Automata Theory and Applications*, J. Janoušek, C. Martín-Vide (editors), Prague, Czech Republic, March 2016, [hal:hal-01218030].

[33] 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, [hal:hal-01590350].

[34] 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*, p. 1–16, Cali, Colombia, May 2018, [hal:hal-01775199].

[35] 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, [hal:hal-01609498].

[36] 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, March 2019, [hal:hal-01811835].

[37] P. BOURHIS, M. BENEDIKT, B. TEN CATE, G. PUPPIS, Querying Visible and Invisible Information, *in : LICS 2016 - 31st Annual ACM/IEEE Symposium on Logic in Computer Science*, p. 297–306, New York City, United States, July 2016, [doi:10.1145/2933575.2935306], [hal:hal-01411118].

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

[39] P. BOURHIS, M. LECLÈRE, M.-L. MUGNIER, S. TISON, F. ULLIANA, L. GALOIS, Oblivious and Semi-Oblivious Boundedness for Existential Rules, *in : IJCAI: International Joint Conference on Artificial Intelligence*, Macao, China, August 2019, [hal:lirmm-02148142].

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

[41] P. BOURHIS, M. MORAK, A. PIERIS, Making Cross Products and Guarded Ontology Languages Compatible, *in : IJCAI 2017 - Twenty-Sixth International Joint Conference on Artificial Intelligence*, p. 880–886, Melbourne, Australia, August 2017, [doi:10.24963/ijcai.2017/122], [hal:hal-01638346].

[42] P. BOURHIS, J. L. REUTTER, F. SUÁREZ, D. VRGOČ, JSON: Data model, Query languages and Schema specification, *in : PODS 2017 - Proceedings of the Thirty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, Chicago, United States, May 2017, [doi:10.1145/3034786.3056120], [hal:hal-01639182].

[43] P. BUNEMAN, S. STAWORKO, RDF Graph Alignment with Bisimulation, *in : VLDB 2016 - 42nd International Conference on Very Large Databases*, *Proceedings of the VLDB Endowment*, *9*, 12, p. 1149 – 1160, New Dehli, India, September 2016, [doi:10.14778/2994509.2994531], [hal:hal-01417156].

[44] F. CAPELLI, S. MENGEL, Knowledge Compilation, Width and Quantification, *in : 36th International Symposium on Theoretical Aspects of Computer Science (STACS 2019)*, Berlin, Germany, March 2019. https://arxiv.org/abs/1807.04263, [hal:hal-01836402].

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

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

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

[48] 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*, ACM Press, p. 303–319, Amsterdam, Netherlands, June 2019, [doi:10.1145/3294052.3319687], [hal:hal-01959143].

National Conferences

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

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

Scientific Books or Book Chapters)

[51] J. E. L. GAYO, E. PRUD'HOMMEAUX, I. BONEVA, D. KONTOKOSTAS, Validating RDF Data, *7*, 1, Morgan & Claypool, September 2017, [doi:10.2200/S00786ED1V01Y201707WBE016], [hal:hal-01667426].

Research Reports

[52] S. SALVATI, On is an n-MCFL, *Research report*, Université de Lille, INRIA, CRIStAL CNRS, April 2018, [hal:hal-01771670].

Miscellaneous

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

[54] A. AMARILLI, P. BOURHIS, M. MONET, P. SENELLART, Combined Tractability of Query Evaluation via Tree Automata and Cycluits (Extended Version), https://arxiv.org/abs/1612.04203 - 69 pages, accepted at ICDT'17. Appendix F contains results from an independent upcoming journal paper by Michael Benedikt, Pierre Bourhis, Georg Gottlob, and Pierre Senellart, December 2016, [hal:hal-01439309].

[55] A. BOIRET, A. LEMAY, J. NIEHREN, Learning Top-Down Tree Transformations with Regular Inspection, working paper or preprint, August 2016, [hal:hal-01357631].

[56] A. BOIRET, A. LEMAY, J. NIEHREN, A Learning Algorithm for Top-Down Tree Transducers, working paper or preprint, July 2017, [hal:hal-01357627].

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

[58] I. BONEVA, J. DUSART, D. FERNÁNDEZÁLVAREZ, J. E. L. GAYO, Shape Designer for ShEx and SHACL Constraints, ISWC 2019, October 2019, Poster, [hal:hal-02268667].

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

[60] I. BONEVA, J. NIEHREN, M. SAKHO, Approximating Certain Query Answers on Nested Hyperstreams, working paper or preprint, April 2019, [hal:hal-02092276].

[61] I. BONEVA, Comparative expressiveness of ShEx and SHACL (Early working draft), working paper or preprint, March 2016, [hal:hal-01288285].

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

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

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