RESEARCH CENTRE

Inria Center
at the University of Lille

IN PARTNERSHIP WITH:

CNRS, Université de Lille

2022 ACTIVITY REPORT

Project-Team LINKS

Linking Dynamic Data

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

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Inria

Contents

Project-Team LINKS

Creation of the Project-Team: 2016 June 01

Keywords

Computer sciences and digital sciences

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 members, visitors, external collaborators

Research Scientists

- Mikael Monet [INRIA, Researcher]
- Joachim Niehren [INRIA, Senior Researcher, HDR]

Faculty Members

- Sylvain Salvati [Team leader, UNIV LILLE, Professor, HDR]
- Iovka Boneva [UNIV LILLE, Associate Professor]
- Florent Capelli [UNIV LILLE, Associate Professor]
- Aurélien Lemay [UNIV LILLE, Associate Professor, HDR]
- Charles Paperman [UNIV LILLE, Associate Professor]
- Slawomir Staworko [UNIV LILLE, Associate Professor, HDR]
- Sophie Tison [UNIV LILLE, Professor, HDR]

PhD Students

- Antonio Al Serhali [INRIA]
- Corentin Barloy [UNIV LILLE]
- Oliver Irwin [UNIV LILLE, from Dec 2022]
- Claire Soyez-Martin [INRIA]

Technical Staff

- Nicolas Crosetti [INRIA, Engineer, until Aug 2022]

Administrative Assistants

- Nathalie Bonte [INRIA]
- Karine Lewandowski [INRIA]

2 Overall objectives

We 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.

2.1 Presentation

The following three items summarize our main research objectives.

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

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

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

3 Research program

3.1 Background

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

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

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

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

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

semi-structured datasets. We will study how to automatize parts of this search process by developing symbolic machine learning techniques for linked data collections.

3.2 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 low level representations 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.

3.2.1 AI: Circuits for Data Analysis

Circuits are concise 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 disjoint 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 based on representation that are often much smaller. Among others, such applications range from knowledge representation/compilation, counting the number of solutions of queries, efficient query answering, factorized databases.

In a first line of research, we want to study novel problems on circuits, in which database queries are relevant to data analysis tasks from artificial intelligence, in machine learning or data mining in particular. In particular we propose to study optimization problems on answer sets of database queries based on circuits, i.e. how to find optimal solutions in the answer set for a given set of conditions. Decompressing small circuits into large answer sets would make the optimization problem unfeasible in many cases. We believe that circuits can structure certain optimization problems in such a way that it can be phrased concisely and then solved efficiently.

Second, we propose to develop a tighter integration between circuits and databases. Indeed query-related circuits are generally produced from a database. This requires that the data is copied within the circuits. This memory cost is accompanied with the loss of the environment of the DBMS which allows many optimizations and uses many low level optimizations that are hard to implement. We propose then to encode circuits directly within the database using materialized views and index structures. We shall also develop the required computational tools for maintaining and exploiting these embedded circuits.

3.2.2 Path Query Optimization

Graph databases are easily queried using 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. The natural theoretical tool that pertains to recursion in the context of relational data Datalog. There has been a wealth of optimization algorithms that have been proposed for queries written in Datalog. We propose to use Datalog as a low level language to which we will compile path queries of various kinds. The idea is that the compiler will try to obtain Datalog programs that will have low execution complexity taking advantages of optimization techniques such as magic supplementary set rewriting, pre-computed indexes and also statistics computed from the graph. Our goal is to develop a compiler that will be able to efficiently evaluate path queries on large graphs which in particular will explore only a part of it.

3.3 Research Axis: Monitoring Data Graphs

Traditional database applications are programs that interact with database via updates and queries. We are interested in developing programming language techniques so as to interact with datagraphs rather than with traditional relational databases. Moreover, we shall take into account the dynamic aspects of datagraphs which shall evolve through updates. We will develop methods to monitor changes in datagraphs and react according to the modifications.

3.3.1 Functional Programming Languges for Data Graphs

The first question is which kind of programming language to use to enable monitoring processes for data graphs based on query answering. While languages of path queries found quite some interest on data graphs, less attention has been given to the programming language tasks, that needed to be solved to produce structured output and to compose various queries with structured output into a pipeline. We believe that transferring the generalization of ideas developed for data trees in the context of XML to data graphs will allow to solve such problems in a systematic manner.

Our approach will consist in developing a functional programming language based on first principles (the lambda calculus, graph navigation, logical connective) that generalizes full XPath 3.0 to the context of graphs. Here we can rely on own previous work for data trees, such as the language X-Fun and λ-XP. After the language for data graphs is designed we shall study its behavior empirically by means of an implementation. This implementation will help us to design optimization methods so as to evaluate the queries in that language. This will allow us to use a wealth of techniques so as to optimize the computation. Indeed, we can try to compile data structures to imperative ones when possible and also exploit possibilities of parallel executions in certain cases. Functional programming comes with nice verification techniques that we are going to use in several contexts: (i) in optimizing queries (e.g. stop the evaluation when it is possible to know that no more data can contribute to the output) and (ii) to verify that the query behaves correctly. The verification methods we shall focus on will be mainly related to automata and transducers.

Finally we shall also develop a programming language that allows to describe services that use datagraphs as a backend for storing data. Here again, functional programming seems a good candidate, we would need however to orchestrate the concurrent executions of queries so as to ensure the correct behavior of services. This means that we should have concurrent constructs that are built in the language. The high level of concurrence enabled by the notion of *futures* seems an interesting candidate to adapt to the context of service orchestration.

3.3.2 Hyperstreaming Program Evaluation

Complex-event processing requires to monitor data graphs that are produced on input streams and to write data graphs to some output stream, which can then be used as inputs again. A major problem here is to reduce the high risk of blocking, which arises when the writing of some of the output stream suspends on a data value that will become available only in the future on some input stream. In such cases, all monitoring processes reading the output stream may have to suspend as well. In order to reduce the risk of blocking, we propose to develop the hyperstreaming approach further, of which we layed the foundations in the evaluation period based on automata techniques. The idea is to generalize streams to hyperstreams, i.e. to add holes to streams that can be filled by some other stream in the future. In order to avoid suspension as possible, a monitoring process on hyperstream must then be able to jump over the holes, and to perform some speculative computation. The objective for the next period are to develop tools for hyperstreaming query answering and to lift these to hyperstreaming program evaluation. Furthermore, on the conceptual side, the notion of certain query answers on hyperstreams needs to be lifted to certain program outputs on hyperstreams.

3.4 Research Axis: Graph Data Integration

We intend to continue to develop 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.

3.4.1 Data Quality with Schemas and Repairing with Inference

The data quality of RDF may suffer due to a number of reasons. Impurities may arise due to data value errors (misspellings, errors during data entry etc.). Such data quality problems have been thoroughly investigated in literature for relational databases and solutions include dictionary methods etc. However, it remains to be seen if the challenges of adapting the existing solutions for relational databases can be easily addressed.

One particular challenge comes from the fact that RDF allows a higher degree of structural freedom in how information is represented as opposed to relation databases, where the choice is strongly limited to flat tables. We plan to investigate suitability of existing data cleaning methods to tackle the problems of data value impurities in RDF. The structural freedom of RDF is a source of data quality issues on its own. With the recent emergence of schema formalisms for RDF, it becomes evident that significant parts of existing RDF repositories do not necessarily satisfy schemas prepared by domain experts.

In the first place, we intend to investigate defining suitable measures of quality for 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 issue here is repairing an RDF document w.r.t. schema by identifying essential fragments of the RDF that fail to satisfy the schema. Once such fragments are identified, repairing actions can be applied however there might be a significant number of alternatives. We intend to explore enumeration approaches where the space of repairing alternatives is intelligently browsed by the user and the most suitable one is chosen. Furthermore, we intend to propose a rule language for choosing the most suitable repairing action and will investigate inference methods to derive from interactions with user the optimal order in which various repairing actions are presented to the user and derive the rules for the choice of the preferred repairing action for repeating types of fragments that do not satisfy the schema.

3.4.2 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 need to define interactive protocols that allows 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. We intend to identify

RDF Data Quality. Approach based on a schema language (ShEx or SHACL) used to identify errors and giving a notion of a measure of quality of an RDF database. Impurities in RDF may come from data value errors (misspellings etc.) but also from the fact that RDF imposes fewer constraints on how data is structured which is a consequence of a significantly different use philosophy (just publish your data anyway you want). Repairing of RDF errors would be modeled with a localized rules (transformations that operate within a small radius of an affected node) and if several rules apply, preferences are used to identify the most desirable one. Both the repairing rules and preferences can be inferred with the help of inference algorithms in an interactive setting. Smart tools for LOD integration. Assuming that the LOD sources are of good quality, we want to build tools that assist the user in constructing mappings that integrate data in the user database. For this, we want to define inference algorithms which are guided by schemas, and which are based on comprehensible interactions with the user. For this, we need to define interactions that are rich enough to inform the algorithm, while simple enough to be understandable by a non-expert user. In particular, that means that we need to present data (nodes in a graph for instance) in a readable way. Also, we want to investigate how the - possibly inferred - schema can be used to guide the inference.

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 Social and environmental responsibility

5.1 Footprint of research activities

Tison is a member of the general assembly of the European Association for Theoretical Computer Science (EATCS) (elected in 2019).

5.2 Impact of research results

Databases and methods from Artificial Intelligence are used in virtually all aspects of the modern digitalized world, from companies' web services to governments' institutions.

6 Highlights of the year

The team has had three accepted articles at the STACS'2023 conference, a renowned international conference in theoretical computer science. This makes LINKS represent about 5% of STACS'2023 accepted papers. The articles in question are:

- Florent Capelli and Yann Strozecki's paper on *Geometric Amortization of Enumeration Algorithms* [28];
- Antoine Amarilli and Mikaël Monet's paper on *Enumerating Regular Languages with Bounded Delay* [29];
- Charles Paperman, Sylvain Salvati and Claire Soyez-Martin's paper on *An algebraic approach to vectorial programs* [10].

We refer to Section 8 for short summaries of these contributions.

7 New software and platforms

7.1 New software

7.1.1 NetworkDisk

Name: NetworkDisk

Keywords: Large graphs, Python, Databases

Functional Description: NetworkDisk provides a way to manipulate graphs on disk. The goal is to be as much as possible compatible with (Di)Graph objects of the NetworkX Python package but lifting memory requirement and providing persistence of the Graph.

URL: https://networkdisk.inria.fr/

Contact: Charles Paperman

7.1.2 Bibendum

Name: Bibendum

Keyword: Bibliography

Functional Description: Small app to fetch bibtex from a short label with the format: LastName.Year.PublicationTerm where the . denote the concatenation. For instance Codd1970Relational. LastName is the last name of one of the authors. Year is the publication year. PublicatonTerm is one meaningful word in the title of the publication.

In case of ambiguity, an extra integer is used. Ambiguous entries are resolved by sorting dois under lexicographical order. The api is idempotent, every decision taken is recorded and replayed.

URL: https://gitlab.inria.fr/cpaperma/bibendum

Contact: Charles Paperman

7.1.3 XPath AutoBench

Name: A Benchmark Collection of Deterministic Automata for XPath Queries

Keywords: XML, Querying, Tree Automata

Functional Description: We provide a benchmark collection of deterministic automata for regular XPath queries. For this, we select the subcollection of forward navigational XPath queries from a corpus that Lick and Schmitz extracted from real-world XSLT and XQuery programs, compile them to stepwise hedge automata (SHAs), and determinize them. Large blowups by automata determinization are avoided by using schema-based determinization. The schema captures the XML data model and the fact that any answer of a path query must return a single node. Our collection also provides deterministic nested word automata that we obtain by compilation from deterministic SHAs.

URL: https://archive.softwareheritage.org/browse/origin/directory/?origin_url=https://gitlab.inria.fr/aalserha/xpath-benchmark

Contact: Joachim Niehren

7.1.4 rsonpath

Keywords: JSon, Streaming, SIMD, Rust

Functional Description: The rsonpath crate provides a JSONPath parser and a query execution engine, which utilizes SIMD instructions to provide massive throughput improvements over conventional engines.

URL: https://github.com/V0ldek/rsonpath

Contact: Charles Paperman

Partner: Warsaw University

7.1.5 Coussinet

Name: Coussinet

Keywords: Enumeration, Complexity

Functional Description: Coussinet is a demo illustrating a technique called geometric amortization for enumeration algorithms introduced in the paper Geometric Amortization for Enumeration Algorithms, Florent Capelli, Yann Strozecki. The result presented in this paper is about making the delay of enumeration algorithms more regular.

URL: http://florent.capelli.me/coussinet/coussinet.html

Contact: Florent Capelli

Participants: Florent Capelli, Yann Strozecki

7.1.6 ShEx validator

Name: 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 Contributions: 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.

URL: https://github.com/iovka/shex-java

Contact: Iovka Boneva

7.1.7 gMark

Name: 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

URL: https://github.com/graphMark/gmark

Contact: Aurélien Lemay

8 New results

Participants: Antonio Al Serhali, Corentin Barloy, Iovka Boneva, Florent Capelli, Nicolas Crosetti, Aurélien Lemay, Mikael Monet, Joachim Niehren, Charles Paperman, Sylvain Salvati, Claire Soyez-Martin, Slawomir Staworko, Sophie Tison.

8.1 Querying Data Graphs

8.1.1 Circuits for Data Analysis in Artificial Intelligence

In their ICDT'2022 article [21], Capelli, Crosetti, Niehren and Ramon study the problem of optimizing a linear program whose variables are answers to a conjunctive query. For this they propose a new language for specifying linear programs whose constraints and objective functions depend on the answer sets of conjunctive queries. They developed an efficient algorithm for solving programs in a fragment of this language. Using tools from knowledge compilation, and exploiting the structure of queries having small *(fractional) treewidth* (a hypergraph parameter, intuitively measuring how far a hypergraph is from being acyclic), they are able to constructs a linear program having the same optimal value but fewer variables, thus nontrivially improving the asymptotic complexity of solving this task. They moreover illustrate the application of their language on three examples: optimizing deliveries of resources, minimizing noise for differential privacy, and computing the s-measure of patterns in graphs as needed for data mining.

In a LICS'2022 article [20], Charles Paperman, his PhD student Corentin Barloy, and others, characterize the regular languages with a neutral letter expressible in firstorder logic with one alternation. Specifically, they show that if an arbitrary Σ_2 formula (a syntactical fragment of first order logic over words) defines a regular language with a neutral letter, then there is an equivalent Σ_2 formula that only uses the order predicate. This shows that the so-called Central Conjecture of Straubing holds for Σ_2 over languages with a neutral letter, the first progress on the Conjecture in more than 20 years. To show the characterization, they establish new lower bounds against polynomial-size depth-3 Boolean circuits with

constant top fan-in.

In an article recently accepted at STACS'2023 [30], Charles Paperman, Sylvain Salvati, and their PhD student Claire Soyez-Martin develop theoretical aspects of *vectorial programming*, the combination of SIMD instructions with usual processor instructions. This new technology is known to speed-up many standard algorithms, such as for simple regular languages. Their idea is to take advantage of the inner algebraic structure of regular languages and produce high level representations of efficient vectorial programs that recognize certain classes of regular languages. As a technical ingredient, they establish equivalences between classes of vectorial circuits and logical formalisms, namely unary temporal logic and first order logic. Their main result is the construction of compilation procedures that turns syntactic semigroups into vectorial circuits. The circuits they obtain are small in that they improve known upper-bounds on representations of automata within the logical formalisms. The gain is mostly due to a careful sharing of sub-formulas based on algebraic tools.

In an Acta Informatica journal article [14], Paperman and Cadilhac characterize the regular languages of wire linear AC0 (a certain class of Boolean circuits) as the languages expressible in the two-variable fragment of first-order logic with regular predicates. Equivalently, they show that this corresponds to languages that are recognized by the algebraic class QLDA, which they prove is decidable. Examples of languages in and outside of this class are presented.

8.1.2 Uncertainty and Explanations

In a SIGMOD'2022 paper [22], Monet et al. use the framework of *Shapley values* to assign and compute contributions of input facts of a database to the results of a query. The goal is, intuitively, to *explain* the results of a query by computing a score for every input fact. The Shapley value is a game-theoretic notion for wealth distribution that is nowadays extensively used to explain complex data-intensive computation, for instance, in network analysis or machine learning. Monet et al. present in this paper two practically effective solutions for computing Shapley values in query answering. First, they establish a tight theoretical connection to the extensively studied problem of *query evaluation over probabilistic databases*, which allows then to obtain a polynomial-time algorithm for the class of queries for which probability computation is tractable. They then propose a first practical solution for computing Shapley values that adopts tools from probabilistic query evaluation and knowledge compilation. Experiments are caried that demonstrate the practical effectiveness of their solutions.

In an MFCS'2022 article [19], Amarilli and Monet study the computational complexity of computing the probability of a obtaining a *matching* in a graph with probabilistic edges. Specifically, they consider the problem denoted PrMatching($\mathcal{G}$), on an arbitrary fixed graph family $\mathcal{G}$. The input consists of a graph $G \in \mathcal{G}$ and of rational probabilities of existence on every edge of G, assuming independence. The output is the probability of obtaining a *matching* of G in the resulting distribution, i.e., a set of edges that are pairwise disjoint. It is known that, if $\mathcal{G}$ has bounded treewidth, then PrMatching($\mathcal{G}$) can be solved in polynomial time. In this paper they establish that, under some assumptions, bounded treewidth in fact *characterizes* the tractable graph families for this problem. More precisely, they show intractability for all graph families $\mathcal{G}$ satisfying the following *treewidth-constructibility* requirement: given an integer k in unary, we can construct in polynomial time a graph $G \in \mathcal{G}$ with treewidth at least k. Their hardness result is then the following: for *any* treewidth-constructible graph family $\mathcal{G}$, the problem PrMatching($\mathcal{G}$) is intractable. This generalizes known hardness results for weighted matching counting under some restrictions that do not bound treewidth, e.g., being planar, 3-regular, or bipartite. They also obtain a similar lower bound for the weighted counting of edge covers.

8.1.3 Query Optimization

In a PVLDB article [13], Staworko et al. study *threshold queries*, that is, queries that only require computing or counting answers up to a specified threshold value. This type of query is very common in practice, yet has been little studied. In this paper, they present a theoretical analysis of threshold query evaluation and show that thresholds can be used to significantly improve the asymptotic bounds of state-of-the-art query evaluation algorithms. In surprising contrast to conventional wisdom, they found important scenarios

in real-world data sets in which users are interested in computing the results of queries up to a certain threshold, independent of a ranking function that orders the query results by importance.

8.2 Monitoring Data Graphs

8.2.1 Functional Programming Languages for Data Trees

In an ICLP'2022 article [25], Niehren and Salvati propose a new algorithm for evaluating nested regular path queries on graphs from a set of start nodes in combined linear time. They show that this complexity upper bound can be reduced by making it dependent on the size of the query's top-down needed subgraph, a notion that they introduce. For many queries in practice, the top-down needed subgraph is way smaller than the whole graph. Their algorithm is based on a novel compilation schema from nested regular path queries to monadic datalog queries. Its complexity upper bound follows from known properties of top-down datalog evaluation.

8.2.2 Query Answering on Streams

Boneva, Niehren et al. [12] study the complexity of regular matching and inclusion for compressed tree patterns with context variables subject to regular constraints. Context variables with regular constraints permit to properly generalize on unranked tree patterns with hedge variables. Regular inclusion on unranked tree patterns is relevant to certain query answering on Xml streams with references. They prove that regular matching and inclusion with regular constraints can be reduced in polynomial time to the corresponding problem without regular constraints.

Niehren and his PhD student Antonio Serhali study schema-based automaton determination in a Gandalf'2022 article [24]. They propose an algorithm for this task, over finite automata on words and stepwise hedge automata on nested words. Their idea is to integrate schema-based cleaning directly into automata determinization. They prove the correctness of their new algorithm and show that it is always more efficient than standard determinization followed by schema-based cleaning. Their implementation permits to obtain a small deterministic automaton for an example of an XPath query, where standard determinization yields a huge stepwise hedge automaton for which schema-based cleaning runs out of memory.

In an XML Prague 2022 paper [18], Niehren and Serhali provide a benchmark collection of deterministic automata for regular XPath queries. The benchmark is constructed by selecting the subcollection of forward navigational XPath queries from a pre-existing corpus that was extracted from real-world XSLT and XQuery programs, compiling them to stepwise hedge automata (Shas), and determinizing them. They are able to avoid large blowups by automata determinization by using schema-based determinization.

8.3 Graph Data Integration

8.3.1 Integration and Graph Mappings with Schemas and Inference

In an ICDT'2022 paper [23], Lemay and Staworko, together with Benoît Groz and Piotr Wieczorek, investigate the problem of constructing a shape graph that describes the structure of a given graph database. They employ the framework of grammatical inference, where the objective is to find an inference algorithm that is both sound, i.e., always producing a schema that validates the input graph, and complete, i.e., able to produce any schema, within a given class of schemas, provided that a sufficiently informative input graph is presented. They identify a number of fundamental limitations that preclude feasible inference, and present inference algorithms based on natural approaches that allow to infer schemas that they argue to be of practical importance.

In a PODS'2023 article [27], Boneva, Staworko and others investigate graph transformations defined using Datalog-like rules based on acyclic conjunctive two-way regular path queries (acyclic C2RPQs). They study two fundamental static analysis problems: type checking and equivalence of transformations

in the presence of graph schemas. Additionally, they investigate the problem of target schema elicitation, which aims to construct a schema that closely captures all outputs of a transformation over graphs conforming to the input schema. They show that all these problems are in EXPTIME by reducing them to C2RPQ containment modulo schema; and also provide matching lower bounds.

8.4 Others

Niehren is cooperating with the BioComputing team of the Cristal Lab at the University of Lille since many years. He uses abstract interpretation of logical formulas for predicting gene knockouts based on formal models of reaction network. In cooperation with bio-engineers from Clermont Ferrant and Lille, Niehren presented a novel application of his prediciton algorithms to the overproduction of Mycosubtilin isoforms [16]. This required an extension to the prediction of gene-overexpressions.
In [26], Niehren et al. propose to simulate chemical reaction networks with the deterministic semantics abstractly, without any precise knowledge on the initial concentrations. For this, the concentrations of species are abstracted to Booleans stating whether the species is present or absent, and the derivatives of the concentrations are abstracted to signs saying whether the concentration is increasing, decreasing, or unchanged. They use abstract interpretation over the structure of signs for mapping the ODEs of a reaction network to a Boolean network with nondeterministic updates. Constraints on the abstraction of the initial concentrations can be added naturally, leading to an abstract simulation algorithm that produces only the part of the abstract state transition graph that is reachable from the abstraction of the initial state. They prove the soundness of their abstract simulation algorithm, discuss its implementation, and show its applicability to reaction networks in some particular format from an existing database.

In an Information Processing Letters article [17], Staworko et al. study computational models that perform a folding operations on words of a given language, following directions coded on words of another given language. They consider the case in which both given languages are regular, and show that the class of languages generated by such F-systems is a proper subset of the class of linear context-free languages.

In an article recently accepted at STACS'2023 [28], Florent Capelli and Yann Strozecki introduce the technique of *geometric amortization* for enumeration algorithms. This technique can be used to make the delay of enumeration algorithms more regular without much overhead on the space it uses. More precisely, they are interested in enumeration algorithms having incremental linear delay, that is, algorithms enumerating a set A of size K such that for every $t \leq K$, it outputs at least t solutions in time $O(tp)$, where p is the incremental delay of the algorithm. While it is folklore that one can transform such an algorithm into an algorithm with delay $O(p)$, the naive transformation may blow the space exponentially. They show that, using geometric amortization, such an algorithm can be transformed into an algorithm with delay $O(p \log K)$ and $O(s \log K)$ space, where s is the space used by the original algorithm. They apply geometric amortization to show that one can trade the delay of flashlight search algorithms for their average delay modulo a factor of $O(\log K)$. They illustrate how this tradeoff may be advantageous for the enumeration of solutions of DNF formulas.

In an article recently accepted at STACS'2023 [29], Antoine Amarilli and Mikaël Monet study the task, for a given language L, of enumerating the (generally infinite) sequence of its words, without repetitions, while bounding the delay between two consecutive words. To allow for delay bounds that do not depend on the current word length, they assume a model where one produces each word by editing the preceding word with a small edit script, rather than writing out the word from scratch. In particular, this witnesses that the language is orderable, i.e., one can write its words as an infinite sequence such that the Levenshtein edit distance between any two consecutive words is bounded by a value that depends only on the language. For instance, the language $(a+b)^*$ is orderable (with a variant of the Gray code), but $a^* + b^*$ is not. They characterize which regular languages are enumerable in this sense, and show that this can be decided in PTIME in an input deterministic finite automaton (DFA) for the language. In fact, they show that, given a DFA A, one can compute in PTIME automata $A_1, \ldots, A_t$ such that $L(A)$ is partitioned as $L(A_1) \sqcup \ldots \sqcup L(A_t)$ and every $L(A_i)$ is orderable in this sense. Further, they show that the value of t obtained is optimal, i.e., it is not possible to partition $L(A)$ into less than t orderable

languages. In the case where $L(A)$ is orderable (i.e., $t = 1$), they show that the ordering can be produced by a bounded-delay algorithm: specifically, the algorithm runs in a suitable pointer machine model, and produces a sequence of bounded-length edit scripts to visit the words of $L(A)$ without repetitions, with bounded delay – exponential in $|A|$ – between each script. In fact, they show that we can achieve this while only allowing the edit operations push and pop at the beginning and end of the word, which implies that the word can in fact be maintained in a double-ended queue.

In a paper appeared in the Journal of Computer and System Sciences [15], Kilian Gebhardt, Frédéric Meunier and Sylvain Salvati solved a problem related to commutative properties in formal languages. These properties pose problems at the frontier of computer science, computational linguistics and computational group theory. A prominent problem of this kind is the position of the language O_n, the language that contains the same number of letters a_i and $\bar{a}_i$ with $1 \leq i \leq n$, in the known classes of formal languages. It has recently been shown that O_n is a Multiple Context-Free Language (MCFL). However the more precise conjecture of Nederhof that O_n is an MCFL of dimension n was left open. The paper presents two proofs of this conjecture, both relying on tools from algebraic topology. On the way, it proves a variant of the necklace splitting theorem.

9 Bilateral contracts and grants with industry

Participants: Slawomir Staworko, Sophie Tison.

9.1 Bilateral contracts with industry

Stawarko Academic member of Linked Data Benchmark Council (LDBC).

Staworko Member of Work Group on Property Graph Schemas (standardisation effort).

Tison Vice-president of the Force Awards association.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Participation in other International Programs

Informal International Partners

Tel Aviv, Israel Monet works with Benny Kimelfeld from Technion (Israel) and Daniel Deutch from Tel Aviv University on computing Shapley values for database query answers. A joint paper has been published at SIGMOD'2022 [22].

Warsaw, Poland Paperman cooperates with Filip Murlak on query evaluation on streams.

Saint Petersburg, Russia Niehren and Salvati cooperate with R. Azimov from the University of Saint Petersburg leading to a common publication at ICLP'2022 [25].

10.2 International research visitors

10.2.1 Visits of international scientists

Antoine Amarilli Télécom Paris. Regularly visits the team to collaborate with Paperman and with Monet.

Yann Strozecki Université de Versailles. Regularly visits the team to collaborate with Capelli.

Michaël Cadilhac DePaul University, USA. Research visit from October 15th to November 15th to collaborate with Paperman.

Howard Straubing Boston College, USA. Research visit from October 15th to November 15th to collaborate with Paperman.

In addition, the following researchers have visited LINKS in-person to give talks for the team's seminar.

Arnaud Durand Université Paris Cité. July 2022.

Luis Galárraga Inria Rennes. September 2022.

Liat Peterfreund CNRS. September 2022.

Rémi Morvan Université de Bordeaux. December 2022.

Sarah Winter Université libre de Bruxelles. January 2023.

10.3 National initiatives

ANR JCJC KCODA

Participants: Florent Capelli *(correspondent)*, Charles Paperman, Sylvain Salvati.

- **Duration:** 2021–2025
- **Objectives:** The aim of KCODA is to study how succinct representations can be used to efficiently solve modern optimization and AI problems that use a lot of data. We suggest using data structures from the field of compilation of knowledge that can represent large datasets succinctly by factoring certain parts while allowing efficient analysis of the represented data. The first goal of KCODA is to understand how one can efficiently solve optimization and training problems for data represented by these structures. The second goal of KCODA is to offer better integration of these techniques into the systems of database management by proposing new algorithms allowing to build factorized representations of the data responses to DB requests and by proposing encodings of these representations inside the DB.

ANR Headwork

Participants: Joachim Niehren *(correspondent)*, Momar Ndiouga Sakho, Nicolas Crosetti, Florent Capelli.

- **Duration:** 2016–2022
- **Coordinator:** D. Gross-Amblard, Druid Team, Université de Rennes 1
- **Scientific partners:** Dahu project-team (Inria Saclay) and Sumo project-team (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 Bravas

Participants: Sylvain Salvati *(correspondent)*.

- **Duration:** 2017–2022
- **Coordinator:** Jérôme Leroux, LaBRI, Université de Bordeaux
- **Scientific Partner:** 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.

10.4 Regional initiatives

CPER Cornelia on Artificial Intelligence (2021-2025)

Participants: Joachim Niehren *(correspondent)*.

The whole LINKS project is partner of this new CPER project.

PhD project Antonio al Serhali (2020-...) Cofunded by the Region Hauts-de-France.

Participants: Joachim Niehren.

11 Dissemination

Participants: Iovka Boneva, Florent Capelli, Aurélien Lemay, Mikael Monet, Joachim Niehren, Charles Paperman, Sylvain Salvati, Slawomir Staworko, Sophie Tison.

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

Capelli and Paperman Co-organizers of TUDASTIC, a thematic school on data structures for text indexation and compression. Held in Lille, 9th and 10th May 2022. 1st edition of the thematic school TUtorials on DAta Structures for Text Indexation and Compression.

Capelli and Paperman Co-organizers of TUDASTIC 2023, to be announced.

Tison General Chair with Cédric Lhoussaine of JNIM'22 (JOURNÉES NATIONALES DU GDR IM (March 29th - April 1st))

Tison General Chair with Jacques Sakarovitch of IFIP general Assembly 2022 (September 19-21)

11.1.2 Scientific events: selection

Chair of conference program committees

Niehren is Co-Chair of the International Conference on Systems Biologoy (CMSB'2023) in Luxembourg.

Tison Program Chair of Highlights of Logic, Games and Automata 2022

Member of conference program committees

Lemay Member of the International Conference on Grammatical Inference (ICGI'2022) program committee.

Monet Member of the ACM SIGMOD/PODS'2022 program committee.

Monet Member of the International Conference on Database Theory (ICDT'2022) program committee.

Staworko Member of the ACM SIGMOD/PODS'2022 program committee.

Capelli Member of the IJCAI'2022 program committee.

Capelli Member of the IJCAI'2023 program committee.

11.1.3 Journal

Member of the editorial boards

Niehren Editorial Board of Fundamenta Informaticae.

Niehren Editorial Board of Algorithms.

11.1.4 Scientific expertise

Niehren Reviewer for project proposals by the Vienna Science and Technology Fund in Austria.

Capelli Member of Inria Lille CER (Commission des Emplois de Recherche)

Salvati Member of Inria's Evaluation Committee.

Tison Elected member of CNU 27.

11.1.5 Research administration

Salvati Member of the joint and restricted commissions of the computer science department of Université de Lille (FIL) and of CRIStAL for recruitments.

Staworko Member of the Parity/Equality commission of the CRIStAL laboratory.

Tison Member of the coordinating team of ISite Université de Lille - Nord Europe. Until March 2022.

Tison Member of the school board of IMT Nord Europe. Until July 2022.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching Responsibilities

Capelli Reponsible for Parcoursup for LEA department, Université de Lille.

Salvati Co-director of studies for the Master MIAGE FA, Université de Lille.

Salvati Director of studies for the mathematics and computer science bachelor's degree of Université de Lille.

Salvati Co-responsible for the research track of the computer science bachelor's degree of Université de Lille.

Salvati Board member of the computer science departement of Université de Lille (FIL).

Salvati Director of studies for the Master Informatique, Université de Lille.

Staworko Coordinator of International Relationships at the Department of Computer Science, Université de Lille (FIL).

Tison Member of the selection board for «Capes» in computer science.

11.2.2 Teaching Activities

Boneva Teaches computer science in DUT Informatique of Université de Lille

Capelli Teaches computer science in the LEA department of Université de Lille for around 200h per year (Licence and Master).

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

Monet Teaches computer science as a temporary lecturer for a total of about 80h per year – about 60h for the computer science department of Université de Lille (FIL), and 20h for the computer science department of Centrale Lille. That includes Compilers (M1, 24h), Introduction to Security (18h), Advanced Databases (M1, 27h), Databases 1 (M1, 8h), SQL and Databases (G3 Centrale Lille, 8h).

Niehren Gives lessons for the 2nd year students of the Master Machine Learning (Université de Lille): on Logical foundations of databases (21h).

Paperman Teaches computer science for around 200h per year. He gives lectures for the computer science and math departments. Topics includes an Advance Databases lecture (M2, 24hx3), Algorithmic and Programming in L2 MIASHS (24h), Web Programming in L3 MIASHS and Introduction to Web Technology in L1 SESI (32h).

Salvati Teaches computer science for a total of around 230h per year in computer science departement of Université de Lille. That includes Introduction to Computer Science (L1, 50h), Logic (L3, 50h), Algorithmic and operational research (L3, 36h), Functional Programming (L3, 35h), Research Option (L3, 10h), Semantic Web (M2, 30h), Advanced Databases (M1, 20h).

Staworko Teaches computer science for a total of around 200h in the MIME department (Université de Lille).

Tison Teaches computer science for around 200h at the Universite‘ de Lille. That includes Advanced Algorithms and Complexity (42h, Master), Databases (60h, L2), and Logic (25h, L2) and Algorithmic and Programming in L1 SESI (36h).

11.2.3 Supervision

Al Serhali PhD project started 2020. On hyperstream programming. Supervised by Niehren.

Barloy PhD project started 2021. On circuits and lower complexity bounds. Supervised by Paperman and Salvati.

Hugo Peyraud Magnin L3 intern from ENS Paris, 8 weeks. On perfect matchings in the powerset. Supervised by Monet.

Soyez-Martin PhD project started 2020. On streaming with vectors and circuits. Supervised by Salvati and Paperman.

Oliver Irwin PhD project started 2022. On compilation and aggregation in databases. Supervised by Capelli.

Crosetti PhD project started 2018. On enriching and solving linear programs with conjunctive queries. Supervised by Capelli, Niehren, and Tison. PhD defense planned February 27th.

11.2.4 Juries

Tison Expert for the "Programme Jeunes Talents - For Women in Science Programme" (Loreal Foundation).

Tison Member of the selection committees for: PR Université Côte d'Azur, PR Université d'Aix-Marseille, Pr ENS Saclay, CPJ Université de Lille.

Capelli Member of the PhD committee (examiner) of Alexis de Colnet, on "Hard Functions in Knowledge Compilation: from Lower Bounds to Applications". Université d'Artois, September 2022.

Tison President of the PhD committee of Alexis de Colnet, on "Hard Functions in Knowledge Compilation: from Lower Bounds to Applications". Université d'Artois, September 2022.

Tison President of the PhD committee of Nicolas Berveglieri, on "Optimisation coûteuse multi-objectifs assistée par des modèles de substitution". Université de Lille, November 2022.

Boneva Member of the selection jury for "Prix de thèse BDA 2022"

11.2.5 Internal or external Inria responsibilities

Tison Member of the Scientific Board of XPerium and member of the steering committee of Xperium challenge

11.2.6 Miscellaneous

Tison Member of the steering committee of the mentoring circle "Femmes et sciences", Université de Lille (since July 22).

12 Scientific production

12.1 Major publications

[1] A. Amarilli, L. Jachiet and C. Paperman. 'Dynamic Membership for Regular Languages'. In: ICALP. Vol. 48. International Colloquium on Automata, Languages, and Programming (ICALP 2021). Glasgow, Scotland, France, 2nd July 2021, 116:1–116:17. DOI: 10.4230/LIPIcs.ICALP.2021.116. URL: https://hal.archives-ouvertes.fr/hal-03466453.

[2] M. Arenas, P. Barceló, L. Bertossi and M. Monet. 'The Tractability of SHAP-Score-Based Explanations over Deterministic and Decomposable Boolean Circuits'. In: *Thirty-Fifth AAAI Conference on Artificial Intelligence*. Held online, France, Feb. 2021. URL: https://hal.inria.fr/hal-03147623.

[3] C. Barloy, M. Cadilhac, C. Paperman and T. Zeume. 'The Regular Languages of First-Order Logic with One Alternation'. In: LICS 2022 - 37th Annual ACM/IEEE Symposium on Logic in Computer Science. Haïfa, Israel, 2nd Aug. 2022, pp. 1–11. DOI: 10.1145/3531130.3533371. URL: https://hal.science/hal-03934389.

[4] C. Barloy, F. Murlak and C. Paperman. 'Stackless Processing of Streamed Trees'. In: *2021 PODS*. Xi'an, Shaanx, China, June 2021. DOI: 10.4230/LIPIcs. URL: https://hal.archives-ouvertes.fr/hal-03021960.

[5] I. Boneva, J. G. Labra Gayo and E. G. Prud 'hommeaux. 'Semantics and Validation of Shapes Schemas for RDF'. In: *ISWC2017 - 16th International semantic web conference*. Vienna, Austria, Oct. 2017. URL: https://hal.archives-ouvertes.fr/hal-01590350.

[6] P. Bourhis, M. Leclère, M.-L. Mugnier, S. Tison, F. Ulliana and L. Gallois. 'Oblivious and Semi-Oblivious Boundedness for Existential Rules'. In: *IJCAI 2019 - International Joint Conference on Artificial Intelligence*. Macao, China, Aug. 2019. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02148142.

[7] F. Capelli, N. Crosetti, J. Niehren and J. Ramon. 'Linear Programs with Conjunctive Queries'. In: 25th Internationcal Conference on Database Theory (ICDT 2022). Edinburgh, United Kingdom, 29th Mar. 2022. URL: https://hal.archives-ouvertes.fr/hal-01981553.

[8] P. D. Gallot, A. Lemay and S. Salvati. 'Linear high-order deterministic tree transducers with regular look-ahead'. In: *MFCS 2020 : The 45th International Symposium on Mathematical Foundations of Computer Science*. Andreas Feldmann, Michal Koucky and Anna Kotesovcova. Prague, Czech Republic, Aug. 2020. DOI: 10.4230/LIPIcs.MFCS.2020.34. URL: https://hal.archives-ouvertes.fr/hal-02902853.

[9] J. Niehren and M. Sakho. 'Determinization and Minimization of Automata for Nested Words Revisited'. In: *Algorithms* (Feb. 2021). URL: https://hal.inria.fr/hal-03134596.

[10] C. Paperman, S. Salvati and C. Soyez-Martin. *An algebraic approach to vectorial programs*. 27th Oct. 2022. DOI: 10.4230/LIPIcs.STACS.2023.14. URL: https://hal.archives-ouvertes.fr/hal-03831752.

[11] S. Staworko and P. Wieczorek. 'Containment of Shape Expression Schemas for RDF'. In: *SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*. Amsterdam, Netherlands, June 2019. URL: https://hal.inria.fr/hal-01959143.

12.2 Publications of the year

International journals

[12] I. Boneva, J. Niehren and M. Sakho. 'Regular Matching and Inclusion on Compressed Tree Patterns with Constrained Context Variables'. In: *Information and Computation* 286 (2022). DOI: 10.1016/j.ic.2021.104776. URL: https://hal.inria.fr/hal-03151014.

[13] A. Bonifati, S. Dumbrava, G. Fletcher, J. Hidders, M. Hofer, W. Martens, F. Murlak, J. Shinavier, S. Staworko and D. Tomaszuk. 'Threshold Queries in Theory and in the Wild'. In: *Proceedings of the VLDB Endowment (PVLDB)* (2022). DOI: 10.14778/3510397.3510407. URL: https://hal.inria.fr/hal-03516360.

[14] M. Cadilhac and C. Paperman. 'The regular languages of wire linear AC0'. In: *Acta Informatica* 59.4 (Aug. 2022), pp. 321–336. DOI: 10.1007/s00236-022-00432-2. URL: https://hal.science/hal-03941070.

[15] K. Gebhardt, F. Meunier and S. Salvati. 'On is an n-MCFL'. In: *Journal of Computer and System Sciences* 127 (1st Aug. 2022), pp. 41–52. DOI: 10.1016/j.jcss.2022.02.003. URL: https://hal.science/hal-01771670.

[16] J. S. Guez, F. Coucheney, J. Castéra-Guy, M. Béchet, P. Fontanille, N.-E. Chihib, J. Niehren, F. Coutte and P. Jacques. 'Bioinformatics modelling and metabolic engineering of thebranched chain amino acid pathway for specific production of mycosubtilin isoforms in Bacillus subtilis'. In: *Metabolites* 12.2 (19th Jan. 2022). DOI: 10.3390/metabo12020107. URL: https://hal.inria.fr/hal-03498125.

[17] J. Lucero and S. Staworko. 'A note on the class of languages generated by F-systems over regular languages'. In: *Information Processing Letters* 179 (Jan. 2023), p. 106283. DOI: 10.1016/j.ipl.2022.106283. URL: https://hal.science/hal-03696314.

International peer-reviewed conferences

[18] A. Al Serhali and J. Niehren. 'A Benchmark Collection of Deterministic Automata for XPath Queries'. In: XML Prague 2022. Prague, Czech Republic, 9th June 2022. URL: https://hal.inria.fr/hal-03527888.

[19] A. Amarilli and M. Monet. 'Weighted Counting of Matchings in Unbounded-Treewidth Graph Families'. In: *Proceedings of MFCS*. MFCS. Vienna, Austria, 22nd Aug. 2022. DOI: 10.4230/LIPIcs.MFCS.2022.9. URL: https://hal.telecom-paris.fr/hal-03712197.

[20] C. Barloy, M. Cadilhac, C. Paperman and T. Zeume. 'The Regular Languages of First-Order Logic with One Alternation'. In: LICS 2022 - 37th Annual ACM/IEEE Symposium on Logic in Computer Science. Haïfa, Israel, 2nd Aug. 2022, pp. 1–11. DOI: 10.1145/3531130.3533371. URL: https://hal.science/hal-03934389.

[21] F. Capelli, N. Crosetti, J. Niehren and J. Ramon. 'Linear Programs with Conjunctive Queries'. In: 25th Internationcal Conference on Database Theory (ICDT 2022). Edinburgh, United Kingdom, 29th Mar. 2022. URL: https://hal.science/hal-01981553.

[22] D. Deutch, N. Frost, B. Kimelfeld and M. Monet. 'Computing the Shapley Value of Facts in Query Answering'. In: SIGMOD Conference 2022. Philadelphia, United States, 12th June 2022. URL: https://hal.inria.fr/hal-03514297.

[23] B. Groz, A. Lemay, S. Staworko and P. Wieczorek. 'Inference of Shape Graphs for Graph Databases'. In: International Conference on Database Theory. Edinbourgh, United Kingdom, 2022. DOI: 10.4230/LIPIcs.ICDT.2022.7. URL: https://hal.inria.fr/hal-03559309.

[24] J. Niehren, M. Sakho and A. Al Serhali. 'Schema-Based Automata Determinization'. In: Gandalf 2022: 13th International Symposium on Games, Automata, Logics, and Formal Verification. Madrid, Spain: EPTCS, 21st Sept. 2022. URL: https://hal.inria.fr/hal-03536045.

[25] J. Niehren, S. Salvati and R. Azimov. 'Jumping Evaluation of Nested Regular Path Queries'. In: 38th International Conference on Logic Programming (ICLP'2022). Haifa, Israel, 27th July 2022. URL: https://hal.inria.fr/hal-02492780.

[26] J. Niehren, A. Vaginay and C. Versari. 'Abstract Simulation of Reaction Networks via Boolean Networks'. In: CMSB2022: 20th International Conference on Computational Methods in Systems Biology. Bucarest, Romania: Springer, 14th Sept. 2022. URL: https://hal.science/hal-02279942.

Conferences without proceedings

[27] I. Boneva, B. Groz, J. Hidders, F. Murlak and S. Staworko. 'Static Analysis of Graph Database Transformations'. In: Symposium on Principles of Database Systems. Seattle, United States, 18th June 2023. URL: https://hal.science/hal-03937274.

[28] F. Capelli and Y. Strozecki. 'Geometric Amortization of Enumeration Algorithms'. In: 40th International Symposium on Theoretical Aspects of Computer Science (STACS 2023). Hamburg, Germany, 7th Mar. 2023. URL: https://hal.science/hal-03955911.

Reports & preprints

[29] A. Amarilli and M. Monet. *Enumerating Regular Languages with Bounded Delay*. Sept. 2022. URL: https://hal.science/hal-03940590.

[30] C. Paperman, S. Salvati and C. Soyez-Martin. *An algebraic approach to vectorial programs*. 27th Oct. 2022. DOI: 10.4230/LIPIcs.STACS.2023.14. URL: https://hal.science/hal-03831752.

12.3 Other

Softwares

[31] [SW] C. Paperman, S. Salvati and C. Soyez-Martin, *Addition Lemma*, 23rd Sept. 2022. LIC: MIT License. HAL: ⟨hal-03787033⟩, URL: https://hal.science/hal-03787033, VCS: https://gitlab.inria.fr/ssalvati/addition-lemma, SWHID: ⟨swh:1:dir:96d9d877d1614a94ce56bc8cec4c941adfaa1af3⟩.