RESEARCH CENTRE

Paris

IN PARTNERSHIP WITH:

CNRS, Ecole normale supérieure de Paris

2020 ACTIVITY REPORT

Project-Team VALDA

Value from Data

IN COLLABORATION WITH: Département d'Informatique de l'Ecole Normale Supérieure

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Contents

Project-Team VALDA

Creation of the Team: 2016 December 01, updated into Project-Team: 2018 January 01

Keywords

Computer sciences and digital sciences

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

A3.1.10. – Heterogeneous data

A3.1.11. – Structured data

A3.2. – Knowledge

A3.2.1. – Knowledge bases

A3.2.2. – Knowledge extraction, cleaning

A3.2.3. – Inference

A3.2.4. – Semantic Web

A3.2.5. – Ontologies

A3.2.6. – Linked data

A3.3.2. – Data mining

A3.4.3. – Reinforcement learning

A3.4.5. – Bayesian methods

A3.5.1. – Analysis of large graphs

A4.7. – Access control

A7.2. – Logic in Computer Science

A7.3. – Calculability and computability

A9.1. – Knowledge

A9.8. – Reasoning

Other research topics and application domains

B6.3.1. – Web

B6.3.4. – Social Networks

B6.5. – Information systems

B9.5.6. – Data science

B9.6.5. – Sociology

B9.6.10. – Digital humanities

B9.7.2. – Open data

B9.9. – Ethics

B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientists

- Serge Abiteboul [Inria, Emeritus, HDR]
- Camille Bourgaux [CNRS, Researcher]
- Olivier Cappé [CNRS, Senior Researcher]
- Luc Segoufin [Inria, Senior Researcher]
- Michaël Thomazo [Inria, Researcher]
- Victor Vianu [Inria, Advanced Research Position, from Sep 2020 until Oct 2020]

Faculty Members

- Pierre Senellart [Team leader, École Normale Supérieure de Paris, Professor]
- Leonid Libkin [École Normale Supérieure de Paris, Professor]
- Silviu Maniu [Université Paris-Saclay, Associate Professor, until Aug 2020]

Post-Doctoral Fellows

- Ashish Deepak Dandekar [École Normale Supérieure de Paris, until Sep 2020]
- Nathan Grosshans [École Normale Supérieure de Paris, until Aug 2020]
- Liat Peterfreund [CNRS]

PhD Students

- Juliette Achddou [1000 Mercis, CIFRE]
- Anatole Dahan [Université de Paris, from October 2020]
- Julien Grange [Université Denis Diderot, until Aug 2020]
- Yann Ramusat [École Normale Supérieure de Paris]
- Yoan Russac [École Normale Supérieure de Paris]

Interns and Apprentices

- Raphaël Chekroun [ENS, Intern, from Oct 2020]
- Théo Delmazure [École Normale Supérieure de Paris, Intern, from Apr 2020 until Aug 2020]
- Lucas Pluvinage [École Normale Supérieure de Paris, Intern, from Apr 2020 until Sep 2020]

Administrative Assistant

- Meriem Guemair [Inria]

External Collaborator

- Victor Vianu [Université de Californie, from Oct 2020]

2 Overall objectives

2.1 Objectives

Valda's focus is on both *foundational and systems aspects of* complex *data management*, especially *human-centric data*. The data we are interested in is typically heterogeneous, massively distributed, rapidly evolving, intensional, and often subjective, possibly erroneous, imprecise, incomplete. In this setting, Valda is in particular concerned with the optimization of complex resources such as computer time and space, communication, monetary, and privacy budgets. The goal is to extract *value from data*, beyond simple query answering.

Data management [44, 53] is now an old, well-established field, for which many scientific results and techniques have been accumulated since the sixties. Originally, most works dealt with static, homogeneous, and precise data. Later, works were devoted to heterogeneous data [42] [45], and possibly distributed [75] but at a small scale.

However, these classical techniques are poorly adapted to handle the new challenges of data management. Consider human-centric data, which is either produced by humans, e.g., emails, chats, recommendations, or produced by systems when dealing with humans, e.g., geolocation, business transactions, results of data analysis. When dealing with such data, and to accomplish any task to extract value from such data, we rapidly encounter the following facets:

- *Heterogeneity*: data may come in many different structures such as unstructured text, graphs, data streams, complex aggregates, etc., using many different schemas or ontologies.
- *Massive distribution*: data may come from a large number of autonomous sources distributed over the web, with complex access patterns.
- *Rapid evolution*: many sources may be producing data in real time, even if little of it is perhaps relevant to the specific application. Typically, recent data is of particular interest and changes have to be monitored.
- *Intensionality* [1]: in a classical database, all the data is available. In modern applications, the data is more and more available only intensionally, possibly at some cost, with the difficulty to discover which source can contribute towards a particular goal, and this with some uncertainty.
- *Confidentiality and security*: some personal data is critical and need to remain confidential. Applications manipulating personal data must take this into account and must be secure against linking.
- *Uncertainty*: modern data, and in particular human-centric data, typically includes errors, contradictions, imprecision, incompleteness, which complicates reasoning. Furthermore, the subjective nature of the data, with opinions, sentiments, or biases, also makes reasoning harder since one has, for instance, to consider different agents with distinct, possibly contradicting knowledge.

These problems have already been studied individually and have led to techniques such as *query rewriting* [66] or *distributed query optimization* [71].

Among all these aspects, intensionality is perhaps the one that has least been studied, so we pay particular attention to it. Consider a user's query, taken in a very broad sense: it may be a classical database query, some information retrieval search, a clustering or classification task, or some more advanced knowledge extraction request. Because of intensionality of data, solving such a query is a typically dynamic task: each time new data is obtained, the partial knowledge a system has of the world is revised, and query plans need to be updated, as in adaptive query processing [59] or aggregated search [83]. The system then needs to decide, based on this partial knowledge, of the best next access to perform. This is reminiscent of the central problem of reinforcement learning [81] (train an agent to accomplish a task in a partially known world based on rewards obtained) and of active learning [77] (decide which

[1] We use the spelling *intensional*, as in mathematical logic and philosophy, to describe something that is neither available nor defined in *extension*; *intensional* is derived from *intension*, while *intentional* is derived from *intent*.

action to perform next in order to optimize a learning strategy) and we intend to explore this connection further.

Uncertainty of the data interacts with its intensionality: efforts are required to obtain more precise, more complete, sounder results, which yields a trade-off between *processing cost* and *data quality*.

Other aspects, such as heterogeneity and massive distribution, are of major importance as well. A standard data management task, such as query answering, information retrieval, or clustering, may become much more challenging when taking into account the fact that data is not available in a central location, or in a common format. We aim to take these aspects into account, to be able to apply our research to real-world applications.

2.2 The Issues

We intend to tackle hard technical issues such as query answering, data integration, data monitoring, verification of data-centric systems, truth finding, knowledge extraction, data analytics, that take a different flavor in this modern context. In particular, we are interested in designing strategies to *minimize data access cost towards a specific goal, possibly a massive data analysis task*. That cost may be in terms of communication (accessing data in distributed systems, on the Web), of computational resources (when data is produced by complex tools such as information extraction, machine learning systems, or complex query processing), of monetary budget (paid-for application programming interfaces, crowdsourcing platforms), or of a privacy budget (as in the standard framework of differential privacy).

A number of data management tasks in Valda are inherently intractable. In addition to properly characterizing this intractability in terms of complexity theory, we intend to develop solutions for solving these tasks in practice, based on approximation strategies, randomized algorithms, enumeration algorithms with constant delay, or identification of restricted forms of data instances lowering the complexity of the task.

3 Research program

3.1 Scientific Foundations

We now detail some of the scientific foundations of our research on complex data management. This is the occasion to review connections between data management, especially on complex data as is the focus of Valda, with related research areas.

Complexity & Logic Data management has been connected to logic since the advent of the relational model as main representation system for real-world data, and of first-order logic as the logical core of database querying languages [44]. Since these early developments, logic has also been successfully used to capture a large variety of query modes, such as data aggregation [70], recursive queries (Datalog), or querying of XML databases [53]. Logical formalisms facilitate reasoning about the expressiveness of a query language or about its complexity.

The main problem of interest in data management is that of query evaluation, i.e., computing the results of a query over a database. The complexity of this problem has far-reaching consequences. For example, it is because first-order logic is in the AC_0 complexity class that evaluation of SQL queries can be parallelized efficiently. It is usual [82] in data management to distinguish *data complexity*, where the query is considered to be fixed, from *combined complexity*, where both the query and the data are considered to be part of the input. Thus, though conjunctive queries, corresponding to a simple SELECT-FROM-WHERE fragment of SQL, have PTIME data complexity, they are NP-hard in combined complexity. Making this distinction is important, because data is often far larger (up to the order of terabytes) than queries (rarely more than a few hundred bytes). Beyond simple query evaluation, a central question in data management remains that of complexity; tools from algorithm analysis, and complexity theory can be used to pinpoint the tractability frontier of data management tasks.

Automata Theory Automata theory and formal languages arise as important components of the study of many data management tasks: in temporal databases [43], queries, expressed in temporal logics,

can often by compiled to automata; in graph databases [49], queries are naturally given as automata; typical query and schema languages for XML databases such as XPath and XML Schema can be compiled to tree automata [74], or for more complex languages to data tree automata[39]. Another reason of the importance of automata theory, and tree automata in particular, comes from Courcelle's results [57] that show that very expressive queries (from the language of monadic second-order language) can be evaluated as tree automata over *tree decompositions* of the original databases, yielding linear-time algorithms (in data complexity) for a wide variety of applications.

Verification Complex data management also has connections to verification and static analysis. Besides query evaluation, a central problem in data management is that of deciding whether two queries are *equivalent* [44]. This is critical for query optimization, in order to determine if the rewriting of a query, maybe cheaper to evaluate, will return the same result as the original query. Equivalence can easily be seen to be an instance of the problem of (non-)satisfiability: $q \equiv q'$ if and only if $(q \wedge \neg q') \vee (\neg q \wedge q')$ is not satisfiable. In other words, some aspects of query optimization are static analysis issues. Verification is also a critical part of any database application where it is important to ensure that some property will never (or always) arise [55].

Workflows The orchestration of distributed activities (under the responsibility of a conductor) and their choreography (when they are fully autonomous) are complex issues that are essential for a wide range of data management applications including notably, e-commerce systems, business processes, health-care and scientific workflows. The difficulty is to guarantee consistency or more generally, quality of service, and to statically verify critical properties of the system. Different approaches to workflow specifications exist: automata-based, logic-based, or predicate-based control of function calls [41].

Probability & Provenance To deal with the uncertainty attached to data, proper models need to be used (such as attaching *provenance* information to data items and viewing the whole database as being *probabilistic*) and practical methods and systems need to be developed to both reliably estimate the uncertainty in data items and properly manage provenance and uncertainty information throughout a long, complex system.

The simplest model of data uncertainty is the NULLs of SQL databases, also called Codd tables [44]. This representation system is too basic for any complex task, and has the major inconvenient of not being closed under even simple queries or updates. A solution to this has been proposed in the form of *conditional tables* [68] where every tuple is annotated with a Boolean formula over independent Boolean random events. This model has been recognized as foundational and extended in two different directions: to more expressive models of *provenance* than what Boolean functions capture, through a semiring formalism [64], and to a probabilistic formalism by assigning independent probabilities to the Boolean events [65]. These two extensions form the basis of modern provenance and probability management, subsuming in a large way previous works [56, 50]. Research in the past ten years has focused on a better understanding of the tractability of query answering with provenance and probabilistic annotations, in a variety of specializations of this framework [80] [69, 47].

Machine Learning Statistical machine learning, and its applications to data mining and data analytics, is a major foundation of data management research. A large variety of research areas in complex data management, such as wrapper induction [76], crowdsourcing [48], focused crawling [63], or automatic database tuning [51] critically rely on machine learning techniques, such as classification [67], probabilistic models [62], or reinforcement learning [81].

Machine learning is also a rich source of complex data management problems: thus, the probabilities produced by a conditional random field [72] system result in probabilistic annotations that need to be properly modeled, stored, and queried.

Finally, complex data management also brings new twists to some classical machine learning problems. Consider for instance the area of *active learning* [77], a subfield of machine learning concerned with how to optimally use a (costly) oracle, in an interactive manner, to label training data that will be used to build a learning model, e.g., a classifier. In most of the active learning literature, the cost model is very basic (uniform or fixed-value costs), though some works [78] consider more realistic costs. Also,

oracles are usually assumed to be perfect with only a few exceptions [60]. These assumptions usually break when applied to complex data management problems on real-world data, such as crowdsourcing.

3.2 Research Directions

At the beginning of the Valda team, the project was to focus on the following directions:

- foundational aspects of data management, in particular related to query enumeration and reasoning on data, especially regarding security issues;
- implementation of provenance and uncertainty management, real-world applications, other aspects of uncertainty and incompleteness, in particular dynamic;
- development of personal information management systems, integration of machine learning techniques.

We believe the first two directions have been followed in a satisfactory manner. The focus on personal information management has not been kept for various organizational reasons, however, but the third axis of the project is reoriented to more general aspects of Web data management.

New permanent arrivals in the group since its creation have impacted its research directions in the following manner:

- Camille BOURGAUX and Michaël THOMAZO are both specialists of knowledge representation and formal aspects of knowledge bases, which is an expertise that did not exist in the group. They are also both interested in, and have started working on aspects related to connecting their research with database theory, and investigating aspects of uncertainty and incompleteness in their research. This will lead to more work on knowledge representation and symbolic AI aspects, while keeping the focus of Valda on foundations of data management and uncertainty.
- Olivier CAPPÉ is a specialist in statistics and machine learning, in particular multi-armed bandits and reinforcement learning. He is also interested in applications of these learning techniques to data management problems. His arrival in the group therefore complements the expertise of other researchers, and will lead to more work on machine learning issues.
- Leonid LIBKIN is a specialist of database theory, of incomplete data management, and has a line of current research on graph data management. His profile fits very well with the original orientation of the Valda project.

We intend to keep producing leading research on the foundations of data management. Generally speaking, the goal is to investigate the borders of feasibility of various tasks. For instance, what are the assumptions on data that allow for computable problems? When is it not possible at all? When can we hope for efficient query answering, when is it hopeless? This is a problem of theoretical nature which is necessary for understanding the limit of the methods and driving research towards the scenarios where positive results may be obtainable. Only when we have understood the limitation of different methods and have many examples where this is possible, we can hope to design a solid foundation that allowing for a good trade-off between what can be done (needs from the users) and what can be achieved (limitation from the system).

Similarly, we will continue our work, both foundational and practical, on various aspects of provenance and uncertainty management. One overall long-term goal is to reach a full understanding of the interactions between query evaluation or other broader data management tasks and uncertain and annotated data models. We would in particular want to go towards a full classification of tractable (typically polynomial-time) and intractable (typically NP-hard for decision problems, or #P-hard for probability evaluation) tasks, extending and connecting the query-based dichotomy [58] on probabilistic query evaluation with the instance-based one of [46, 47]. Another long-term goal is to consider more dynamic scenarios than what has been considered so far in the uncertain data management literature: when following a workflow, or when interacting with intensional data sources, how to properly represent and update uncertainty annotations that are associated with data. This is critical for many complex data management scenarios where one has to maintain a probabilistic current knowledge of the world, while

obtaining new knowledge by posing queries and accessing data sources. Such intensional tasks requires minimizing jointly data uncertainty and cost to data access.

As application area, in addition to the historical focus on personal information management which is now less stressed, we target Web data (Web pages, the semantic Web, social networks, the deep Web, crowdsourcing platforms, etc.).

We aim at keeping a delicate balance between theoretical, foundational research, and systems research, including development and implementation. This is a difficult balance to find, especially since most Valda researchers have a tendency to favor theoretical work, but we believe it is also one of the strengths of the team.

4 Application domains

4.1 Personal Information Management Systems

We recall that Valda's focus is on human-centric data, i.e., data produced by humans, explicitly or implicitly, or more generally containing information about humans. Quite naturally, we have used as a privileged application area to validate Valda's results that of personal information management systems (Pims for short) [40].

A Pims is a system that allows a user to integrate her own data, e.g., emails and other kinds of messages, calendar, contacts, web search, social network, travel information, work projects, etc. Such information is commonly spread across different services. The goal is to give back to a user the control on her information, allowing her to formulate queries such as "What kind of interaction did I have recently with Alice B.?", "Where were my last ten business trips, and who helped me plan them?". The system has to orchestrate queries to the various services (which means knowing the existence of these services, and how to interact with them), integrate information from them (which means having data models for this information and its representation in the services), e.g., align a GPS location of the user to a business address or place mentioned in an email, or an event in a calendar to some event in a Web search. This information must be accessed intensionally: for instance, costly information extraction tools should only be run on emails which seem relevant, perhaps identified by a less costly cursory analysis (this means, in turn, obtaining a cost model for access to the different services). Impacted people can be found by examining events in the user's calendar and determining who is likely to attend them, perhaps based on email exchanges or former events' participant lists. Of course, uncertainty has to be maintained along the entire process, and provenance information is needed to explain query results to the user (e.g., indicate which meetings and trips are relevant to each person of the output). Knowledge about services, their data models, their costs, need either to be provided by the system designer, or to be automatically learned from interaction with these services, as in [76].

One motivation for that choice is that Pims concentrate many of the problems we intend to investigate: heterogeneity (various sources, each with a different structure), massive distribution (information spread out over the Web, in numerous sources), rapid evolution (new data regularly added), intensionality (knowledge from Wikidata, OpenStreetMap...), confidentiality and security (mostly private data), and uncertainty (very variable quality). Though the data is distributed, its size is relatively modest; other applications may be considered for works focusing on processing data at large scale, which is a potential research direction within Valda, though not our main focus. Another strong motivation for the choice of Pims as application domain is the importance of this application from a societal viewpoint.

A Pims is essentially a system built on top of a user's *personal knowledge base*; such knowledge bases are reminiscent of those found in the Semantic Web, e.g., linked open data. Some issues, such as ontology alignment [79] exist in both scenarios. However, there are some fundamental differences in building personal knowledge bases vs collecting information from the Semantic Web: first, the scope is quite smaller, as one is only interested in knowledge related to a given individual; second, a small proportion of the data is already present in the form of semantic information, most needs to be extracted and annotated through appropriate wrappers and enrichers; third, though the linked open data is meant to be read-only, the only update possible to a user being adding new triples, a personal knowledge base is very much something that a user needs to be able to edit, and propagating updates from the knowledge base to original data sources is a challenge in itself.

4.2 Web Data

The choice of Pims is not exclusive. We also consider other application areas as well. In particular, we have worked in the past and have a strong expertise on Web data [45] in a broad sense: semi-structured, structured, or unstructured content extracted from Web databases [76]; knowledge bases from the Semantic Web [79]; social networks [73]; Web archives and Web crawls [61]; Web applications and deep Web databases [54]; crowdsourcing platforms [48]. We intend to continue using Web data as a natural application domain for the research within Valda when relevant. For instance [52], deep Web databases are a natural application scenario for intensional data management issues: determining if a deep Web database contains some information requires optimizing the number of costly requests to that database.

A common aspect of both personal information and Web data is that their exploitation raises ethical considerations. Thus, a user needs to remain fully in control of the usage that is made of her personal information; a search engine or recommender system that ranks Web content for display to a specific user needs to do so in an unbiased, justifiable, manner. These ethical constraints sometimes forbid some technically solutions that may be technically useful, such as sharing a model learned from the personal data of a user to another user, or using blackboxes to rank query result. We fully intend to consider these ethical considerations within Valda. One of the main goals of a Pims is indeed to empower the user with a full control on the use of this data.

5 Highlights of the year

5.1 Awards

- Leonid Libkin was awarded the *Gems of PODS award* at PODS 2020 for his work on incomplete data, and invited to write a survey article for the occasion. [21]
- Pierre Senellart was named a junior member of Institut Universitaire de France.

6 New software and platforms

6.1 New software

6.1.1 ProvSQL

Keywords: Databases, Provenance, Probability

Functional Description: The goal of the ProvSQL project is to add support for (m-)semiring provenance and uncertainty management to PostgreSQL databases, in the form of a PostgreSQL extension/module/plugin.

News of the Year: Implementation of an in-memory storage of the provenance circuit. Implementation of aggregate provenance. Major performance enhancements. Support for PostgreSQL 13. Miscellaneous enhancements and bug fixes.

URL: https://github.com/PierreSenellart/provsql

Publications: hal-01672566, hal-01851538

Contact: Pierre Senellart

Participants: Pierre Senellart, Silviu Maniu, Yann Ramusat

6.1.2 apxproof

Keyword: LaTeX

Functional Description: apxproof is a LaTeX package facilitating the typesetting of research articles with proofs in appendix, a common practice in database theory and theoretical computer science in general. The appendix material is written in the LaTeX code along with the main text which it naturally complements, and it is automatically deferred. The package can automatically send proofs to the appendix, can repeat in the appendix the theorem environments stated in the main text, can section the appendix automatically based on the sectioning of the main text, and supports a separate bibliography for the appendix material.

Release Contributions: Compatibility fixes with xypic, fancyvrb, memoir, natbib

News of the Year: Minor 1.2.1 release: compatibility fixes with xypic, fancyvrb, memoir, natbib

URL: https://github.com/PierreSenellart/apxproof

Contact: Pierre Senellart

Participant: Pierre Senellart

6.1.3 TheoremKB

Keyword: Information extraction

Functional Description: TheoremKB is a collection of tools to extract semantic information from (mathematical) research articles.

News of the Year: Initial version. Initial version of extractors of theorems and proofs from PDF and LaTeX. Construction and analysis of relations between theorems.

URL: https://github.com/PierreSenellart/theoremkb

Publications: hal-02956526, hal-02940819

Contact: Pierre Senellart

Participants: Pierre Senellart, Theo Delemazure, Lucas Pluvinage

7 New results

We present the results we obtained and published in 2020 in four directions: the management of incomplete and uncertain data; the complexity of query languages over restricted structures; information extraction; and some other works in the area of database theory.

7.1 Incompleteness, Uncertainty, and Provenance of Data

A major research area within Valda is the management of incomplete, missing, imprecise, uncertain data, along with the tracking of *data provenance*, a tool often necessary for uncertain data management.

Incomplete data. The standard notion of query answering over incomplete database is that of *certain answers*, guaranteeing correctness regardless of how incomplete data is interpreted. In [22], we consider databases with numerical data and queries with arithmetic and comparisons. Even though the notion of certain answers still applies,we explain that it becomes much more problematic in situations when missing data occurs in numerical columns. We propose a new general framework that allows us to assign a measure of certainty to query answers. We test it in the agnostic scenario where we do not have prior information about values of numerical attributes, similarly to the predominant approach in handling incomplete data which assumes that each null can be interpreted as an arbitrary value of the domain.

In [29], we consider incomplete databases whose information content may be enriched by additional knowledge. The knowledge order among them is derived from their semantics, rather than being fixed a priori. The resulting framework allows us to capture and justify existing notions of certainty, and extend

these concepts to other data models and query languages. As natural applications, we provide for the first time a well-founded definition of certain answers for the relational bag data model and for value-inventing queries on incomplete databases, addressing the key shortcomings of previous approaches.

Missing data. When dealing with missing data, one regularly estimates likelihoods of certain events by computing volumes of sets that serve as a mathematical representation of such events. Such sets need to be measurable, which is usually achieved by putting bounds, sometimes ad hoc, on them. In [23], we address the question how unbounded or unmeasurable sets can be measured nonetheless. Intuitively, we want to know how likely a randomly chosen point is to be in a given set, even in the absence of a uniform distribution over the entire space.

Inconsistent data. Another form of uncertainty in databases arises in the presence of *inconsistencies*; a way to address such inconsistencies is by way of *optimal repairs*. In [17], we explore the issue of inconsistency handling over prioritized knowledge bases (KBs), which consist of an ontology, a set of facts, and a priority relation between con- flicting facts. After transfer- ring the notions of globally-, Pareto- and completion-optimal repairs from the database literature to our setting, we study the data complexity of the core reasoning tasks: query entailment under inconsistency-tolerant semantics based upon optimal repairs, existence of a unique optimal repair, and enumeration of all optimal repairs. Our results provide a nearly complete picture of the data complexity of these tasks for ontologies formulated in common DL-Lite dialects.

Data provenance. In [20, 19], we address the problem of handling provenance information in ELHr ontologies. We consider a setting recently introduced for ontology-based data access, based on semirings and extending classical data provenance, in which ontology axioms are annotated with provenance tokens. A consequence inherits the provenance of the axioms involved in deriving it, yielding a provenance polynomial as annotation. We analyse the semantics for the ELHr case and show that the presence of conjunctions poses various difficulties for handling provenance, some of which are mitigated by assuming multiplicative idempotency of the semiring. Under this assumption, we study three problems: ontology completion with provenance, computing the set of relevant axioms for a consequence, and query answering.

In [12], we investigate compact representations of *Boolean provenance* represented as Boolean circuits, by providing a systematic picture of many circuit classes considered in knowledge compilation and how they can be systematically connected to width measures, through upper and lower bounds. Our upper bounds show that bounded-treewidth circuits can be constructively converted to d-SDNNFs, in time linear in the circuit size and singly exponential in the treewidth; and that bounded-pathwidth circuits can similarly be converted to uOBDDs. We show matching lower bounds on the compilation of monotone DNF or CNF formulas to structured targets, assuming a constant bound on the arity (size of clauses) and degree (number of occurrences of each variable): any d-SDNNF (resp., SDNNF) for such a DNF (resp., CNF) must be of exponential size in its treewidth, and the same holds for uOBDDs (resp., n-OBDDs) when considering pathwidth.

7.2 Query Languages over Restricted Structures

Another major line of research within Valda is to investigate the complexity of classical database problems (query evaluation, query enumeration, query containement), and the expressive power of quer classes, when data is assumed to have a restricted structure: trees, relations with bounded treewidth, bounded expansion, bounded degree...

Complexity. In [15], we consider the evaluation of first-order queries over classes of databases with bounded expansion. The notion of bounded expansion is fairly broad and generalizes bounded degree, bounded treewidth and exclusion of at least one minor. It was known that over a class of databases with bounded expansion, first-order sentences could be evaluated in time linear in the size of the database. We give a different proof of this result. Moreover, we show that answers to first-order queries can be

enumerated with constant delay after a linear time preprocessing. We also show that counting the number of answers to a query can be done in time linear in the size of the database.

In [14], we consider the problem of containment of monadic datalog (MDL) queries in unions of conjunctive queries (UCQs). We start by revisiting the connection between MDL/UCQ containment and containment problems involving regular tree languages. We then present a general approach for getting tighter bounds on the complexity of query containment, based on analysis of the number of mappings of queries into tree-like instances. We give two applications of the machinery. We first give an important special case of the MDL/UCQ containment problem that is in EXPTIME, and use this bound to show an EXPTIME bound on containment under access patterns. Secondly we show that the same technique can be used to get a new tight upper bound for containment of tree automata in UCQs. We finally show that the new MDL/UCQ upper bounds are tight. We establish a 2EXPTIME lower bound on the MDL/UCQ containment problem, resolving an open problem from the early 1990s.

Expresive power. Julien Grange's PhD thesis [32] focused on the expressive power of invariant logics over sparse classes of structures. In [26], we show that the expressive power of order-invariant first-order logic collapses to first-order logic over hollow trees. A hollow tree is an unranked ordered tree where every non leaf node has at most four adjacent nodes: two siblings (left and right) and its first and last children. In particular there is no predicate for the linear order among siblings nor for the descendant relation. Moreover only the first and last nodes of a siblinghood are linked to their parent node, and the parent-child relation cannot be completely reconstructed in first-order. In [25], we study the expressive power of successor-invariant first-order logic, which is an extension of first-order logic where the usage of an additional successor relation on the structure is allowed, as long as the validity of formulas is independent on the choice of a particular successor. We show that when the degree is bounded, successor-invariant first-order logic is no more expressive than first-order logic.

7.3 Information Extraction

Information extraction consists in extracting structured data and knowledge from unstructured text or semi-structured documents.

The framework of document spanners abstracts the task of information extraction from text as a function that maps every document (a string) into a relation over the document's spans (intervals identified by their start and end indices). In [24], we embark on the investigation of document spanners that can annotate extractions with auxiliary information such as confidence, support, and confidentiality measures. To this end, we adopt the abstraction of provenance semirings. Hence, the proposed spanner extension, referred to as an annotator, maps every string into an annotated relation over the spans. We investigate key aspects of expressiveness, such as the closure under the positive RA, and key aspects of computational complexity, such as the enumeration of annotated answers and their ranked enumeration in the case of numeric semirings.

Beyond these formal approaches, we also consider a practical application of information extraction: building a knowledge base of (mathematical) results in the scientific literature. This is the goal of the TheoremKB project. In [38], we start on the task of extracting theorems and proofs from the PDF version or LaTeX source of mathematical articles. In [37], we aim at building a graph of interconnected results from these extracted theorems and proofs.

One standard approach to information extraction has been to rely on human intelligence by using the power of *crowd data sourcing* platforms. In [13], we discuss challenges of such platforms involving *humans in the loop*, and more generally how such platforms can evolve to support the *future of work*.

7.4 Other Topics in Database Theory

Finally, we also dealt with other research problems, mostly in the field of database theory.

Register automata have been used as a convenient model for specifying and verifying database driven systems. An important problem in such systems is to provide views that hide or restructure certain information about the data or process, extending classical notions of database views. In [28] we carry out a formal investigation of views of register automata by considering simple views that project away some of the registers. We show that classical register automata are not able to describe such projections and

introduce more powerful register automata that are able to do so. We also show useful properties of these automata such as closure under projection and decidability of verifying temporal properties of their runs.

Ontology-mediated query answering (OMQA) is a promising approach to data access and integration that has been actively studied in the knowledge representation and database communities for more than a decade. The vast majority of work on OMQA focuses on conjunctive queries, whereas more expressive queries that feature counting or other forms of aggregation remain largely unex-plored. In [18], we introduce a general form of counting query, relate it to previous proposals, and study the complexity of answering such queries in the presence of DL-Lite ontologies. As it follows from existing work that query answering is intractable and often of high complexity, we consider some practically relevant restrictions, for which we establish improved complexity bounds.

The *program-over-monoid* model of computation originates with Barrington's proof that it captures the complexity class NC^1. In [27] we make progress in understanding the subtleties of the model. First, we identify a new tameness condition on a class of monoids that entails a natural characterization of the regular languages recognizable by programs over monoids from the class. Second, we prove that the class known as DA satisfies tameness and hence that the regular languages recognized by programs over monoids in DA are precisely those recognizable in the classical sense by morphisms from QDA. Third, we show by contrast that the well studied class of monoids called J is not tame. Finally, we exhibit a program-length-based hierarchy within the class of languages recognized by programs over monoids from DA.

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

Numberly:

- Duration: 2019–2022
- Local coordinator: Olivier Cappé
- Juliette Achddou's PhD research is set up as a CIFRE contract and supervision agreement between her employer, the Numberly company, and École normale supérieure.

Neo4j:

- Duration: 2020–2021
- Local coordinator: Leonid Libkin
- A contract has been established with Neo4j, the leading company in the field of graph databases, to work towards the creation of a new standard for graph languages called GQL, building on Neo4j's Cypher query language. Leonid Libkin is chairing a working group on the formal semantics of GQL. In addition to Valda, it involves researchers from Edinburgh, Santiago, Warsaw, and other universities in Paris (UPEM, Université de Paris). This project is supported by a grant from Neo4j. Leonid Libkin is also a scientific advisor of Neo4j.

8.2 Standardization activities

Leonid Libkin is involved in the standardization process of the GQL and SQL query languages. In particular, he is a chair of the LDBC working group on semantics of GQL, and a member of ISO/IEC JTC1 SC32 WG3 (SQL committee).

9 Partnerships and cooperations

9.1 International initiatives

Informal international partners Valda has strong collaborations with the following international groups:

Univ. Edinburgh, United Kingdom: Paolo Guagliardo, Andreas Pieris

Univ. Oxford, United Kingdom: Michael Benedikt, Dan Olteanu, and Georg Gottlob

TU Dresden, Germany: Markus Krötzsch and Sebastian Rudolph

Dortmund University, Germany: Thomas Schwentick

Bayreuth University, Germany: Wim Martens

Univ. Bergen, Norway: Ana Ozaki

Univ. Roma La Sapienza, Italy: Marco Console

Warsaw University, Poland: Mikołaj Bojańczyk and Szymon Toruńczyk

Tel Aviv University, Israel: Daniel Deutch and Tova Milo

NYU, USA: Julia Stoyanovich

Univ. California San Diego, USA: Victor Vianu

Pontifical Catholic University of Chile: Marcelo Arenas, Pablo Barceló

National University of Singapore: Stéphane Bressan

9.2 International research visitors

9.2.1 Visits of international scientists

Victor Vianu, Professor at UC San Diego and former holder of an Inria international chair, spent a few months within Valda, employed on a short-term Advanced Research Position on the HeadWork project.

9.3 European initiatives

9.3.1 Collaborations in European programs, except FP7 and H2020

A bilateral French–German ANR project, entitled EQUUS – Efficient Query answering Under UpdateS has started in 2020. It involves CNRS (CRIL, CRIStAL, IMJ), Télécom Paris, HU Berlin, and Bayreuth University, in addition to Inria Valda.

9.4 National initiatives

9.4.1 ANR

Valda has been part of four national ANR projects in 2020:

HEADWORK (2016–2021; 38 k€ for Valda, budget managed by Inria), together with IRISA (Druid, coordinator), Inria Lille (Links & Spirals), and Inria Rennes (Sumo), and two application partners: MNHN (Cesco) and FouleFactory. The topic is workflows for crowdsourcing. See http://headwork.gforge.inria.fr/.

BioQOP (2017–2021; 66 k€ for Valda, budget managed by ENS), with Idemia (coordinator) and GREYC, on the optimization of queries for privacy-aware biometric data management. See http://bioqop.di.ens.fr/.

CQFD (2018–2022; 19 k€ for Valda, budget managed by Inria), with Inria Sophia (GraphIK, coordinator), LaBRI, LIG, Inria Saclay (Cedar), IRISA, Inria Lille (Spirals), and Télécom ParisTech, on complex ontological queries over federated and heterogeneous data. See http://www.lirmm.fr/cqfd/.

QUID (2018–2022; 49 k€ for Valda, budget managed by Inria), LIGM (coordinator), IRIF, and LaBRI, on incomplete and inconsistent data. See https://quid.labri.fr/home.html.

Camille Bourgaux has been participating in the AI Chair of Meghyn Bienvenu on *INTENDED (Intelligent handling of imperfect data)* since 2020.

9.5 Regional initiatives

Liat Peterfreund obtained a post-doc scholarship from FSMP from 2020 to 2022.

Pierre Senellart has held a Chair of the PaRis Artificial Intelligence Research InstitutE, PRAIRIE since the fall of 2019.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- Leonid Libkin, general chair of PODS 2021 and chair of the PODS Executive Committee
- Luc Segoufin, chair of the steering committee of the conference series Highlights of Logic, Games and Automata
- Pierre Senellart, co-organizer and chief judge of the ICPC (International Collegiate Programming Contest) Southwestern Europe 2019-2020 competition

Member of the organizing committees

- Leonid Libkin, member of the SIGMOD Executive Committee.
- Pierre Senellart, member of the steering committee of BDA, the French scientific community on data management.
- Pierre Senellart, co-organizer and secretary of the ICPC (International Collegiate Programming Contest) Southwestern Europe 2020-2021 competition.

10.1.2 Scientific events: selection

Chair of conference program committees

- Leonid Libkin, LICS 2021

Member of the conference program committees

- Camille Bourgaux, AAAI 2021, DL 2020, IJCAI 2020, KR 2020, TIME 2020
- Leonid Libkin, FOSSACS 2020, IJCAI 2020, KR 2020 (track chair)
- Olivier Cappé, NeurIPS 2020 (area chair)
- Liat Peterfreund, PODS 2021
- Luc Segoufin, ICALP 2020
- Pierre Senellart, BDA 2020, ICDT 2021 Test-of-Time Committee, PODS 2021
- Michaël Thomazo, AAAI 2021, IJCAI 2020, RJCIA 2020

10.1.3 Journal

Member of the editorial boards

- Olivier Cappé, *Annals of the Institute of Statistical Mathematics*
- Leonid Libkin, *Bulletin of Symbolic Logic*
- Leonid Libkin, *Acta Informatica*
- Leonid Libkin, *RAIRO Theoretical Informatics and Applications*
- Leonid Libkin, *Journal of Applied Logic*
- Leonid Libkin, *SN Computer Science*

10.1.4 Leadership within the scientific community

- Serge Abiteboul is a member of the French Academy of Sciences, of the Academia Europaea, of the scientific council of the Société Informatique de France, and an ACM Fellow.
- Leonid Libkin is a Fellow of the Royal Society of Edinburgh, a member of the Academia Europaea, of the UK Computing research committee, and an ACM Fellow.
- Pierre Senellart is a junior member of the Institut Universitaire de France.

10.1.5 Scientific expertise

- Pierre Senellart, reviews for ANR (*Flash Covid-19*), FONDECYT (Chili)

10.1.6 Research administration

- Olivier Cappé is a scientific deputy director of CNRS division of Information Sciences and Technologies (INS2I).
- Luc Segoufin is a member of the CNHSCT of Inria.
- Pierre Senellart is a member of the board of section 6 of the National Committee for Scientific Research.
- Pierre Senellart is deputy director of the DI ENS laboratory, joint between ENS, CNRS, and Inria.
- Pierre Senellart is a member of the board of the DIM RFSI (Réseau Francilien en Sciences Informatiques).

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Licence: *Databases*, 74 heqTD, L3, École normale supérieure – Pierre Senellart, Nathan Grosshans, Leonid Libkin, Michaël Thomazo
- Licence: *Data Structures*, NYU Paris – Ashish Dandekar
- Master: *Data wrangling, Data privacy*, 36 heqTD, M2, IASD – Leonid Libkin, Pierre Senellart
- Master: *Anonymization, privacy*, 36 heqTD, M2, IASD – Ashish Dandekar, Pierre Senellart
- Master: *Knowledge graphs, description logics, reasoning on data*, 72 heqTD, M2, IASD – Camille Bourgaux, Michaël Thomazo

- Other: invited one-day mini course on graph data at Peking University (Beijing, online) – Leonid Libkin

Pierre Senellart has had various teaching responsibilities (L3 internships, M1 projects, M2 administration, entrance competition) at ENS. Leonid Libkin is responsible of the graduate program in computer science of PSL University, and co-responsible of the international entrance competition at ENS. Nathan Grosshans was the secretary of the entrance competition at ENS for computer science. Most members of the group are also involved in tutoring ENS students, advising them on their curriculum, their internships, etc. They are also occasionally involved with reviewing internship reports, supervising student projects, etc.

10.2.2 Supervision

- PhD: Julien Grange, *Successor-Invariant First-Order Logic on Classes of Bounded Degree*, PSL University, 29 June 2020, Luc Segoufin
- PhD in progess: Juliette Achddou, *Application of reinforcement learning strategies to the context of Real-Time Bidding*, started in September 2018, Olivier Cappé & Aurélien Garivier
- PhD in progress: Anatole Dahan, *Logical foundations of the polynomial hierarchy*, started in October 2020, Arnaud Durand & Luc Segoufin
- PhD in progress: Yann Ramusat, *Provenance-based routing in probabilistic graphs*, started in September 2018, Silviu Maniu & Pierre Senellart
- PhD in progess: Yoan Russac, *Sequential methods for robust decision making*, started in December 2018, Olivier Cappé

10.2.3 Juries

- PhD: Julien Romero [president], Institut Polytechnique de Paris, Pierre Senellart

10.3 Popularization

10.3.1 Internal or external Inria responsibilities

- Serge Abiteboul is the president of the strategic committee of the Blaise Pascal foundation for scientific mediation.
- Pierre Senellart is a research fellow within the CERRE (Centre on Regulation in Europe), a European think tank that produces policy papers and organize events about the regulation of network industries. He contributes in particular to reflections on the use of artificial intelligence techniques and on the interoperability of software platforms.

10.3.2 Articles and contents

- Serge Abiteboul is a founding editor of the binaire blog for popularizing computer science. See https://www.lemonde.fr/blog/binaire/.
- Serge Abiteboul co-edited a special issue of a magazine on the industrial heritage of information technology [16] in which he also co-wrote an article on *Pictures of the digital transformation* [11].
- Olivier Cappé co-wrote two research reports on population mobility in France during the Covid-19 pandemic [34, 33].
- Pierre Senellart co-wrote a CERRE report on *Making data portability more effective for the digital economy* [35].

11 Scientific production

11.1 Major publications

[1] S. Abiteboul, P. Bourhis and V. Vianu. 'Explanations and Transparency in Collaborative Workflows'. In: *PODS 2018 - 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles Of Database Systems*. Houston, Texas, United States, June 2018. URL: https://hal.inria.fr/hal-01744978.

[2] M. Benedikt, P. Bourhis, G. Gottlob and P. Senellart. 'Monadic Datalog, Tree Validity, and Limited Access Containment'. In: *ACM Transactions on Computational Logic* 21.1 (2020), 6:1–6:45. DOI: 10.1145/3344514. URL: https://hal.inria.fr/hal-02307999.

[3] M. Bienvenu, Q. Manière and M. Thomazo. 'Answering Counting Queries over DL-Lite Ontologies'. In: *IJCAI 2020 - Twenty-Ninth International Joint Conference on Artificial Intelligence*. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. Reportée de juillet 2020 à janvier 2021 en raison de la COVID. Yokohama, Japan, July 2020. URL: https://hal.inria.fr/hal-02927913.

[4] C. Bourgaux, A. Ozaki, R. Peñaloza and L. Predoiu. 'Provenance for the Description Logic ELHr'. In: *IJCAI-PRICAI-20 - Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence*. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. Reportée de juillet 2020 à janvier 2021 en raison de la COVID. Yokohama, Japan, July 2020, pp. 1862–1869. DOI: 10.24963/ijcai.2020/258. URL: https://hal.archives-ouvertes.fr/hal-02899464.

[5] M. Console, P. Guagliardo, L. Libkin and E. Toussaint. 'Coping with Incomplete Data: Recent Advances'. In: *SIGMOD/PODS 2020 - International Conference on Management of Data*. Portland / Virtual, United States: ACM, June 2020, pp. 33–47. DOI: 10.1145/3375395.3387970. URL: https://hal.inria.fr/hal-03127726.

[6] W. Kazana and L. Segoufin. 'First-order queries on classes of structures with bounded expansion'. In: *Logical Methods in Computer Science* 16.1 (2020). URL: https://hal.inria.fr/hal-01706665.

[7] P. Lagrée, O. Cappé, B. Cautis and S. Maniu. 'Algorithms for Online Influencer Marketing'. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13.1 (Jan. 2019), pp. 1–30. DOI: 10.1145/3274670. URL: https://hal.inria.fr/hal-01478788.

[8] Y. Russac, C. Vernade and O. Cappé. 'Weighted Linear Bandits for Non-Stationary Environments'. In: *NeurIPS 2019 - 33rd Conference on Neural Information Processing Systems*. https://arxiv.org/abs/1909.09146. Vancouver, Canada, Dec. 2019. URL: https://hal.inria.fr/hal-02291460.

[9] N. Schweikardt, L. Segoufin and A. Vigny. 'Enumeration for FO Queries over Nowhere Dense Graphs'. In: *PODS 2018 - Principles Of Database Systems*. Houston, United States, June 2018. URL: https://hal.inria.fr/hal-01895786.

[10] P. Senellart, L. Jachiet, S. Maniu and Y. Ramusat. 'ProvSQL: Provenance and Probability Management in PostgreSQL'. In: *Proceedings of the VLDB Endowment (PVLDB)* 11.12 (Aug. 2018), pp. 2034–2037. DOI: 10.14778/3229863.3236253. URL: https://hal.inria.fr/hal-01851538.

11.2 Publications of the year

International journals

[11] S. Abiteboul and C. Mathieu. 'Pictures of the digital transformation'. In: *Patrimoine industriel* (2020). URL: https://hal.inria.fr/hal-02613482.

[12] A. Amarilli, F. Capelli, M. Monet and P. Senellart. 'Connecting Knowledge Compilation Classes and Width Parameters'. In: *Theory of Computing Systems* (1st Aug. 2020). DOI: 10.1007/s00224-019-09930-2. URL: https://hal.inria.fr/hal-02163749.

[13] S. Basu Roy, L. Chen, A. Morishima, J. A. Monedero, P. Bourhis, F. Charoy, M. Danilevsky, G. Das, G. Demartini, A. Dubey, S. Elbassuoni, D. Gross-Amblard, E. Hoareau, M. Inoguchi, J. Kenworthy, I. Kitahara, D. Lee, Y. Li, R. M. Borromeo, P. Papotti, R. Rao, S. Roy, P. Senellart, K. Tajima, S. Thirumuruganathan, M. Tommasi, K. Umemoto, A. Wiggins, K. Yoshida and S. Amer-Yahia. 'Making AI Machines Work for Humans in FoW'. In: *SIGMOD record* 49.2 (9th Dec. 2020), pp. 30–35. DOI: 10.1145/3442322.3442327. URL: https://hal.inria.fr/hal-03103700.

[14] M. Benedikt, P. Bourhis, G. Gottlob and P. Senellart. 'Monadic Datalog, Tree Validity, and Limited Access Containment'. In: *ACM Transactions on Computational Logic* 21.1 (2020), 6:1–6:45. DOI: 10.1145/3344514. URL: https://hal.inria.fr/hal-02307999.

[15] W. Kazana and L. Segoufin. 'First-order queries on classes of structures with bounded expansion'. In: *Logical Methods in Computer Science* 16.1 (2020). URL: https://hal.inria.fr/hal-01706665.

National journals

[16] S. Abiteboul and F. Hachez-Leroy. 'What Heritage for Information Technology ? Introduction to the journal'. In: *Patrimoine industriel*. Patrimoine industriel informatique 73 (2020). URL: https://hal.inria.fr/hal-02613465.

International peer-reviewed conferences

[17] M. Bienvenu and C. Bourgaux. 'Querying and Repairing Inconsistent Prioritized Knowledge Bases: Complexity Analysis and Links with Abstract Argumentation'. In: KR 2020 - 17th International Conference on Principles of Knowledge Representation and Reasoning. Rhodes / Virtual, Greece, 2020, pp. 141–151. DOI: 10.24963/kr.2020/15. URL: https://hal.inria.fr/hal-02947251.

[18] M. Bienvenu, Q. Manière and M. Thomazo. 'Answering Counting Queries over DL-Lite Ontologies'. In: IJCAI 2020 - Twenty-Ninth International Joint Conference on Artificial Intelligence. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. Yokohama, Japan, 11th July 2020. URL: https://hal.inria.fr/hal-02927913.

[19] C. Bourgaux, A. Ozaki, R. Peñaloza and L. Predoiu. 'Provenance for the Description Logic ELHr'. In: IJCAI-PRICAI-20 - Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. Yokohama, Japan, 11th July 2020, pp. 1862–1869. DOI: 10.24963/ijcai.2020/258. URL: https://hal.archives-ouvertes.fr/hal-02899464.

[20] C. Bourgaux, A. Ozaki, R. Peñaloza and L. Predoiu. 'Provenance for the Description Logic ELHr (Extended Abstract)'. In: DL 2020 - 33rd International Workshop on Description Logics. Rhodes / Virtual, Greece, 12th Sept. 2020. URL: https://hal.inria.fr/hal-02947265.

[21] M. Console, P. Guagliardo, L. Libkin and E. Toussaint. 'Coping with Incomplete Data: Recent Advances'. In: SIGMOD/PODS 2020 - International Conference on Management of Data. Portland / Virtual, United States, 29th May 2020, pp. 33–47. DOI: 10.1145/3375395.3387970. URL: https://hal.inria.fr/hal-03127726.

[22] M. Console, M. Hofer and L. Libkin. 'Queries with Arithmetic on Incomplete Databases'. In: SIGMOD/PODS 2020 : International Conference on Management of Data. Portland / Virtual, United States, 14th June 2020, pp. 179–189. DOI: 10.1145/3375395.3387666. URL: https://hal.inria.fr/hal-03127717.

[23] M. Console, M. Hofer and L. Libkin. 'Reasoning about Measures of Unmeasurable Sets'. In: KR 2020 - 17th International Conference on Principles of Knowledge Representation and Reasoning. Rhodes / Virtual, Greece, 12th Sept. 2020, pp. 264–273. DOI: 10.24963/kr.2020/27. URL: https://hal.inria.fr/hal-03128512.

[24] J. Doleschal, B. Kimelfeld, W. Martens and L. Peterfreund. 'Weight Annotation in Information Extraction'. In: ICDT 2020 - 23rd International Conference on Database Theory. Copenhague / Virtual, Denmark: https://diku-dk.github.io/edbticdt2020/, 30th Mar. 2020. DOI: 10.4230/LIPIcs.ICDT.2020.8. URL: https://hal.inria.fr/hal-03104155.

[25] J. Grange. 'Successor-Invariant First-Order Logic on Classes of Bounded Degree'. In: LICS 2020 - Thirty-Fifth Annual ACM/IEEE Symposium on Logic in Computer Science. Vol. 13. Saarbrücken / Virtual, Germany, 8th July 2020. DOI: 10.1145/3373718.3394767. URL: https://hal.inria.fr/hal-02882118.

[26] J. Grange and L. Segoufin. 'Order-Invariant First-Order Logic over Hollow Trees'. In: CSL 2020 - 28th annual conference of the European Association for Computer Science Logic. Vol. 23. Barcelona, Spain, 13th Jan. 2020, pp. 1–23. DOI: 10.4230/LIPIcs.CSL.2020.23. URL: https://hal.inria.fr/hal-02310749.

[27] N. Grosshans. 'The Power of Programs over Monoids in J'. In: LATA 2020 - 14th International Conference on Language and Automata Theory and Applications. Milan, Italy, 4th Mar. 2020. DOI: 10.1007/978-3-030-40608-0_22. URL: https://hal.archives-ouvertes.fr/hal-02414771.

[28] L. Segoufin and V. Vianu. 'Projection Views of Register Automata'. In: PODS'20: Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. Portland / Virtual, United States, 14th June 2020, pp. 299–313. DOI: 10.1145/3375395.3387651. URL: https://hal.inria.fr/hal-02947172.

[29] E. Toussaint, P. Guagliardo and L. Libkin. 'Knowledge-Preserving Certain Answers for SQL-like Queries'. In: KR 2020 - 17th International Conference on Principles of Knowledge Representation and Reasoning. Rhodes / Virtual, Greece, 12th Sept. 2020, pp. 758–767. DOI: 10.24963/kr.2020/78. URL: https://hal.inria.fr/hal-03128504.

National peer-reviewed Conferences

[30] A. Dandekar, D. Basu, P. Senellart and S. Bressan. 'Confidentialité différentielle à risque : Relier les sources d'aléa et un budget de confidentialité'. In: BDA 2020 - 36ème Conférence sur la Gestion de Données – Principes, Technologies et Applications. Paris / Virtuel, France, 27th Oct. 2020. URL: https://hal.inria.fr/hal-03103528.

[31] Y. Ramusat, S. Maniu and P. Senellart. 'Algorithmes à base de provenance pour des requêtes enrichies sur les bases de données graphes'. In: BDA 2020 - 36ème Conférence sur la Gestion de Données – Principes, Technologies et Applications. Paris / Virtuel, France, 27th Oct. 2020. URL: https://hal.inria.fr/hal-03103509.

Doctoral dissertations and habilitation theses

[32] J. Grange. 'On the Expressive Power of Invariant Logics over Sparse Classes of Structures'. ENS Paris, 29th June 2020. URL: https://hal.inria.fr/tel-02947853.

Reports & preprints

[33] J. Atif, B. Cabot, O. Cappé, O. Mula and R. Pinot. *Initiative face au virus. Regards croisés sur l'épidemie de Covid-19 apportés par les données sanitaires et de géolocalisation (mars à octobre 2020)*. Université PSL; Inria; CNRS, 21st Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03084832.

[34] J. Atif, O. Cappé, A. Kazakçi, Y. Léo, L. Massoulié and O. Mula. *Initiative face au virus Observations sur la mobilité pendant l'épidémie de Covid-19*. Université PSL, 25th May 2020. URL: https://hal.archives-ouvertes.fr/hal-02921194.

[35] J. Krämer, P. Senellart and A. de Streel. *Making data portability more effective for the digital economy*. CERRE, 15th June 2020. URL: https://hal.inria.fr/hal-03151335.

[36] Y. Russac, O. Cappé and A. Garivier. *Algorithms for Non-Stationary Generalized Linear Bandits*. 21st Mar. 2020. URL: https://hal.archives-ouvertes.fr/hal-02514151.

Other scientific publications

[37] T. Delemazure. 'A Knowledge Base of Mathematical Results'. Ecole Normale Supérieure (ENS), 21st Sept. 2020. URL: https://hal.inria.fr/hal-02940819.

[38] L. Pluvinage. 'Extracting scientific results from research articles'. Ecole Normale Supérieure (ENS), 21st Sept. 2020. URL: https://hal.inria.fr/hal-02956526.

11.3 Cited publications

[39] F. Jacquemard, L. Segoufin and J. Dimino. 'FO2(<, +1, ~) on data trees, data tree automata and branching vector addition systems'. In: *Logical Methods in Computer Science* 12.2 (2016). DOI: 10.2168/LMCS-12(2:3)2016. URL: https://doi.org/10.2168/LMCS-12(2:3)2016.

[40] S. Abiteboul, B. André and D. Kaplan. 'Managing your digital life'. In: *Commun. ACM* 58.5 (2015), pp. 32–35. DOI: 10.1145/2670528. URL: http://doi.acm.org/10.1145/2670528.

[41] S. Abiteboul, P. Bourhis and V. Vianu. 'Comparing workflow specification languages: A matter of views'. In: *ACM Trans. Database Syst.* 37.2 (2012), 10:1–10:59. DOI: 10.1145/2188349.2188352. URL: http://doi.acm.org/10.1145/2188349.2188352.

[42] S. Abiteboul, P. Buneman and D. Suciu. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann, 1999.

[43] S. Abiteboul, L. Herr and J. Van den Bussche. 'Temporal Versus First-Order Logic to Query Temporal Databases'. In: *Proceedings of the Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, 1996, Montreal, Canada*. 1996, pp. 49–57. DOI: 10.1145/237661.237674. URL: http://doi.acm.org/10.1145/237661.237674.

[44] S. Abiteboul, R. Hull and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995. URL: http://webdam.inria.fr/Alice/.

[45] S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset and P. Senellart. *Web Data Management*. Cambridge University Press, 2011. URL: http://webdam.inria.fr/Jorge.

[46] A. Amarilli, P. Bourhis and P. Senellart. 'Provenance Circuits for Trees and Treelike Instances'. In: *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part II*. 2015, pp. 56–68. DOI: 10.1007/978-3-662-47666-6_5. URL: https://doi.org/10.1007/978-3-662-47666-6_5.

[47] A. Amarilli, P. Bourhis and P. Senellart. 'Tractable Lineages on Treelike Instances: Limits and Extensions'. In: *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016, San Francisco, CA, USA, June 26 - July 01, 2016*. 2016, pp. 355–370. DOI: 10.1145/2902251.2902301. URL: http://doi.acm.org/10.1145/2902251.2902301.

[48] Y. Amsterdamer, Y. Grossman, T. Milo and P. Senellart. 'CrowdMiner: Mining association rules from the crowd'. In: *PVLDB* 6.12 (2013), pp. 1250–1253. URL: http://www.vldb.org/pvldb/vol6/p1250-amsterdamer.pdf.

[49] P. B. Baeza. 'Querying graph databases'. In: *Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2013, New York, NY, USA - June 22 - 27, 2013*. 2013, pp. 175–188. DOI: 10.1145/2463664.2465216. URL: http://doi.acm.org/10.1145/2463664.2465216.

[50] D. Barbará, H. Garcia-Molina and D. Porter. 'The Management of Probabilistic Data'. In: *IEEE Trans. Knowl. Data Eng.* 4.5 (1992), pp. 487–502. DOI: 10.1109/69.166990. URL: https://doi.org/10.1109/69.166990.

[51] D. Basu, Q. Lin, W. Chen, H. T. Vo, Z. Yuan, P. Senellart and S. Bressan. 'Regularized Cost-Model Oblivious Database Tuning with Reinforcement Learning'. In: *T. Large-Scale Data- and Knowledge-Centered Systems* 28 (2016), pp. 96–132. DOI: 10.1007/978-3-662-53455-7_5. URL: https://doi.org/10.1007/978-3-662-53455-7_5.

[52] M. Benedikt, G. Gottlob and P. Senellart. 'Determining relevance of accesses at runtime'. In: *Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2011, June 12-16, 2011, Athens, Greece*. 2011, pp. 211–222. DOI: 10.1145/1989284.1989309. URL: http://doi.acm.org/10.1145/1989284.1989309.

[53] M. Benedikt and P. Senellart. 'Databases'. In: *Computer Science, The Hardware, Software and Heart of It*. Springer, 2011, pp. 169–229. DOI: 10.1007/978-1-4614-1168-0_10. URL: https://doi.org/10.1007/978-1-4614-1168-0_10.

[54] M. Bienvenu, D. Deutch, D. Martinenghi, P. Senellart and F. M. Suchanek. 'Dealing with the Deep Web and all its Quirks'. In: *Proceedings of the Second International Workshop on Searching and Integrating New Web Data Sources, Istanbul, Turkey, August 31, 2012*. 2012, pp. 21–24. URL: http://ceur-ws.org/Vol-884/VLDS2012_p21_Bienvenu.pdf.

[55] M. Bojańczyk, L. Segoufin and S. Toruńczyk. 'Verification of database-driven systems via amalgamation'. In: *Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2013, New York, NY, USA - June 22 - 27, 2013*. 2013, pp. 63–74. DOI: 10.1145/2463664.2465228. URL: http://doi.acm.org/10.1145/2463664.2465228.

[56] P. Buneman, S. Khanna and W.-C. Tan. 'Why and Where: A Characterization of Data Provenance'. In: *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings*. 2001, pp. 316–330. DOI: 10.1007/3-540-44503-X_20. URL: https://doi.org/10.1007/3-540-44503-X_20.

[57] B. Courcelle. 'The Monadic Second-Order Logic of Graphs. I. Recognizable Sets of Finite Graphs'. In: *Inf. Comput.* 85.1 (1990), pp. 12–75. DOI: 10.1016/0890-5401(90)90043-H. URL: https://doi.org/10.1016/0890-5401(90)90043-H.

[58] N. N. Dalvi and D. Suciu. 'The dichotomy of probabilistic inference for unions of conjunctive queries'. In: *J. ACM* 59.6 (2012), 30:1–30:87. DOI: 10.1145/2395116.2395119. URL: http://doi.acm.org/10.1145/2395116.2395119.

[59] A. Deshpande, Z. G. Ives and V. Raman. 'Adaptive Query Processing'. In: *Foundations and Trends in Databases* 1.1 (2007), pp. 1–140. DOI: 10.1561/1900000001. URL: https://doi.org/10.1561/1900000001.

[60] P. Donmez and J. G. Carbonell. 'Proactive learning: cost-sensitive active learning with multiple imperfect oracles'. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*. 2008, pp. 619–628. DOI: 10.1145/1458082.1458165. URL: http://doi.acm.org/10.1145/1458082.1458165.

[61] M. Faheem and P. Senellart. 'Adaptive Web Crawling Through Structure-Based Link Classification'. In: *Digital Libraries: Providing Quality Information - 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, Seoul, Korea, December 9-12, 2015, Proceedings*. 2015, pp. 39–51. DOI: 10.1007/978-3-319-27974-9_5. URL: https://doi.org/10.1007/978-3-319-27974-9_5.

[62] L. Getoor. *Introduction to statistical relational learning*. MIT Press, 2007.

[63] G. Gouriten, S. Maniu and P. Senellart. 'Scalable, generic, and adaptive systems for focused crawling'. In: *25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014*. 2014, pp. 35–45. DOI: 10.1145/2631775.2631795. URL: http://doi.acm.org/10.1145/2631775.2631795.

[64] T. J. Green, G. Karvounarakis and V. Tannen. 'Provenance semirings'. In: *Proceedings of the Twenty-Sixth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 11-13, 2007, Beijing, China*. 2007, pp. 31–40. DOI: 10.1145/1265530.1265535. URL: http://doi.acm.org/10.1145/1265530.1265535.

[65] T. J. Green and V. Tannen. 'Models for Incomplete and Probabilistic Information'. In: *IEEE Data Eng. Bull.* 29.1 (2006), pp. 17–24. URL: http://sites.computer.org/debull/A06mar/green.ps.

[66] A. Y. Halevy. 'Answering queries using views: A survey'. In: *VLDB J.* 10.4 (2001), pp. 270–294. DOI: 10.1007/s007780100054. URL: https://doi.org/10.1007/s007780100054.

[67] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf. 'Support vector machines'. In: *IEEE Intelligent Systems* 13.4 (1998), pp. 18–28. DOI: 10.1109/5254.708428. URL: https://doi.org/10.1109/5254.708428.

[68] T. Imielinski and W. Lipski Jr. 'Incomplete Information in Relational Databases'. In: *J. ACM* 31.4 (1984), pp. 761–791. DOI: 10.1145/1634.1886. URL: http://doi.acm.org/10.1145/1634.1886.

[69] B. Kimelfeld and P. Senellart. 'Probabilistic XML: Models and Complexity'. In: *Advances in Probabilistic Databases for Uncertain Information Management*. Springer, 2013, pp. 39–66. DOI: 10.1007/978-3-642-37509-5_3. URL: https://doi.org/10.1007/978-3-642-37509-5_3.

[70] A. C. Klug. 'Equivalence of Relational Algebra and Relational Calculus Query Languages Having Aggregate Functions'. In: *J. ACM* 29.3 (1982), pp. 699–717. DOI: 10.1145/322326.322332. URL: http://doi.acm.org/10.1145/322326.322332.

[71] D. Kossmann. 'The State of the art in distributed query processing'. In: *ACM Comput. Surv.* 32.4 (2000), pp. 422–469. DOI: 10.1145/371578.371598. URL: http://doi.acm.org/10.1145/371578.371598.

[72] J. D. Lafferty, A. McCallum and F. C. N. Pereira. 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data'. In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*. 2001, pp. 282–289.

[73] S. Lei, S. Maniu, L. Mo, R. Cheng and P. Senellart. 'Online Influence Maximization'. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*. 2015, pp. 645–654. DOI: 10.1145/2783258.2783271. URL: http://doi.acm.org/10.1145/2783258.2783271.

[74] F. Neven. 'Automata Theory for XML Researchers'. In: *SIGMOD Record* 31.3 (2002), pp. 39–46. DOI: 10.1145/601858.601869. URL: http://doi.acm.org/10.1145/601858.601869.

[75] M. T. Özsu and P. Valduriez. *Principles of Distributed Database Systems, Third Edition*. Springer, 2011. DOI: 10.1007/978-1-4419-8834-8. URL: https://doi.org/10.1007/978-1-4419-8834-8.

[76] P. Senellart, A. Mittal, D. Muschick, R. Gilleron and M. Tommasi. 'Automatic wrapper induction from hidden-web sources with domain knowledge'. In: *10th ACM International Workshop on Web Information and Data Management (WIDM 2008), Napa Valley, California, USA, October 30, 2008*. 2008, pp. 9–16. DOI: 10.1145/1458502.1458505. URL: http://doi.acm.org/10.1145/1458502.1458505.

[77] B. Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012. DOI: 10.2200/S00429ED1V01Y201207AIM018. URL: https://doi.org/10.2200/S00429ED1V01Y201207AIM018.

[78] B. Settles, M. Craven and L. Friedland. 'Active learning with real annotation costs'. In: *NIPS 2008 Workshop on Cost-Sensitive Learning*. 2008. URL: http://burrsettles.com/pub/settles.nips08ws.pdf.

[79] F. M. Suchanek, S. Abiteboul and P. Senellart. 'PARIS: Probabilistic Alignment of Relations, Instances, and Schema'. In: *PVLDB* 5.3 (2011), pp. 157–168. URL: http://www.vldb.org/pvldb/vol5/p157_fabianmsuchanek_vldb2012.pdf.

[80] D. Suciu, D. Olteanu, C. Ré and C. Koch. *Probabilistic Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011. DOI: 10.2200/S00362ED1V01Y201105DTM016. URL: https://doi.org/10.2200/S00362ED1V01Y201105DTM016.

[81] R. S. Sutton and A. G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998. URL: http://www.worldcat.org/oclc/37293240.

[82] M. Y. Vardi. 'The Complexity of Relational Query Languages (Extended Abstract)'. In: *Proceedings of the 14th Annual ACM Symposium on Theory of Computing, May 5-7, 1982, San Francisco, California, USA*. 1982, pp. 137–146. DOI: 10.1145/800070.802186. URL: http://doi.acm.org/10.1145/800070.802186.

[83] K. Zhou, M. Lalmas, T. Sakai, R. Cummins and J. M. Jose. 'On the reliability and intuitiveness of aggregated search metrics'. In: *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*. 2013, pp. 689–698. DOI: 10.1145/2505515.2505691. URL: http://doi.acm.org/10.1145/2505515.2505691.