RESEARCH CENTRE

Inria Paris Center

IN PARTNERSHIP WITH:

Ecole normale supérieure de Paris, CNRS

2022 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

Inria

Contents

Project-Team VALDA

Creation of the 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]
- Luc Segoufin [Inria, Senior Researcher, HDR]
- Michael Thomazo [Inria, Researcher]

Faculty Members

- Pierre Senellart [Team leader, ENS Paris, Professor, HDR]
- Leonid Libkin [ENS Paris, Professor]
- Cristina Sirangelo [Université Paris-Cité, Professor, from Feb 2022 until Jul 2022, Secondment to Inria, HDR]
- Victor Vianu [ENS Paris, Professor, from Jul 2022 until Nov 2022, Visiting professor]

Post-Doctoral Fellows

- Nofar Carmeli [ENS Paris, until Sep 2022]
- Shufan Jiang [ENS Paris, from Dec 2022, ATER]
- Anantha Padmanabha [ENS Paris]

PhD Students

- Anatole Dahan [Université Paris-Cité]
- Baptiste Lafosse [ENS Paris]
- Shrey Mishra [ENS Paris]
- Yann Ramusat [ENS Paris, ATER, until Aug 2022]
- Alexandra Rogova [Université Paris-Cité]

Technical Staff

- N. Smith [ENS Paris, Engineer, from Feb 2022]

Interns and Apprentices

- Yacine Brihmouche [Université Paris-Dauphine, Intern, from May 2022 until Sep 2022]
- Antoine Gauquier [Télécom Paris, Intern, from Oct 2022, Part-time]
- Siméon Gheorgin [PSL, Intern, until Jun 2022, Part-time]

Administrative Assistant

- Meriem Guemair [Inria]

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 [41, 50] 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 [39] [42], and possibly distributed [72] 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* [63] or *distributed query optimization* [68].

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 [56] or aggregated search [80]. 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 [78] (train an agent to accomplish a task in a partially known world based on rewards obtained) and of active learning [74] (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 [41]. Since these early developments, logic has also been successfully used to capture a large variety of query modes, such as data aggregation [67], recursive queries (Datalog), or querying of XML databases [50]. 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 [79] 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 [40], queries, expressed in temporal logics,

can often by compiled to automata; in graph databases [46], 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 [71], or for more complex languages to data tree automata[36]. Another reason of the importance of automata theory, and tree automata in particular, comes from Courcelle's results [54] 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* [41]. 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 [52].

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 [38].

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 [41]. 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* [65] 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 [61], and to a probabilistic formalism by assigning independent probabilities to the Boolean events [62]. These two extensions form the basis of modern provenance and probability management, subsuming in a large way previous works [53, 47]. 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 [77] [66, 44].

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 [73], crowdsourcing [45], focused crawling [60], or automatic database tuning [48] critically rely on machine learning techniques, such as classification [64], probabilistic models [59], or reinforcement learning [78].

Machine learning is also a rich source of complex data management problems: thus, the probabilities produced by a conditional random field [69] 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* [74], 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 [75] consider more realistic costs. Also,

oracles are usually assumed to be perfect with only a few exceptions [57]. 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.
- 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 [55] on probabilistic query evaluation with the instance-based one of [43, 44]. 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) [37].

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 [73].

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 [76] 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 [42] in a broad sense: semi-structured, structured, or unstructured content extracted from Web databases [73]; knowledge bases from the

Semantic Web [76]; social networks [70]; Web archives and Web crawls [58]; Web applications and deep Web databases [51]; crowdsourcing platforms [45]. We intend to continue using Web data as a natural application domain for the research within Valda when relevant. For instance [49], 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 Social and environmental responsibility

Data-driven algorithmic systems raise ethical and legal concerns, that need to be taken into account within research. Serge Abiteboul, with collaborators from NYU, U. Washington, U. Michigan, U. Amsterdam, wrote a position article detailing the role that data management research needs to play in ensuring responsible design and use of algorithmic data-driven systems. [17]

6 Highlights of the year

6.1 Awards

Michaël Thomazo, together with Maxime Buron and Marie-Laure Mugnier, received the BDA (French database community) award for their work on *Parallelisable Existential Rules: a Story of Pieces* [31], also published at KR 2021 [31]

6.2 Broader Inria Context

The work of the Valda team in 2022 was affected by several issues within Inria; in particular major issues with the deployment of a new information system (Eksae) negatively impacted the work of our administrative assistant and made it impossible for the team leader to keep track of expenses.

The team also would like to thank the Inria evaluation committee for its admirable work in support of the research community, for its transparency, and for the integrity in which it conducts its activities.

7 New software and platforms

7.1 New software

7.1.1 ORBITS

Name: Optimal Repair-Based Inconsistency-Tolerant Semantics

Keywords: Knowledge Bases, Databases

Scientific Description: ORBITS (Optimal Repair-Based Inconsistency-Tolerant Semantics) is a tool for filtering answers that hold under a given inconsistency-tolerant semantics among AR, IAR and brave with standard repairs or Pareto- or completion-optimal repairs in the case where a priority relation between the conflicting facts is given. ORBITS implements a variety of algorithms and propositional encoding variants for each semantics and type of repairs.

Functional Description: ORBITS is a tool for filtering answers that hold under a given inconsistency-tolerant semantics based on some kind of optimal repairs in the case where a priority relation between the conflicting facts is given.

URL: https://github.com/bourgaux/orbits

Publication: hal-03770516

Contact: Camille Bourgaux

Participant: Camille Bourgaux

7.1.2 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: Support for PostgreSQL 15. Miscellaneous enhancements and bug fixes.

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

Publications: hal-01672566, hal-01851538

Contact: Pierre Senellart

Participants: Pierre Senellart, Baptiste Lafosse

7.1.3 TheoremKB

Keyword: Information extraction

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

News of the Year: Improvements to theorem extraction, preliminary work on multimodal approach.

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

Publications: hal-02956526, hal-02940819, hal-03293643, hal-03897168

Contact: Pierre Senellart

Participants: Pierre Senellart, Shrey Mishra, Yacine Brihmouche

7.1.4 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: Support for lipcs's claimproof, support optional arguments in proofs

News of the Year: 1.2.4 release: support for claimproof environment from lipics, support optional arguments in proofs.

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

Contact: Pierre Senellart

Participant: Pierre Senellart

7.1.5 dissem.in

Name: Dissemin

Keywords: Open Access, Publishing, HAL

Functional Description: Dissemin is a web platform gathering metadata from many sources to analyze the open-access full text availability of publications of researchers. It has been designed to foster the use of repositories such as HAL (rather than preprints posted on personal homepages). It allows deposit on these repositories.

News of the Year: Support for a large variety of IdP from Shibboleth. Various Shibboleth fixes. Support for v2 of Sherpa/Romeo API. Other small improvements, bug fixes, maintainance.

URL: https://gitlab.com/dissemin/dissemin

Contact: Pierre Senellart

Participant: Pierre Senellart

Partner: CAPSH

7.2 New platforms

7.2.1 dissem.in

dissem.in, the openly accessible platform for promoting full-text deposit of scientific articles of researchers, which is based on the dissem.in (7.1.5) software, has been maintained by Valda since 2021. Works on the platform in 2022, in addition to works on the base software, include updating information about journals and publisher policies from the Sherpa/Romeo API.

Participants: Pierre Senellart, N. Smith.

8 New results

We present the results we obtained and published in 2022. Much research within Valda centers around the central problem of query answering in databases, but exploring various side questions: How to handle incomplete or inconsistent information? How to efficiently access query results when there are many of them? How to incorporate external ontologies within query answering? How to keep track of the provenance of queries? We describe our works in each of these areas in turn, and finish with other theoretical research conducted in the team, beyond data management.

8.1 Incomplete and inconsistent information

We first consider databases containing incomplete (missing) or inconsistent (contradictory) information.

One of the most common scenarios of handling incomplete information occurs in relational databases. They describe incomplete knowledge with three truth values, using Kleene's logic for propositional formulae and a rather peculiar extension to predicate calculus. This design by a committee from several decades ago is now part of the standard adopted by vendors of database management systems. But is it really the right way to handle incompleteness in propositional and predicate logics? Our goal in [13] is to answer this question. Using an epistemic approach, we first characterize possible levels of partial knowledge about propositions, which leads to six truth values. We impose rationality conditions on the semantics of the connectives of the propositional logic, and prove that Kleene's logic is the maximal sublogic to which the standard optimization rules apply, thereby justifying this design choice. For extensions to predicate logic, however, we show that the additional truth values are not necessary: every many-valued extension of first-order logic over databases with incomplete information represented by

null values is no more powerful than the usual two-valued logic with the standard Boolean interpretation of the connectives. We use this observation to analyze the logic underlying SQL query evaluation, and conclude that the many-valued extension for handling incompleteness does not add any expressiveness to it.

We continue on the topic of incomplete information in [18], where our goal is to collect and analyze the shortcomings of nulls and their treatment by SQL, and to re-evaluate existing research in this light. To this end, we designed and conducted a survey on the everyday usage of null values among database users. From the analysis of the results we reached two main conclusions. First, null values are ubiquitous and relevant in real-life scenarios, but SQL's features designed to deal with them cause multiple problems. The severity of these problems varies depending on the SQL features used, and they cannot be reduced to a single issue. Second, foundational research on nulls is misdirected and has been addressing problems of limited practical relevance. We urge the community to view the results of this survey as a way to broaden the spectrum of their researches and further bridge the theory-practice gap on null values.

To answer database queries over incomplete data the gold standard is finding certain answers: those that are true regardless of how incomplete data is interpreted. Such answers can be found efficiently for conjunctive queries and their unions, even in the presence of constraints such as keys or functional dependencies. With negation added, the complexity of finding certain answers becomes intractable however. In [28] we exhibit a well-behaved class of queries that extends unions of conjunctive queries with a limited form of negation and that permits efficient computation of certain answers even in the presence of constraints by means of rewriting into Datalog with negation. The class consists of queries that are the closure of conjunctive queries under Boolean operations of union, intersection and difference. We show that for these queries, certain answers can be expressed in Datalog with negation, even in the presence of functional dependencies, thus making them tractable in data complexity. We show that in general Datalog cannot be replaced by first-order logic, but without constraints such a rewriting can be done in first-order.

While all relational database systems are based on the bag data model, much of theoretical research still views relations as sets. Recent attempts to provide theoretical foundations for modern data management problems under the bag semantics concentrated on applications that need to deal with incomplete relations, i.e., relations populated by constants and nulls. Our goal in [12] is to provide a complete characterization of the complexity of query answering over such relations in fragments of bag relational algebra. The main challenges that we face are twofold. First, bag relational algebra has more operations than its set analog (e.g., additive union, max-union, min-intersection, duplicate elimination) and the relationship between various fragments is not fully known. Thus we first fill this gap. Second, we look at query answering over incomplete data, which again is more complex than in the set case: rather than certainty and possibility of answers, we now have numerical information about occurrences of tuples. We then fully classify the complexity of finding this information in all the fragments of bag relational algebra.

Finally, we turn to inconsistent data. In [20, 19], we investigate practical algorithms for inconsistency-tolerant query answering over prioritized knowledge bases, which consist of a logical theory, a set of facts, and a priority relation between conflicting facts. We consider three well-known semantics (AR, IAR and brave) based upon two notions of optimal repairs (Pareto and completion). Deciding whether a query answer holds under these semantics is (co)NP-complete in data complexity for a large class of logical theories, and SAT-based procedures have been devised for repair-based semantics when there is no priority relation, or the relation has a special structure. We introduce the first SAT encodings for Pareto- and completion-optimal repairs w.r.t. general priority relations and proposes several ways of employing existing and new encodings to compute answers under (optimal) repair-based semantics, by exploiting different reasoning modes of SAT solvers. The comprehensive experimental evaluation of our implementation compares both (i) the impact of adopting semantics based on different kinds of repairs, and (ii) the relative performances of alternative procedures for the same semantics.

8.2 Enumeration and direct access to query results

Many queries have as output sets of results which are too big to be generated at once. Two strategies can then be used: either to design algorithms for efficient *enumeration* of the query results, one after the other, or for efficient *direct access* to one specific result among the set of results.

In [16], we consider the evaluation of first-order queries over classes of databases that are nowhere

dense. The notion of nowhere-dense classes was introduced by Nešetřil and Ossona de Mendez as a formalization of classes of "sparse" graphs and generalizes many well-known classes of graphs, such as classes of bounded degree, bounded tree-width, or bounded expansion. It has recently been shown by Grohe, Kreutzer, and Siebertz that over nowhere-dense classes of databases, first-order sentences can be evaluated in pseudo-linear time (pseudo-linear time means that for all ϵ there exists an algorithm working in time $O(n^{1+\epsilon})$, where n is the size of the database). For first-order queries of higher arities, we show that over any nowhere dense class of databases, the set of their solutions can be enumerated with constant delay after a pseudo-linear time preprocessing. In the same context, we also show that after a pseudo-linear time preprocessing we can, on input of a tuple, test in constant time whether it is a solution to the query.

A class of relational databases has low degree if for all $\delta > 0$, all but finitely many databases in the class have degree at most n^{δ} , where n is the size of the database. Typical examples are databases of bounded degree or of degree bounded by $\log n$. It is known that over a class of databases having low degree, first-order boolean queries can be checked in pseudo-linear time, i.e. for all $\epsilon > 0$ in time bounded by $n^{1+\epsilon}$. We generalize this result in [14] by considering query evaluation. We show that counting the number of answers to a query can be done in pseudo-linear time and that after a pseudo-linear time preprocessing we can test in constant time whether a given tuple is a solution to a query or enumerate the answers to a query with constant delay.

Finally, we consider in [25] the task of lexicographic direct access to query answers. That is, we want to simulate an array containing the answers of a join query sorted in a lexicographic order chosen by the user. A recent dichotomy showed for which queries and orders this task can be done in polylogarithmic access time after quasilinear preprocessing, but this dichotomy does not tell us how much time is required in the cases classified as hard. We determine the pre-processing time needed to achieve polylogarithmic access time for all self-join free queries and all lexicographical orders. To this end, we propose a decomposition-based general algorithm for direct access on join queries. We then explore its optimality by proving lower bounds for the preprocessing time based on the hardness of a certain online Set-Disjointness problem, which shows that our algorithm's bounds are tight for all lexicographic orders on self-join free queries. Then, we prove the hardness of Set-Disjointness based on the Zero-Clique Conjecture which is an established conjecture from fine-grained complexity theory. We also show that similar techniques can be used to prove that, for enumerating answers to Loomis-Whitney joins, it is not possible to significantly improve upon trivially computing all answers at preprocessing. This, in turn, gives further evidence (based on the Zero-Clique Conjecture) to the enumeration hardness of self-join free cyclic joins with re-spect to linear preprocessing and constant delay.

8.3 Ontology-mediated query answering

We know consider cases where to answer a query, we need to take into account external knowledge given in the form of a logical *ontology* (e.g., described in description logics, or through *existential rules*).

While ontology-mediated query answering most often adopts (unions of) conjunctive queries as the query language, some recent works have explored the use of counting queries coupled with DL-Lite ontologies. The aim of [22, 21] is to extend the study of counting queries to Horn description logics outside the DL-Lite family. Through a combination of novel techniques, adaptations of existing constructions, and new connections to closed predicates, we achieve a complete picture of the data and combined complexity of answering counting conjunctive queries (CCQs) and cardinality queries (a restricted class of CCQs) in $\mathcal{ELHI}_{\bot}$ and its various sublogics. Notably, we show that CCQ answering is 2EXP-complete in combined complexity for $\mathcal{ELHI}_{\bot}$ and every sublogic that extends EL or DL-Lite$^{\mathcal{H}}_{\text{pos}}$. Our study not only provides the first results for counting queries beyond DL-Lite, but it also closes some open questions about the combined complexity of CCQ answering in DL-Lite.

Existential rules are a very popular ontology-mediated query language for which the chase represents a generic computational approach for query answering. It is straightforward that existential rule queries exhibiting chase termination are decidable and can only recognize properties that are preserved under homomorphisms. [24] is an extended abstract of our eponymous publication at KR 2021 where we show the converse: every decidable query that is closed under homomorphism can be expressed by an existential rule set for which the standard chase universally terminates. Membership in this fragment is not decidable, but we show via a diagonalisation argument that this is unavoidable.

In the literature, existential rules are often supposed to be in some normal form that simplifies technical developments. For instance, a common assumption is that rule heads are atomic, i.e., restricted to a single atom. Such assumptions are considered to be made without loss of generality as long as all sets of rules can be normalised while preserving entailment. However, an important question is whether the properties that ensure the decidability of reasoning are preserved as well. We provide in [26] a systematic study of the impact of these procedures on the different chase variants with respect to chase (non-)termination and FO-rewritability. This also leads us to study open problems related to chase termination of independent interest.

8.4 Provenance for recursive queries

Data provenance consists in bookkeeping meta information during query evaluation, in order to enrich query results with their trust level, likelihood, evaluation cost, and more. The framework of semiring provenance abstracts from the specific kind of meta information that annotates the data.

While the definition of semiring provenance is uncontroversial for unions of conjunctive queries, the picture is less clear for Datalog. Indeed, the original definition might include infinite computations, and is not consistent with other proposals for Datalog semantics over annotated data. In [23], we propose and investigate several provenance semantics, based on different approaches for defining classical Datalog semantics. We study the relationship between these semantics, and introduce properties that allow us to analyze and compare them.

In [30, 33], we establish a translation between a formalism for dynamic programming over hypergraphs and the computation of semiring-based provenance for Datalog programs. The benefit of this translation is a new method for computing the provenance of Datalog programs for specific classes of semirings, which we apply to provenance-aware querying of graph databases. Theoretical results and practical optimizations lead to an efficient implementation using Soufflé, a state-of-the-art Datalog interpreter. Experimental results on real-world data suggest this approach to be efficient in practical contexts, competing with dedicated solutions for graphs.

8.5 Theoretical computer science beyond databases

Valda's research has always encompassed other foundational topics. We conclude with the description of other theoretical computer science works (namely, in algebraic automata theory and logic), which does not fit within the previous areas of research.

The program-over-monoid model of computation originates with Barrington's proof that the model captures the complexity class NC^1. In [15] 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.

When we bundle quantifiers and modalities together (as in $\exists x \Box$, $\Diamond \forall x$ etc.) in first-order modal logic (FOML), we get new logical operators whose combinations produce interesting bundled fragments of FOML. It is well-known that finding decidable fragments of FOML is hard, but existing work shows that certain bundled fragments are decidable, without any restriction on the arity of predicates, the number of variables, or the modal scope. In [29], we explore generalized bundles such as $\forall x \forall y \Box$, $\forall x \exists y \Diamond$ etc., and map the terrain with regard to decidability, presenting both decidability and undecidability results. In particular, we propose the loosely bundled fragment, which is decidable over increasing domains and encompasses all known decidable bundled fragments.

9 Bilateral contracts and grants with industry

9.1 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). He is also a member of INCITS, the US InterNational Committee for Information Technology Standards.

As part of this standardization effort, [27] presents the key elements of the graph pattern matching language at the core of both SQL/PGQ and GQL, in advance of the publication of the corresponding new standards.

Participants: Leonid Libkin.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

GQA

Title: Languages for Graph Querying and Analytics

Duration: 2022 ->

Coordinator: Pablo Barceló (pbarcelo@dcc.uchile.cl)

Partners:

- Pontificia Universidad Católica de Chile Santiago (Chili)

Inria contact: Leonid Libkin

Summary: The project brings together experts in graph databases, in particular in the new generation of query languages currently standardized by the ISO. The history of collaboration between the two groups goes back many years and pre-dates our current collaboration on graph data; having started in the areas of tree-structured data and data interoperability. Our main objective is to combine the graph query languages expertise of the Inria group with the machine learning and graphs analytics expertise of the Chilean group to come up with a new generation of query languages that seamlessly integrate graph querying with analytics.

10.1.2 Participation in other International Programs

DesCartes (2021–2026) is a project managed by CNRS@CREATE, a CNRS subsidiary in Singapore and funded by Singapore's National Research Foundation, with 50 million total budget. Pierre Senellart is involved in the project as one of the French PIs.

10.1.3 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 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

10.2 International research visitors

10.2.1 Visits of international scientists

Visits of international scientists

- Victor Vianu, Professor at UCSD, visited the group during several months in 2022. He was also hired on a fixed-term contract by ENS.
- Yael Amsterdamer, Senior Lecturer at Bar-Ilan University & Daniel Deutch, Professor at Tel Aviv University jointly visited Valda in July 2022.
- Dan Suciu, Professor at University of Washington, visited Valda in November 2022.

10.3 European initiatives

10.3.1 Other european programs/initiatives

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

10.4 National initiatives

10.4.1 ANR

Valda has been part of three national ANR projects in 2022:

CQFD (2018–2024; 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.

QUID (2018–2024; 49 k€ for Valda, budget managed by Inria), LIGM (coordinator), IRIF, and LaBRI, on incomplete and inconsistent data.

VERIGRAPH (2022–2026; 150 k€ for Valda (coordinator), budget managed by ENS), LIG, and LIRIS, on verifiable graph queries and transformations

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

Pierre Senellart has held a chair within the PR[AI]RIE institute for artificial intelligence in Paris since 2019.

10.4.2 Others

Dissemin (2021–2024; 124 k€ for Valda, budget managed by ENS), sole partner, on the development of the dissem.in platform for open science promotion. Funded by the Fonds National Science Ouverte.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- Leonid Libkin, general chair of PODS 2022; chair (until July 2022) and now member of the PODS Executive Committee

Member of the organizing committees

- Leonid Libkin, member of the LICS Steering Committee
- Luc Segoufin, member of the steering committee of the conference series *Highlights of Logic, Games and Automata*

11.1.2 Scientific events: selection

Chair of conference program committees

- Camille Bourgaux, program co-chair of the *Artificial Intelligence in Bergen* research school, AIB 2022

Member of the conference program committees

- Camille Bourgaux, AAAI 2023, IJCAI-ECAI 2022, KR 2022, DL 2022
- Nofar Carmeli, PODS 2023
- Leonid Libkin, KR 2022 (area chair), KR 2023, The Web Conference 2023 (industry track)
- Pierre Senellart, BDA 2022, SIGMOD 2023
- Michaël Thomazo, IJCAI 2022, KR 2022

11.1.3 Journal

Member of the editorial boards

- Leonid Libkin, *Acta Informatica*
- Leonid Libkin, *Bulletin of Symbolic Logic*
- Luc Segoufin, *ACM Transactions on Computational Logics*

11.1.4 Invited talks

- Nofar Carmeli, Invited Tutorial at ICDT 2022 on *Answering Unions of Conjunctive Queries with Ideal Time Guarantees*
- Leonid Libkin, Invited Talk at KR 2022 on *Graph queries: do we study what matters?*
- Leonid Libkin, Invited Talk at the Workshop on Finite Model Theory and Many-Valued Logics

11.1.5 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.

11.1.6 Research administration

- Luc Segoufin is a member of the CNHSCT of Inria.
- Pierre Senellart is the president of section 6 of the National Committee for Scientific Research.
- Pierre Senellart is a member of the board of the conference of presidents of the national committee (CPCN) and as such a member of the coordination of managing parties of the national committee (C3N).
- Pierre Senellart is deputy director of the DI ENS laboratory, joint between ENS, CNRS, and Inria.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Licence: *Algorithms*, L2, CPES, PSL – Pierre Senellart
- Licence: *Practical Computing*, L3, École normale supérieure – Pierre Senellart
- Licence: *Formal Languages, Computability, Complexity*, L3, École normale supérieure – Michaël Thomazo, Yann Ramusat
- Licence: *Databases*, L3, École normale supérieure – Leonid Libkin, Yann Ramusat
- Master: *Advanced Databases*, M2, IASD – Pierre Senellart, Michaël Thomazo
- Master: *Data wrangling, Data privacy*, M2, IASD – Leonid Libkin, Pierre Senellart
- Master: *Anonymization, privacy*, IASD – Pierre Senellart
- Master: *Knowledge graphs, description logics, reasoning on data*, M2, IASD – Camille Bourgaux, Michaël Thomazo

Pierre Senellart holds various teaching responsibilities (L3 internships, M1 projects, M2 administration, entrance competition) at ENS. Pierre Senellart is in the managing board of the graduate program. Leonid Libkin is co-responsible of the international entrance competition at ENS. Yann Ramusat was the secretary of the entrance competition at ENS for computer science. Michaël Thomazo is an adjunct professor at PSL.

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.

11.2.2 Supervision

- PhD completed: Quentin Manière, Counting queries in ontology-based data access, 2019–2022, Meghyn Bienvenu & Michaël Thomazo (as he was based in Bordeaux, he was not considered a Valda member)
- PhD completed: Yann Ramusat, Provenance-based routing in probabilistic graphs [33], 2018–2022, Silviu Maniu & Pierre Senellart
- PhD in progress: Anatole Dahan, Logical foundations of the polynomial hierarchy, started in October 2020, Arnaud Durand & Luc Segoufin
- PhD in progress: Baptiste Lafosse, Compiler dedicated to the evaluation of SQL queries, started in October 2021, Pierre Senellart & Jean-Marie Lagniez
- PhD in progress: Shrey Mishra, Towards a knowledge base of mathematic results, started in January 2021, Pierre Senellart
- PhD in progress: Alexandra Rogova, Query analytics in Cypher, started October 2021, Amelie Gheerbrant & Leonid Libkin
- PhD in progress: Étienne Toussaint, Paolo Guagliardo & Leonid Libkin (as he is based in Edinburgh, he is not considered a Valda member)
- Internship: Yacine Brihmouche, M1 internship, Pierre Senellart [34]
- Internship: Siméon Gheorghin, L3 internship, Pierre Senellart [35]

11.2.3 Juries

- PhD: Sajad Nazari [reviewer], INSA Centre Val de Loire, Pierre Senellart

11.3 Popularization

11.3.1 Responsibilities

- Serge Abiteboul is a member of the strategic committee of the Blaise Pascal foundation for scientific mediation.
- Pierre Senellart is a scientific expert advising the Scientific and Ethical Committee of Parcoursup, the platform for the selection of first-year higher education students.

11.3.2 Articles and contents

- Serge Abiteboul is a founding editor of the binaire blog for popularizing computer science.
- Serge Abiteboul contributed an interview about artificial intelligence to the May 2022 special edition of *Pour la Science*; this article was among the ten 2022 articles recommanded by the editorial team of the magazine.
- Serge Abiteboul wrote a book on the regulation of social networks [32]
- Serge Abiteboul wrote an article on the carbon impact of 5G [11]

12 Scientific production

12.1 Major publications

[1] 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.

[2] 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.

[3] C. Bourgaux, P. Bourhis, L. Peterfreund and M. Thomazo. 'Revisiting Semiring Provenance for Datalog'. In: KR 2022 - 19th International Conference on Principles of Knowledge Representation and Reasoning. Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning. Haifa, Israel, 31st July 2022, pp. 91–101. DOI: 10.24963/kr.2022/10. URL: https://hal.science/hal-03771031.

[4] C. Bourgaux, D. Carral, M. Krötzsch, S. Rudolph and M. Thomazo. 'Capturing Homomorphism-Closed Decidable Queries with Existential Rules'. In: KR 2021 - 18th International Conference on Principles of Knowledge Representation and Reasoning. Virtual, Vietnam, 3rd Nov. 2021, pp. 141–150. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03345614.

[5] M. Buron, M.-L. Mugnier and M. Thomazo. 'Parallelisable Existential Rules: a Story of Pieces'. In: KR 2021 - 18th International Conference on Principles of Knowledge Representation and Reasoning. Virtual, Vietnam, 3rd Nov. 2021. URL: https://hal.inria.fr/hal-03405745.

[6] 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.

[7] N. Grosshans, P. Mckenzie and L. Segoufin. 'Tameness and the power of programs over monoids in DA'. In: *Logical Methods in Computer Science* 18.3 (2nd Aug. 2022), 14:1–14:34. DOI: 10.46298/lmcs-18(3:14)2022. URL: https://hal.science/hal-03114304.

[8] N. Schweikardt, L. Segoufin and A. Vigny. 'Enumeration for FO Queries over Nowhere Dense Graphs'. In: *Journal of the ACM (JACM)* 69.3 (30th June 2022), pp. 1–37. DOI: 10.1145/3517035. URL: https://hal.inria.fr/hal-03809754.

[9] 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.

[10] E. Toussaint, P. Guagliardo, L. Libkin and J. Sequeda. 'Troubles with nulls, views from the users'. In: *Proceedings of the VLDB Endowment (PVLDB)* 15.11 (July 2022), pp. 2613–2625. DOI: 10.14778/3551793.3551818. URL: https://hal.inria.fr/hal-03934346.

12.2 Publications of the year

International journals

[11] S. Abiteboul and P. Lagrange. '5G : amélioration ou aggravation du bilan carbone ?' In: *Polytechnique Insights* (Mar. 2022). URL: https://hal.inria.fr/hal-03833620.

[12] M. Console, P. Guagliardo and L. Libkin. 'Fragments of bag relational algebra: Expressiveness and certain answers'. In: *Information Systems* 105 (Mar. 2022), p. 101604. DOI: 10.1016/j.is.2020.101604. URL: https://hal.inria.fr/hal-03934340.

[13] M. Console, P. Guagliardo and L. Libkin. 'Propositional and predicate logics of incomplete information'. In: *Artificial Intelligence* 302 (Jan. 2022), p. 103603. DOI: 10.1016/j.artint.2021.103603. URL: https://hal.inria.fr/hal-03934331.

[14] A. Durand, N. Schweikardt and L. Segoufin. 'Enumerating Answers to First-Order Queries over Databases of Low Degree'. In: *Logical Methods in Computer Science* 18.2 (10th May 2022), p. 23. DOI: 10.46298/lmcs-18(2:7)2022. URL: https://hal.inria.fr/hal-03809756.

[15] N. Grosshans, P. Mckenzie and L. Segoufin. 'Tameness and the power of programs over monoids in DA'. In: *Logical Methods in Computer Science* 18.3 (2nd Aug. 2022), 14:1–14:34. DOI: 10.46298/lmcs-18(3:14)2022. URL: https://hal.archives-ouvertes.fr/hal-03114304.

[16] N. Schweikardt, L. Segoufin and A. Vigny. 'Enumeration for FO Queries over Nowhere Dense Graphs'. In: *Journal of the ACM (JACM)* 69.3 (30th June 2022), pp. 1–37. DOI: 10.1145/3517035. URL: https://hal.inria.fr/hal-03809754.

[17] J. Stoyanovich, B. Howe, H. V. Jagadish, S. Schelter and S. Abiteboul. 'Responsible data management'. In: *Communications of the ACM* 65.6 (June 2022), pp. 64–74. DOI: 10.1145/3488717. URL: https://hal.inria.fr/hal-03689050.

[18] E. Toussaint, P. Guagliardo, L. Libkin and J. Sequeda. 'Troubles with nulls, views from the users'. In: *Proceedings of the VLDB Endowment (PVLDB)* 15.11 (July 2022), pp. 2613–2625. DOI: 10.14778/3551793.3551818. URL: https://hal.inria.fr/hal-03934346.

International peer-reviewed conferences

[19] M. Bienvenu and C. Bourgaux. 'Querying Inconsistent Prioritized Data with ORBITS: Algorithms, Implementation, and Experiments'. In: KR 2022 - 19th International Conference on Principles of Knowledge Representation and Reasoning. Haifa, Israel, 31st July 2022. URL: https://hal.archives-ouvertes.fr/hal-03770516.

[20] M. Bienvenu and C. Bourgaux. 'Querying Inconsistent Prioritized Data with ORBITS: Algorithms, Implementation, and Experiments (Extended Abstract)'. In: DL 2022 - 35th International Workshop on Description Logics. Proceedings of the 35th International Workshop on Description Logics. Haifa, Israel, 7th Aug. 2022. URL: https://hal.archives-ouvertes.fr/hal-03801032.

[21] M. Bienvenu, Q. Manière and M. Thomazo. 'Complexity Landscape for Counting Queries'. In: *Proceedings of the 35th International Workshop on Description Logics*. 35th International Workshop on Description Logics. Haifa, Israel, 7th Aug. 2022. URL: https://hal.inria.fr/hal-03896410.

[22] M. Bienvenu, Q. Manière and M. Thomazo. 'Counting Queries over ELHI⊥ Ontologies'. In: KR 2022 - 19th International Conference on Principles of Knowledge Representation and Reasoning. Haifa, Israel, 31st July 2022, pp. 53–62. DOI: 10.24963/kr.2022/6. URL: https://hal.archives-ouvertes.fr/hal-03820249.

[23] C. Bourgaux, P. Bourhis, L. Peterfreund and M. Thomazo. 'Revisiting Semiring Provenance for Datalog'. In: KR 2022 - 19th International Conference on Principles of Knowledge Representation and Reasoning. Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning. Haifa, Israel, 31st July 2022, pp. 91–101. DOI: 10.24963/kr.2022/10. URL: https://hal.archives-ouvertes.fr/hal-03771031.

[24] C. Bourgaux, D. Carral, M. Krötzsch, S. Rudolph and M. Thomazo. 'Capturing Homomorphism-Closed Decidable Queries with Existential Rules (Extended Abstract)'. In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. IJCAI-ECAI 2022 - 31st International Joint Conference on Artificial Intelligence - 25th European Conference on Artificial Intelligence. Vienna, Austria, 23rd July 2022, pp. 5269–5273. DOI: 10.24963/ijcai.2022/733. URL: https://hal.archives-ouvertes.fr/hal-03801049.

[25] K. Bringmann, N. Carmeli and S. Mengel. 'Tight Fine-Grained Bounds for Direct Access on Join Queries'. In: SIGMOD/PODS '22: International Conference on Management of Data. PODS '22: International Conference on Management of Data. Philadelphia PA, United States: ACM, 12th June 2022, pp. 427–436. DOI: 10.1145/3517804.3526234. URL: https://hal.archives-ouvertes.fr/hal-03864970.

[26] D. Carral, L. Larroque, M.-L. Mugnier and M. Thomazo. 'Normalisations of Existential Rules: Not so Innocuous!' In: KR 2022 - 19th International Conference on Principles of Knowledge Representation and Reasoning. Haïfa, Israel, 2022, pp. 102–111. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03762686.

[27] A. Deutsch, N. Francis, A. Green, K. Hare, B. Li, L. Libkin, T. Lindaaker, V. Marsault, W. Martens, J. Michels, F. Murlak, S. Plantikow, P. Selmer, H. Voigt, O. van Rest, D. Vrgoč, M. Wu and F. Zemke. 'Graph Pattern Matching in GQL and SQL/PGQ'. In: SIGMOD '22: International Conference on Management of Data. Philadelphia, United States, 11th June 2022. DOI: 10.1145/3514221.3526057. URL: https://hal.science/hal-03688214.

[28] A. Gheerbrant, L. Libkin, A. Rogova and C. Sirangelo. 'Certain Answers of Extensions of Conjunctive Queries by Datalog and First-Order Rewriting'. In: 4th International Workshop on the Resurgence of Datalog in Academia and Industry. Genoa, Italy, 5th Sept. 2022. URL: https://hal-cnrs.archives-ouvertes.fr/hal-03793035.

[29] M. Liu, A. Padmanabha, R. Ramanujam and Y. Wang. 'Generalized Bundled Fragments for First-Order Modal Logic'. In: *Leibniz International Proceedings in Informatics Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany*. 47th International Symposium on Mathematical Foundations of Computer Science (MFCS 2022). Vol. 241. 47th International Symposium on Mathematical Foundations of Computer Science (MFCS 2022). Vienna, Austria, 22nd Aug. 2022, 70:1–70:14. DOI: 10.4230/LIPIcs.MFCS.2022.70. URL: https://hal.inria.fr/hal-03765358.

[30] Y. Ramusat, S. Maniu and P. Senellart. 'Efficient Provenance-Aware Querying of Graph Databases with Datalog'. In: GRADES-NDA 2022 - Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA). Philadelphia, United States, 12th June 2022. URL: https://hal.inria.fr/hal-03664928.

Conferences without proceedings

[31] M. Buron, M.-L. Mugnier and M. Thomazo. 'Parallelisable Existential Rules: a Story of Pieces'. In: BDA 2022 - 38ème journée "Gestion de Données – Principes, Technologies et Applications". Clermont-Ferrand, France, 24th Oct. 2022. URL: https://hal.inria.fr/hal-03896402.

Scientific books

[32] S. Abiteboul and J. Cattan. *Nous sommes les réseaux sociaux*. Odile Jacob, 2022. URL: https://hal.inria.fr/hal-03794484.

Doctoral dissertations and habilitation theses

[33] Y. Ramusat. 'The Semiring-Based Provenance Framework for Graph Databases'. Ecole normale supérieure - ENS PARIS; PSL University, 28th Apr. 2022. URL: https://hal.inria.fr/tel-03896482.

Other scientific publications

[34] Y. Brihmouche. 'TheoremKB : une base de connaissance des résultats mathématiques'. Paris IX Dauphine, 5th Sept. 2022. URL: https://hal.inria.fr/hal-03897168.

[35] S. Gheorghin. 'Etude de données Twitter en lien avec l'élection présidentielle française d'avril 2022'. Paris: Paris Sciences et Lettres, 27th June 2022, p. 14. URL: https://hal.inria.fr/hal-03897939.

12.3 Cited publications

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

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

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.