RESEARCH CENTRE

Inria Paris Centre

IN PARTNERSHIP WITH:

Ecole normale supérieure de Paris, CNRS

2023
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 1

1 Team members, visitors, external collaborators 3

2 Overall objectives 4

2.1 Objectives . . . 4

2.2 The Issues . . . 5

3 Research program 5

3.1 Scientific Foundations . . . 5

3.2 Research Directions . . . 7

4 Application domains 8

4.1 Personal Information Management Systems . . . 8

4.2 Web Data . . . 9

5 Highlights of the year 9

5.1 Awards . . . 9

6 New software, platforms, open data 9

6.1 New software . . . 9

6.1.1 ProvSQL . . . 9

6.1.2 TheoremKB . . . 10

6.1.3 dissem.in . . . 10

6.2 New platforms . . . 10

6.2.1 dissem.in . . . 10

6.3 Open data . . . 11

7 New results 11

7.1 Querying property graphs . . . 11

7.2 Incomplete and uncertain information . . . 11

7.3 Information extraction and natural language processing . . . 13

7.4 Database theory . . . 13

7.5 Theoretical computer science beyond databases . . . 14

8 Bilateral contracts and grants with industry 14

8.1 Standardization activities . . . 14

9 Partnerships and cooperations 14

9.1 International initiatives . . . 14

9.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program . . . 14

9.1.2 Participation in other International Programs . . . 15

9.2 International research visitors . . . 15

9.2.1 Visits of international scientists . . . 15

9.2.2 Visits to international teams . . . 15

9.3 European initiatives . . . 15

9.3.1 Other european programs/initiatives . . . 15

9.4 National initiatives . . . 16

9.4.1 ANR . . . 16

9.4.2 Others . . . 16

10 Dissemination **16**
10.1 Promoting scientific activities 16
10.1.1 Scientific events: organisation 16
10.1.2 Scientific events: selection 16
10.1.3 Journal 16
10.1.4 Invited talks 17
10.1.5 Leadership within the scientific community 17
10.1.6 Research administration 17
10.2 Teaching - Supervision - Juries 17
10.2.1 Teaching 17
10.2.2 Supervision 18
10.2.3 Juries 19
10.3 Popularization 19
10.3.1 Responsibilities 19
10.3.2 Articles and contents 19
10.3.3 Education 19
10.3.4 Interventions 19

11 Scientific production **19**
11.1 Major publications 19
11.2 Publications of the year 20
11.3 Cited publications 22

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. – Data and knowledge analysis

A3.3.1. – On-line analytical processing

A3.3.2. – Data mining

A3.3.3. – Big data analysis

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

- B2. – Health
- B3.3. – Geosciences
- B4. – Energy
- B4.2. – Nuclear Energy Production
- B9.3. – Medias
- 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]
- Paul Boniol [Inria, ISFP, from Dec 2023]
- Camille Bourgaux [CNRS, Researcher]
- Luc Segoufin [Inria, Senior Researcher, HDR]
- Michaël Thomazo [Inria, Researcher, HDR]

Faculty Members

- Pierre Senellart [Team leader, ENS, Professor, HDR]
- Leonid Libkin [ENS, Professor, until Aug 2023]
- Cristina Sirangelo [Université Paris-Cité, Professor Delegation, from Feb 2023 until Jul 2023, HDR]

Post-Doctoral Fellows

- Sven Dziadek [Inria, Post-Doctoral Fellow, from May 2023]
- Shufan Jiang [ENS, until Sep 2023, ATER]
- Anantha Matikurke Shankara Narayana [ENS, Post-Doctoral Fellow, until Jun 2023]

PhD Students

- Anatole Dahan [Université Paris-Cité & Inria]
- Antoine Gauquier [ENS, from Sep 2023]
- Robin Jean [CNRS, from Oct 2023]
- Baptiste Lafosse [ENS]
- Lucas Larroque [ENS, from Sep 2023]
- Shrey Mishra [ENS]
- Alexandra Rogova [Université Paris-Cité, until Aug 2023]

Technical Staff

- N. Smith [ENS, Engineer]

Interns and Apprentices

- Sarah Benamara [Inria, Intern, from Apr 2023 until Sep 2023]
- Belkis Djeffal [ENS, Intern, from Feb 2023 until Jun 2023]
- Antoine Gauquier [ENS, Intern, until Jul 2023]
- Atefe Khodadaditaghanaki [ENS, Intern, from Apr 2023 until Sep 2023]
- Antonia Labarca Sanchez [Inria, Intern, until Apr 2023]
- Lucas Larroque [ENS, Intern, from May 2023 until Jul 2023]
- Ilyas Lebleu [ENS, Intern, from May 2023 until Sep 2023]

Administrative Assistant

- Meriem Guemair [Inria]

Visiting Scientists

- Thomas Schwentick [TU Dortmund, from Oct 2023]
- Victor Vianu [UC San Diego & Inria, from Jul 2023 until Sep 2023]

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 [38, 47] 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 [36] [39], and possibly distributed [70] 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.

[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*.

These problems have already been studied individually and have led to techniques such as *query rewriting* [60] or *distributed query optimization* [65].

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 [53] or aggregated search [78]. 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 [76] (train an agent to accomplish a task in a partially known world based on rewards obtained) and of active learning [72] (decide which 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 [38]. Since these early developments, logic has also been successfully used to capture a large variety of query modes, such as data aggregation [64], recursive queries (Datalog), or querying of XML databases [47]. 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 [77] 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 [37], queries, expressed in temporal logics, can often by compiled to automata; in graph databases [43], 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 [69], or for more complex languages to data tree automata[33]. Another reason of the importance of automata theory, and tree automata in particular, comes from Courcelle's results [51] 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* [38]. 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 $(q \land \neg q') \lor (\neg q \land q')$ 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 [49].

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 [35].

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 [38]. 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* [62] 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 [58], and to a probabilistic formalism by assigning independent probabilities to the Boolean events [59]. These two extensions form the basis of modern provenance and probability management, subsuming in a large way previous works [50, 44]. 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 [75] [63, 41].

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 [71], crowdsourcing [42], focused crawling [57], or automatic database tuning [45] critically rely on machine learning techniques, such as classification [61], probabilistic models [56], or reinforcement learning [76].

Machine learning is also a rich source of complex data management problems: thus, the probabilities produced by a conditional random field [66] 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* [72], 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 [73] consider more realistic costs. Also, oracles are usually assumed to be perfect with only a few exceptions [54]. 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 been working on aspects related to connecting their research with database theory, and investigating aspects of uncertainty and incompleteness in their research. This leads 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.

- Paul Boniol is reinforcing the team activity on systems aspects.

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 [52] on probabilistic query evaluation with the instance-based one of [40, 41]. 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) [34].

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 [71].

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 [74] 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 [39] in a broad sense: semi-structured, structured, or unstructured content extracted from Web databases [71]; knowledge bases from the Semantic Web [74]; social networks [67]; Web archives and Web crawls [55]; Web applications and deep Web databases [48]; crowdsourcing platforms [42]. We intend to continue using Web data as a natural application domain for the research within Valda when relevant. For instance [46], 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

- Michaël Thomazo defended his *Habilitation to supervise research* [27] in October 2023.

- Paul Boniol was hired as an ISFP in Valda in December 2023, while Leonid Libkin left Valda in September 2023 to take a position with Relational AI & Université Paris-Cité.

5.1 Awards

- Leonid Libkin received the ICDT test-of-time award for his ICDT 2013 paper [68]

- Leonid Libkin received the SIGMOD 2023 Industry Track best paper award [14]

- Leonid Libkin and Alexandra Rogova received the test-of-time award at the French BDA 2023 conference for their PODS 2023 paper [20]

- As part of the group working on the SQL/PGQ standard, Leonid Libkin received the INCITS 2023 Team award

6 New software, platforms, open data

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: Support for Shapley and expected Shapley value computation. Better continuous integration. Support for PostgreSQL 16. Various optimizations and bug fixes.

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

Publications: hal-01672566, hal-01851538

Contact: Pierre Senellart

Participants: Pierre Senellart, Baptiste Lafosse

6.1.2 TheoremKB

Keyword: Information extraction

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

News of the Year: Full mutlimodal model, see https://github.com/mv96/mm_extraction

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, Antoine Gauquier

6.1.3 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: Various bug fixes. Work on crossref ingestion.

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

Contact: Pierre Senellart

Participant: Pierre Senellart

Partner: CAPSH

6.2 New platforms

6.2.1 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 (6.1.3) software, has been maintained by Valda since 2021. Works on the platform in 2023, in addition to works on the base software, include updating information using recent data from CrossRef.

Participants: Pierre Senellart, N. Smith.

6.3 Open data

- A number of machine learning models, trained in the context of various works related to the TheoremKB project, were released on HuggingFace. **Contacts:** Shrey Mishra, Pierre Senellart
- Instructions to obtain the dataset and reproduce the experiments from [30] are provided in a companion repository on Github. **Contacts:** Antoine Gauquier, Pierre Senellart
- Instructions to obtain the dataset and reproduce the experiments from [22] are provided in a companion repository on Github. **Contacts:** Shufan Jiang, Pierre Senellart

7 New results

We present the results we obtained and published in 2023. In 2023, we achieved results in different areas of data management and data science: property graph databases, management of incomplete and uncertain information, information extraction, as well as foundational problems in database theory. We describe our works in each of these areas in turn, and finish with other theoretical research conducted in the team, beyond data management.

7.1 Querying property graphs

GQL (Graph Query Language) is being developed as a new ISO standard for graph query languages to play the same role for graph databases as SQL plays for relational. In parallel, an extension of SQL for querying property graphs, SQL/PGQ, is added to the SQL standard; it shares the graph pattern matching functionality with GQL. Both standards (not yet published) are hard-to-understand specifications of hundreds of pages. The goal of [13] is to present a digest of the language that is easy for the research community to understand, and thus to initiate research on these future standards for querying graphs. The paper concentrates on pattern matching features shared by GQL and SQL/PGQ, as well as querying facilities of GQL.

Indeed, the development of practical query languages for graph databases runs well ahead of the underlying theory. The main component of both GQL and SQL/PGQ is the pattern matching facility, which is shared by the two standards. In many aspects, it goes well beyond RPQs, CRPQs, and similar queries on which the research community has focused for years. Our main contribution in [20] is to distill the lengthy standard specification into a simple Graph Pattern Calculus (GPC) that reflects all the key pattern matching features of GQL and SQL/PGQ, and at the same time lends itself to rigorous theoretical investigation. We describe the syntax and semantics of GPC, along with the typing rules that ensure its expressions are well-defined, and state some basic properties of the language. With this paper we provide the community a tool to embark on a study of query languages that will soon be widely adopted by industry.

Despite documented demand, schema support is limited both in existing systems and in the first version of the GQL Standard. It is anticipated that the second version of the GQL Standard will include a rich DDL. Aiming to inspire the development of GQL and enhance the capabilities of graph database systems, we propose in [14] PG-Schema, a simple yet powerful formalism for specifying property graph schemas. It features PG-Schema with flexible type definitions supporting multi-inheritance, as well as expressive constraints based on the recently proposed PG-Keys formalism. We provide the formal syntax and semantics of PG-Schema, which meet principled design requirements grounded in contemporary property graph management scenarios, and offer a detailed comparison of its features with those of existing schema languages and graph database systems.

7.2 Incomplete and uncertain information

In [15], we revisit the problem of repairing and querying inconsistent databases equipped with universal constraints. We adopt symmetric difference repairs, in which both deletions and additions of facts can be used to restore consistency, and suppose that preferred repair actions are specified via a binary priority relation over (negated) facts. Our first contribution is to show how existing notions of optimal repairs, defined for simpler denial constraints and repairs solely based on fact deletion, can be suitably extended to our richer setting. We next study the computational properties of the resulting repair notions, in particular, the data complexity of repair checking and inconsistency-tolerant query answering. Finally, we clarify the relationship between optimal repairs of prioritized databases and repair notions introduced in the framework of active integrity constraints. In particular, we show that Pareto-optimal repairs in our setting correspond to founded, grounded and justified repairs w.r.t. the active integrity constraints obtained by translating the prioritized database. Our study also yields useful insights into the behavior of active integrity constraints.

In [19], we consider the dichotomy conjecture for consistent query answering under primary key constraints stating that for every fixed Boolean conjunctive query q, testing whether it is certain over all repairs of a given inconsistent database is either polynomial time or coNP-complete. This conjecture has been verified for self-join-free and path queries. We propose a simple inflationary fixpoint algorithm for consistent query answering which, for a given database, naively computes a set Δ of subsets of database repairs with at most k facts, where k is the size of the query q. The algorithm runs in polynomial time and can be formally defined as: 1. Initialize Δ with all sets S of at most k facts such that S satisfies q. 2. Add any set S of at most k facts to Δ if there exists a block B (ie, a maximal set of facts sharing the same key) such that for every fact a of B there is a set $S' \in \Delta$ contained in $(S \cup \{a\})$. The algorithm answers "q is certain" iff Δ eventually contains the empty set. The algorithm correctly computes certain answers when the query q falls in the polynomial time cases for self-join-free queries and path queries. For arbitrary queries, the algorithm is an under-approximation: The query is guaranteed to be certain if the algorithm claims so. However, there are polynomial time certain queries (with self-joins) which are not identified as such by the algorithm.

The design of SQL is based on a three-valued logic (3VL), rather than the familiar two-valued Boolean logic (2VL). In addition totrue andfalse, 3VL addsunknown to handle nulls. Viewed as indispensable for SQL expressiveness, it is often criticized for unintuitive behavior of queries and for being a source of programmer mistakes. In [23], we show that, contrary to the widely held view, SQL could have been designed based on 2VL, without any loss of expressiveness. Similarly to SQL's WHERE clause, which only keeps true tuples, we conflate false and unknown for conditions involving nulls to obtain an equally expressive 2VL-based version of SQL. This applies to the core of the 1999 SQL Standard. Queries written under the 2VL semantics can be efficiently translated into the 3VL SQL and thus executed on any existing RDBMS. We show that 2VL enables additional optimizations. To gauge its applicability, we establish criteria under which 2VL and 3VL semantics coincide, and analyze common benchmarks such as TPC-H and TPC-DS to show that most of their queries are such. For queries that behave differently under 2VL and 3VL, we undertake a user study to show a consistent preference for the 2VL semantics.

Queries with aggregation and arithmetic operations, as well as incomplete data, are common in real-world database, but we lack a good understanding of how they should interact. On the one hand, systems based on SQL provide ad-hoc rules for numerical nulls, on the other, theoretical research largely concentrates on the standard notions of certain and possible answers. In the presence of numerical attributes and aggregates, however, these answers are often meaningless, returning either too little or too much. Our goal in [17] is to define a principled framework for databases with numerical nulls and answering queries with arithmetic and aggregations over them. Towards this goal, we assume that missing values in numerical attributes are given by probability distributions associated with marked nulls. This yields a model of probabilistic bag databases in which tuples are not necessarily independent, since nulls can repeat. We provide a general compositional framework for query answering, and then concentrate on queries that resemble standard SQL with arithmetic and aggregation. We show that these queries are measurable, and that their outputs have a finite representation. Moreover, since the classical forms of answers provide little information in the numerical setting, we look at the probability that numerical values in output tuples belong to specific intervals. Even though their exact computation is intractable, we show efficient approximation algorithms to compute such probabilities.

Federated knowledge discovery and data mining are challenged to assess the trustworthiness of data originating from autonomous sources while protecting confidentiality and privacy. Truth-finding algorithms help corroborate data from disagreeing sources. For each query it receives, a truth-finding algorithm predicts a truth value of the answer, possibly updating the trustworthiness factor of each source. Few works, however, address the issues of confidentiality and privacy. In [24, 28], we devise and present a secure secret-sharing-based multi-party computation protocol for pseudo-equality tests that are used in truth-finding algorithms to compute additions depending on a condition. The protocol guarantees confidentiality of the data and privacy of the sources. We also present a variants of a truth-finding algorithm that would make the computation faster when executed using secure multi-party computation. We empirically evaluate the performance of the proposed protocol on a state-of-the-art truth-finding algorithm, 3-Estimates, and compare it with that of the baseline plain algorithm. The results confirm that the secret-sharing-based secure multi-party algorithms are as accurate as the corresponding baselines but for proposed numerical approximations that significantly reduce the efficiency loss incurred.

7.3 Information extraction and natural language processing

In [21], we consider the problem of automatically inferring the (LaTeX) document class used to write a scientific article from its PDF representation. Applications include improving the performance of information extraction techniques that rely on the style used in each document class, or determining the publisher of a given scientific article. We introduce two approaches: a simple classifier based on hand-coded document style features, as well as a CNN-based classifier taking as input the bitmap representation of the first page of the PDF article. We experiment on a dataset of around 100k articles from arXiv, where labels come from the source LaTeX document associated to each article. Results show the CNN approach significantly outperforms that based on simple document style features, reaching over 90% average F1-score on a task to distinguish among several dozens of the most common document classes.

We consider in [22] automatically identifying the defined term within a mathematical definition from the text of an academic article. Inspired by the development of transformer-based natural language processing applications, we pose the problem as (a) a token-level classification task using fine-tuned pre-trained transformers; and (b) a question-answering task using a generalist large language model (GPT). We also propose a rule-based approach to build a labeled dataset from the LaTeX source of papers. Experimental results show that it is possible to reach high levels of precision and recall using either recent (and expensive) GPT 4 or simpler pre-trained models fine-tuned on our task.

The Bidirectional Encoder Representations from Transformers (BERT) architecture offers a cutting-edge approach to Natural Language Processing. It involves two steps: 1) pre-training a language model to extract contextualized features and 2) fine-tuning for specific downstream tasks. Although pre-trained language models (PLMs) have been successful in various text-mining applications, challenges remain, particularly in areas with limited labeled data such as plant health hazard detection from individuals' observations. To address this challenge, we propose in [12] to combine GAN-BERT, a model that extends the fine-tuning process with unlabeled data through a Generative Adversarial Network (GAN), with ChouBERT, a domain-specific PLM. Our results show that GAN-BERT outperforms traditional fine-tuning in multiple text classification tasks. In this paper, we examine the impact of further pre-training on the GAN-BERT model. We experiment with different hyper parameters to determine the best combination of models and fine-tuning parameters. Our findings suggest that the combination of GAN and ChouBERT can enhance the generalizability of the text classifier but may also lead to increased instability during training. Finally, we provide recommendations to mitigate these instabilities.

7.4 Database theory

Even though query evaluation is a fundamental task in databases, known classifications of conjunctive queries by their fine-grained complexity only apply to queries without self-joins. We study in [16] how self-joins affect enumeration complexity, with the aim of building upon the known results to achieve general classifications. We do this by examining the extension of two known dichotomies: one with respect to linear delay, and one with respect to constant delay after linear preprocessing. As this turns out to be an intricate investigation, this paper is structured as an example-driven discussion that initiates this analysis. We show enumeration algorithms that rely on self-joins to efficiently evaluate queries that otherwise cannot be answered with the same guarantees. Due to these additional tractable cases, the hardness proofs are more complex than the self-join-free case. We show how to harness a known tagging technique to prove hardness of queries with self-joins. Our study offers sufficient conditions and necessary conditions for tractability and settles the cases of queries of low arity and queries with cyclic cores. Nevertheless, many cases remain open.

In [11], we study the question of when we can provide direct access to the k-th answer to a Conjunctive Query (CQ) according to a specified order over the answers in time logarithmic in the size of the database, following a preprocessing step that constructs a data structure in time quasilinear in database size. Specifically, we embark on the challenge of identifying the tractable answer orderings , that is, those orders that allow for such complexity guarantees. To better understand the computational challenge at hand, we also investigate the more modest task of providing access to only a single answer (i.e., finding the answer at a given position), a task that we refer to as the selection problem , and ask when it can be performed in quasilinear time. We also explore the question of when selection is indeed easier than ranked direct access. We begin with lexicographic orders . For each of the two problems, we give a decidable characterization (under conventional complexity assumptions) of the class of tractable lexicographic orders for every CQ without self-joins. We then continue to the more general orders by the sum of attribute weights and establish the corresponding decidable characterizations, for each of the two problems, of the tractable CQs without self-joins. Finally, we explore the question of when the satisfaction of Functional Dependencies (FDs) can be utilized for tractability and establish the corresponding generalizations of our characterizations for every set of unary FDs.

Existential rules are an expressive knowledge representation language mainly developed to query data. In the literature, they 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 [25] 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.

7.5 Theoretical computer science beyond databases

In [18], we show how to efficiently solve energy Büchi problems in finite weighted automata and in one-clock weighted timed automata. Solving the former problem is our main contribution and is handled by a modified version of Bellman-Ford interleaved with Couvreur's algorithm. The latter problem is handled via a reduction to the former relying on the corner-point abstraction. All our algorithms are freely available and implemented in a tool based on the open-source platforms TChecker and Spot.

8 Bilateral contracts and grants with industry

8.1 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.

Participants: Leonid Libkin.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 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) and Leonid Libkin

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.

9.1.2 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.

9.2 International research visitors

9.2.1 Visits of international scientists

- Victor Vianu, Professor at UCSD, visited the group during several months in 2023.
- Thomas Schwentick, Professor at TU Dortmund, visited the group during several months in 2023.
- Avijeet Gosh, PhD student at the Indian Statistical Institute in Chennai, visited Valda for one week in September 2023.

9.2.2 Visits to international teams

Research stays abroad Pierre Senellart was an invited participant of the *Logic and Algorithms in Database Theory and AI* workshop at the Simons Institute, UC Berkeley, where he spent four weeks.

9.3 European initiatives

9.3.1 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.
- Valda is involved in a PHC AURORA project with Ana Ozaki (University of Bergen), on *Learning and Reasoning in Knowledge Graph Embeddings*.

9.4 National initiatives

9.4.1 ANR

Valda has been part of three national ANR projects in 2023:

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.

9.4.2 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.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

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* (until summer 2023)
- Luc Segoufin, member of the steering committee of STACS (since September 2023)
- Pierre Senellart, editorial board of the *LIPIcs* series of conference proceedings

10.1.2 Scientific events: selection

Member of the conference program committees

- Camille Bourgaux, IJCAI 2023, ECAI 2023, KR 2023, DL 2023, AAAI 2024
- Leonid Libkin, KR 2023 (area chair), WWW (industry track), ICDE 2023 (industry track)
- Michaël Thomazo, IJCAI 2023
- Pierre Senellart, ECML/PKDD 2023, BDA 2023, Provenance Week 2023

10.1.3 Journal

Member of the editorial boards

- Leonid Libkin, *Acta Informatica*
- Luc Segoufin, *ACM Transactions on Computational Logics*

10.1.4 Invited talks

- Shufan Jiang, invited tutorial at the microandbig2023 thematic school
- Leonid Libkin, keynote talk at EDBT/ICDT 2023
- Michaël Thomazo, invited lecture at the *Reasoning Web* 2023 summer school

10.1.5 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.

10.1.6 Research administration

- Luc Segoufin was a member of the CNHSCT of Inria (until summer 2023) and is now a member of the LSS of Inria Paris.
- Pierre Senellart is the president of section 6 of the National Committee for Scientific Research. As a representative of CoNRS, Pierre Senellart was in the Hcéres evaluation committee of the I3S and IRIF research units.
- 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.
- Pierre Senellart is the scientific resource person for *Scientific information & edition* of the Inria Paris centre

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Licence: *Algorithms*, L1, CPES, PSL – Pierre Senellart
- Licence: *Practical Computing*, L3, École normale supérieure – Pierre Senellart
- Licence: *Formal Languages, Computability, Complexity*, L3, École normale supérieure – Michaël Thomazo, Lucas Larroque
- Licence: *Databases*, L3, École normale supérieure – Leonid Libkin

- Master: *Logiques de description*, M1, DCI – Camille Bourgaux
- Master: *Data Acquisition, Extraction, and Storage*, M2, IASD – Pierre Senellart, Michaël Thomazo
- Master: *Data wrangling, Data privacy*, M2, IASD – Leonid Libkin, Pierre Senellart
- Master: *Description logics and reasoning on data*, M2, LMFI & IASD – Camille Bourgaux, Michaël Thomazo

- Professional training: *Web Security*, PESTO (*Corps des Mines* professional training) – Pierre Senellart

Pierre Senellart holds various teaching responsibilities (M1 projects, M2 administration, entrance competition) at ENS. Pierre Senellart is in the managing board of the graduate program of PSL.

As an adjunct professor at PSL, Michaël Thomazo is in charge of PhD committees within DI ENS and co-responsible of the international entrance competition at ENS. Shufan Jiangwas 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 in progress: Anatole Dahan, Logical foundations of the polynomial hierarchy, started in October 2020, Arnaud Durand & Luc Segoufin
- PhD in progress: Antoine Gauquier, Intelligent construction of a multimodal and heterogeneous data warehouse, with data traceability, started in September 2023, Pierre Senellart & Ioana Manolescu
- PhD in progress: Robin Jean, Integration of preferences and domain knowledge in inconsistency-tolerant ontology-based data access , started in October 2023, Meghyn Bienvenu & Camille Bourgaux
- 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: Lucas Larroque, Extension of rewriting procedures for reasoning using existential rules, started in September 2023, Michaël Thomazo
- 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: Pratik Karmakar, Stéphane Bressan & Pierre Senellart (as he is based in Singapore, he is not considered a Valda member)
- PhD in progress: Étienne Toussaint, Paolo Guagliardo & Leonid Libkin (as he is based in Edinburgh, he is not considered a Valda member)
- Internship: Sarah Benamara, M2 internship, Michaël Thomazo
- Internship: Belkis Djeffal, M2 internship, Pierre Senellart [29]
- Internship: Antoine Gauquier, research project & M2 internship, Pierre Senellart [30, 31]
- Internship: Atefe Khodadaditaghanaki, M2 internship, Camille Bourgaux & Meghyn Bienvenu
- Antonia Labarca Sanchez, internship in the framework of the mobility agreements signed between Inria, Inria Chile and Universidad de Chile, Camille Bourgaux & Michaël Thomazo
- Internship: Lucas Larroque, M2 internship, Michaël Thomazo [32]
- Internship: Ilyas Lebleu, M2 internship, Pierre Senellart

10.2.3 Juries

- PhD: Claire Soyez-Martin [president], Université de Lille, Pierre Senellart
- PhD: Pierre Faure–Giovagnoli [president] INSA Lyon, Pierre Senellart
- PhD: Rébecca Zucchini [reviewer], Université Paris-Saclay, Pierre Senellart
- PhD: Antonia Kormpa [reviewer] University of Oxford, Pierre Senellart
- PhD: Gaston Zanitti [reviewer & president], Université Paris-Saclay, Pierre Senellart
- PhD: Nicolas Crosetti [reviewer], Université de Lille, Pierre Senellart

10.3 Popularization

10.3.1 Responsibilities

- Serge Abiteboul is a member of the strategic committee of the Blaise Pascal foundation for scientific mediation.
- Serge Abiteboul is the president of the newly established scientific council of the DGFiP (Direction générale des finances publiques).
- Serge Abiteboul was a member of the scientific board of the exhibition *Évolutions industrielles* at the Cité des Sciences et de l'Industrie
- 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.

10.3.2 Articles and contents

Serge Abiteboul is the author of a theatre play, *Qui a hacké Garoutzia?* on artificial intelligence, with Gilles Dowek and Laurence Devillers. This play was premiered at the "off" festival in Avignon on July 15–17 and is published with C&F Editions

10.3.3 Education

Michaël Thomazo presented the job of a researcher at the Maurice Ravel high school in Paris (2 classes)

10.3.4 Interventions

Serge Abiteboul organized two events as a member of the Académie des Sciences:

- *Lire et écrire au temps des algorithmes* (lectures of the Institut de France)
- *L'intelligence artificielle est-elle intelligente, Des clés pour comprendre*, roundtable within the Institut de France

11 Scientific production

11.1 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.

11.2 Publications of the year

International journals

[11] N. Carmeli, N. Tziavelis, W. Gatterbauer, B. Kimelfeld and M. Riedewald. 'Tractable Orders for Direct Access to Ranked Answers of Conjunctive Queries'. In: *ACM Transactions on Database Systems* 48.1 (31st Mar. 2023), pp. 1–45. DOI: 10.1145/3578517. URL: https://inria.hal.science/hal-04278100.

[12] S. Jiang, S. Cormier, R. Angarita and F. Rousseaux. 'Improving text mining in plant health domain with GAN and/or pre-trained language model'. In: *Frontiers in Artificial Intelligence* 6 (21st Feb. 2023), p. 1072329. DOI: 10.3389/frai.2023.1072329. URL: https://hal.science/hal-04008864.

Invited conferences

[13] N. Francis, A. Gheerbrant, P. Guagliardo, L. Libkin, V. Marsault, W. Martens, F. Murlak, L. Peterfreund, A. Rogova and D. Vrgoč. 'A Researcher's Digest of GQL'. In: *ICDT'23*. 26th International Conference on Database Theory (ICDT 2023). Vol. 255. Ioannina, Greece, 17th Mar. 2023. DOI: 10.4230/LIPIcs.ICDT.2023.1. URL: https://hal.science/hal-04094449.

International peer-reviewed conferences

[14] R. Angles, A. Bonifati, S. Dumbrava, G. Fletcher, A. Green, J. Hidders, B. Li, L. Libkin, V. Marsault, W. Martens, F. Murlak, S. Plantikow, O. Savkovic, M. Schmidt, J. Sequeda, S. Staworko, D. Tomaszuk, H. Voigt, D. Vrgoč, M. Wu and D. Zivkovic. 'PG-Schema: schemas for property graphs'. In: *SIGMOD 2023 : ACM SIGMOD International Conference on Management of Data*. ACM SIGMOD International Conference on Management of Data (SIGMOD). Vol. 1. 2. Seattle, WA, United States: ACM, 20th June 2023, pp. 1–25. DOI: 10.1145/3589778. URL: https://hal.science/hal-04224583.

[15] M. Bienvenu and C. Bourgaux. 'Inconsistency Handling in Prioritized Databases with Universal Constraints: Complexity Analysis and Links with Active Integrity Constraints'. In: *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023)*. 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023). Rhodes, Greece: International Joint Conferences on Artificial Intelligence Organization, 2nd Sept. 2023, pp. 97–106. DOI: 10.24963/kr.2023/10. URL: https://hal.science/hal-04204432.

[16] N. Carmeli and L. Segoufin. 'Conjunctive Queries With Self-Joins, Towards a Fine-Grained Complexity Analysis'. In: PODS'23. Seattle, United States, 18th June 2023. URL: https://inria.hal.science/hal-04136055.

[17] M. Console, L. Libkin and L. Peterfreund. 'Querying Incomplete Numerical Data: Between Certain and Possibile Answers'. In: SIGMOD/PODS '23: International Conference on Management of Data. Seattle WA USA, France: ACM, 18th June 2023, pp. 349–358. DOI: 10.1145/3584372.3588660. URL: https://inria.hal.science/hal-04408298.

[18] S. Dziadek, U. Fahrenberg and P. Schlehuber-Caissier. 'Energy Büchi Problems'. In: *LNCS - Lecture Notes in Computer Science*. FM 2023 - 25th International Symposium on Formal Methods. Vol. 14000. Lübeck, Germany: Springer, 2023, pp. 222–239. DOI: 10.1007/978-3-031-27481-7_14. URL: https://inria.hal.science/hal-04344167.

[19] D. Figueira, A. Padmanabha, L. Segoufin and C. Sirangelo. 'A Simple Algorithm for Consistent Query Answering under Primary Keys'. In: *LIPIcs*. ICDT 2023 - International Conference on Database Theory. Proceedings of the 26th International Conference on Database Theory (ICDT'23). Ioannina, Greece, 28th Mar. 2023. URL: https://hal.science/hal-03953588.

[20] N. Francis, A. Gheerbrant, P. Guagliardo, L. Libkin, V. Marsault, W. Martens, F. Murlak, L. Peterfreund, A. Rogova and D. Vrgoč. 'GPC: A Pattern Calculus for Property Graphs'. In: *Symposium on Principles of Database Systems (PODS)*. Symposium on Principles of Database Systems (PODS). Seattle WA USA, France: ACM, June 2023, pp. 241–250. DOI: 10.1145/3584372.3588662. URL: https://hal.science/hal-03836782.

[21] A. Gauquier and P. Senellart. 'Automatically Inferring the Document Class of a Scientific Article'. In: DocEng 2023 - 23rd ACM Symposium on Document Engineering. Limerick, Ireland, 22nd Aug. 2023. DOI: 10.1145/3573128.3604894. URL: https://inria.hal.science/hal-04138880.

[22] S. Jiang and P. Senellart. 'Extracting Definienda in Mathematical Scholarly Articles with Transformers'. In: *Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications*. The 2nd Workshop on Information Extraction from Scientific Publications at IJCNLP-AACL 2023. Online, Indonesia, 2023. URL: https://hal.science/hal-04282533.

[23] L. Libkin and L. Peterfreund. 'SQL Nulls and Two-Valued Logic'. In: SIGMOD/PODS '23: International Conference on Management of Data. Seattle WA USA, France: ACM, 18th June 2023, pp. 11–20. DOI: 10.1145/3584372.3588661. URL: https://inria.hal.science/hal-04408295.

[24] A. Saadeh, P. Senellart and S. Bressan. 'Confidential Truth Finding with Multi-Party Computation'. In: DEXA 2023 - 34th International Conference on Database and Expert Systems Applications. Penang, Malaysia, 28th Aug. 2023. URL: https://inria.hal.science/hal-04139281.

National peer-reviewed Conferences

[25] D. Carral, L. Larroque, M. Thomazo and M.-L. Mugnier. 'Normalisations of Existential Rules: Not so Innocuous!' In: BDA 2023 - 39e Conférence sur la Gestion de Données – Principes, Technologies et Applications. Montpellier, France, 23rd Oct. 2023. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04315377.

Conferences without proceedings

[26] S. Jiang, R. Angarita, S. Cormier, J. Orensanz and F. Rousseaux. 'ChouBERT : Pré-entraînement d'un modèle de langue française pour le Crowdsensing avec des Tweets dans un contexte phytosanitaire'. In: INFORSID 2023 - INFormatique des ORganisations et Systèmes d'Information et de Décision. La Rochelle, France, 30th May 2023. URL: https://hal.science/hal-04377395.

Doctoral dissertations and habilitation theses

[27] M. Thomazo. 'Ontology-Based Query Answering: Expressivity and Extensions'. Ecole Normale Supérieure de Paris, 4th Oct. 2023. URL: https://hal.science/tel-04275013.

Reports & preprints

[28] A. Saadeh, P. Senellart and S. Bressan. *Confidential Truth Finding with Multi-Party Computation (Extended Version)*. 24th May 2023. DOI: 10.48550/arXiv.2305.14727. URL: https://inria.hal.science/hal-04139243.

Other scientific publications

[29] B. Djeffal. 'Étendre un système de gestion de provenance : ProvSQL'. Université des Sciences et de la Technologie Houari Boumediene (Algérie), 2nd July 2023. URL: https://inria.hal.science/hal-04342025.

[30] A. Gauquier. 'Automatically inferring the document class used in a scientific article'. Paris: Télécom Paris, 28th Feb. 2023, p. 31. URL: https://inria.hal.science/hal-04379415.

[31] A. Gauquier. 'Impact of the document class in the automatic extraction of mathematical environments in the scientific literature'. IMT Nord Europe, 11th July 2023, p. 44. URL: https://inria.hal.science/hal-04220990.

[32] L. Larroque. 'Ontology-Based Query Answering Over Datalog-Expressible Rule Sets Is Undecidable'. Université paris diderot, 20th Sept. 2023, p. 32. URL: https://inria.hal.science/hal-04347020.

11.3 Cited publications

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

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

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] L. Libkin, W. Martens and D. Vrgoc. 'Querying graph databases with XPath'. In: *Joint 2013 EDBT/ICDT Conferences, ICDT '13 Proceedings, Genoa, Italy, March 18-22, 2013.* Ed. by W. Tan, G. Guerrini, B. Catania and A. Gounaris. ACM, 2013, pp. 129–140. DOI: 10.1145/2448496.2448513. URL: https://doi.org/10.1145/2448496.2448513.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.