IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2018

Project-Team VALDA

Value from Data

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

RESEARCH CENTER
Paris

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team VALDA

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

Keywords:

Computer Science and Digital Science:

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, Visitors, External Collaborators

Research Scientists

Serge Abiteboul [Inria, Senior Researcher, until Jan 2018, HDR]
Luc Segoufin [Inria, Senior Researcher, HDR]
Michael Thomazo [Inria, Researcher, from Apr 2018]
Camille Bourgaux [CNRS, Researcher, from Oct 2018]
Olivier Cappe [CNRS, Senior Researcher, from Feb 2018]

Faculty Member

Pierre Senellart [Team leader, École normale supérieure, Professor, HDR]

Post-Doctoral Fellow

Nathan Grosshans [École normale supérieure, from Sep 2018, temporary research and teaching assistant (ATER)]

PhD Students

Julien Grange [École normale supérieure]
Miyoung Han [Télécom ParisTech, until Aug 2018]
Quentin Lobbe [Télécom ParisTech, until Nov 2018]
Mikael Monet [Télécom ParisTech, until Oct 2018]
Karima Rafes [BorderCloud]
Yann Ramusat [École normale supérieure, from Sep 2018]
Alexandre Vigny [Université Paris Diderot, until Sep 2018]
Russac Yoan [École normale supérieure, from Dec 2018]

Administrative Assistants

Meriem Henni [Inria, from Apr 2018]
Sandrine Vergès [Inria, until Jun 2018]

Visiting Scientist

Victor Vianu [UCSD & École normale supérieure, from Jun 2018 until Sep 2018]

External Collaborators

Serge Abiteboul [ARCEP, Board Member, from Feb 2018, HDR]
Yann Ramusat [École normale supérieure, student on a long-term project, until Aug 2018]

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 [43], [55] 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 [40] [45], and possibly distributed [91] 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* [71] or *distributed query optimization* [77].

Among all these aspects, intensionality is perhaps the one that has least been studied, so we will 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 [61] or aggregated search [90]. 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 [88] (train an agent to accomplish a task in a partially known world based on rewards obtained) and of active learning [85] (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

[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*.

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.

3.1.1. 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 [43]. Since these early developments, logic has also been successfully used to capture a large variety of query modes, such as data aggregation [76], recursive queries (Datalog), or querying of XML databases [55]. 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 [89] 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.

3.1.2. Automata Theory.

Automata theory and formal languages arise as important components of the study of many data management tasks: in temporal databases [42], queries, expressed in temporal logics, can often by compiled to automata; in graph databases [51], 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 [81], or for more complex languages to data tree automata[4]. Another reason of the importance of automata theory, and tree automata in particular, comes from Courcelle's results [59] 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.

3.1.3. 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* [43]. 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 [57].

3.1.4. 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 [39].

3.1.5. 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 [43]. 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* [73] 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 [69], and to a probabilistic formalism by assigning independent probabilities to the Boolean events [70]. These two extensions form the basis of modern provenance and probability management, subsuming in a large way previous works [58], [52]. 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 [87] [75], [48].

3.1.6. 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 [83], crowdsourcing [50], focused crawling [68], or automatic database tuning [53] critically rely on machine learning techniques, such as classification [72], probabilistic models [67], or reinforcement learning [88].

Machine learning is also a rich source of complex data management problems: thus, the probabilities produced by a conditional random field [78] 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* [85], 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 [84] consider more realistic costs. Also, oracles are usually assumed to be perfect with only a few exceptions [62]. These assumptions usually break when applied to complex data management problems on real-world data, such as crowdsourcing.

Having situated Valda's research area within its broader scientific scope, we now move to the discussion of Valda's application domains.

3.2. Research Directions

We now detail three main research axes within the research agenda of Valda. For each axis, we first mention the leading researcher, and other permanent members involved.

3.2.1. *Foundations of data management (Luc Segoufin; Serge Abiteboul, Camille Bourgaux, Michaël Thomazo, Pierre Senellart).*

Foundations of data management

The systems we are interested in, i.e., for manipulating heterogeneous and confidential data, rapidly changing and massively distributed, are inherently error-prone. The need for formal methods to verify data management systems is best illustrated by the long list of famous leakages of sensitive or personal data that made the front pages of newspapers recently. Moreover, because of the cost in accessing intensional data, it is important to optimize the resources needed for manipulating them.

This creates a need for solid and high-level foundations of DBMS in a manner that is easier to understand, while also facilitating optimization and verification of its critical properties.

In particular these foundations are necessary for various design and reasoning tasks. It allows for clean specifications of key properties of the system such as confidentiality, access control, robustness etc. Once clean specifications are available, it opens the door for formal and runtime verification of the specification. It also permits the design of appropriate query languages – with good expressive power, with limited usage of resources –, the design of good indexes – for optimized evaluation –, and so on. Note that access control policies currently used in database management systems are relatively crude – for example, PostgreSQL offers access control rules on tables, views, or tuples (*row security policies*), but provides no guarantee that these access methods do not contradict each other, or that a user may have access through a query to information that she is not supposed to have access to.

Valda involves leading researchers in the formal verification of data flow in a system manipulating data. Other notable teams involve the WAVE project [2] at U. C. San Diego, and the Business Artifact [3] research program of IBM. One of Valda's objectives is to continue this line of research.

In the short run, we plan to contribute to the state of the art of foundations of systems manipulating data by identifying new scenarios, i.e., specification formalisms, query languages, index structures, query evaluation plans, etc., that allow for any of the tasks mentioned above: formal or runtime verification, optimization etc. Several such scenarios are already known and Valda researchers contributed significantly to their discovery [57], [74], [64], but this research is still in infancy and there is a clear need for more functionalities and more efficiency. This research direction has many facets.

One of the facet is to develop new logical frameworks and new automaton models, with good algorithmic properties (for instance efficient emptiness test, efficient inclusion test and so on), in order to develop a toolbox for reasoning task around systems manipulating data. This toolbox can then be used for higher level tasks such as optimization, verification [57], or query rewriting using views [64].

Another facet is to develop new index structures and new algorithms for efficient query evaluation. For example the enumeration of the output of a query requires the construction of index structures allowing for efficient compressed representation of the output with efficient streaming decompression algorithms as we aim for a constant delay between any two consecutive outputs [82]. We have contributed a lot to this fields by providing several such indexes [74] but there remains a lot to be investigated.

Our medium-term goal is to investigate the borders of feasibility of all the reasoning tasks above. For instance what are the assumptions on data that allow for computable verification 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.

A typical result would be to show that constant delay enumeration of queries is not possible unless the database verify property A and the query property B. Another typical result would be to show that having a robust access control policy verifying at the same time this and that property is not achievable.

[2] http://db.ucsd.edu/WAVE/default.html
[3] http://researcher.watson.ibm.com/researcher/view_group.php?id=2501

Very few such results exist nowadays. If many problems are shown undecidable or decidable, charting the frontier of tractability (say linear time) remains a challenge.

Only when we will have understood the limitation of the method (medium-term goal) 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). This will be our long-term goal.

3.2.2. Uncertainty and provenance of data (Pierre Senellart; Camille Bourgaux, Olivier Cappé, Michaël Thomazo, Luc Segoufin).

Uncertainty and provenance of data

This research axis deals with the modeling and efficient management of data that come with some uncertainty (probabilistic distributions, logical incompleteness, missing values, open-world assumption, etc.) and with provenance information (indicating where the data originates from), as well as with the extraction of uncertainty and provenance annotations from real-world data. Interestingly, the foundations and tools for uncertainty management often rely on provenance annotations. For example, a typical way to compute the probability of query results in probabilistic databases is first to generate the provenance of these query results (in some Boolean framework, e.g., that of Boolean functions or of provenance semirings), and then to compute the probability of the resulting provenance annotation. For this reason, we will deal with uncertainty and provenance in a unified manner.

Valda researchers have carried out seminal work on probabilistic databases [75], [44][7], provenance management [47], incomplete information [46], and uncertainty analysis and propagation in conflicting datasets [65], [41]. These research areas have reached a point where the foundations are well-understood, and where it becomes critical, while continuing developing the theory of uncertain and provenance data management, to move to concrete implementations and applications to real-world use cases.

In the short term, we will focus on implementing techniques from the database theory literature on provenance and uncertainty data management, in the direction of building a full-featured database management add-on that transparently manages provenance and probability annotations for a large class of querying tasks. This work has started recently with the creation of the ProvSQL extension to PostgreSQL, discussed in more details in the following section. To support this development work, we need to resolve the following research question: what representation systems and algorithms to use to support both semiring provenance frameworks [69], extensions to queries with negation [66], aggregation [49], or recursion [80]?

Next, we will study how to add support for incompleteness, probabilities, and provenance annotations in the scenarios identified in the first axis, and how to extract and derive such annotations from real-world datasets and tasks. We will also work on the efficiency of our uncertain data management system, and compare it to other uncertainty management solutions, in the perspective of making it a fully usable system, with little overhead compared to a classical database management system. This requires a careful choice of the provenance representation system used, which should be both compact and amenable to probability computations. We will study practical applications of uncertainty management. As an example, we intend to consider routing in public transport networks, given a probabilistic model on the reliability and schedule uncertainty of different transit routes. The system should be able to provide a user with itinerary to get to have a (probabilistic) guarantee to be at its destination within a given time frame, which may not be the shortest route in the classical sense.

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 [60] on probabilistic query evaluation with the instance-based one of [47], [48].

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.

3.2.3. Personal information management (Serge Abiteboul; Pierre Senellart).

Personal information management

This is a more applied direction of research that will be the context to study issues of interest (see discussion in application domains further).

A typical person today usually has data on several devices and in a number of commercial systems that function as data traps where it is easy to check in information and difficult to remove it or sometimes to simply access it. It is also difficult, sometimes impossible, to control data access by other parties. This situation is unsatisfactory because it requires users to trade privacy against convenience but also, because it limits the value we, as individuals and as a society, can derive from the data. This leads to the concept of Personal Information Management System, in short, a Pims.

A Pims runs, on a user's server, the services selected by the user, storing and processing the user's data. The Pims centralizes the user's personal information. It is a digital home. The Pims is also able to exert control over information that resides in external services (for example, Facebook), and that only gets replicated inside the Pims. See, for instance, [38] for a discussion on the advantages of Pims, as well as issues they raise, e.g. security issues. It is argued there that the main reason for a user to move to Pims is these systems enable great new functionalities.

Valda will study in particular the integration of the user's data. Researchers in the team have already provided important contributions in the context of data integration, notably in the context of the Webdam ERC (2009–2013).

Based on such an integration, Pims can provide a functions, that goes beyond simple query answering:

- Global search over the person's data with a semantic layer using a personal ontology (for example, the data organization the person likes and the person's terminology for data) that helps give meaning to the data;
- Automatic synchronization of data on different devices/systems, and global task sequencing to facilitate interoperating different devices/services;
- Exchange of information and knowledge between "friends" in a truly social way, even if these use different social network platforms, or no platform at all;
- Centralized control point for connected objects, a hub for the Internet of Things; and
- Data analysis/mining over the person's information.

The focus on personal data and these various aspects raise interesting technical challenges that we intend to address.

In the short term, we intend to continue work on the ThymeFlow system to turn it into an easily extendable and deployable platform for the management of personal information – we will in particular encourage students from the M2 *Web Data Management* class taught by Serge and Pierre in the MPRI programme to use this platform in their course projects. The goal is to make it easy to add new functionalities (such as new source *synchronizers* to retrieve data and propagate updates to original data sources, and *enrichers* to add value to existing data) to considerably broaden the scope of the platform and consequently expand its value.

In the medium term, we will continue the work already started that focuses in turning information into knowledge and in knowledge integration. Issues related to intensionality or uncertainty will in particular be considered, relying on the works produced in the other two research axes. We stress, in particular, the importance of minimizing the cost to data access (or, in specific scenarios, the privacy cost associated with obtaining data items) in the context of personal information management: legacy data is often only available through costly APIs, interaction between several Pims may require sharing information within a strict privacy budget, etc. For these reasons, intensionality of data will be a strong focus of the research.

In the long term, we intend to use the knowledge acquired and machine learning techniques to predict the user's behavior and desires, and support new digital assistant functions, providing real *value from data*. We will also look into possibilities for deploying the ThymeFlow platform at a large scale, perhaps in collaboration with industry partners.

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 will use as a privileged application area to validate Valda's results that of personal information management systems (Pims for short) [38].

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 [83].

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 [86] 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 intend to consider other application areas as well. In particular, we have worked in the past and have a strong expertise on Web data [45] in a broad sense: semi-structured, structured, or unstructured content extracted from Web databases [83]; knowledge bases from the Semantic Web [86]; social networks [79]; Web archives and Web crawls [63]; Web applications and deep Web databases [56]; crowdsourcing platforms [50]. We intend to continue using Web data as a natural application domain for the research within Valda when relevant. For instance [54], 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. New Software and Platforms

5.1. 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 where-provenance has been completed. Numerous bug fixes. A docker version has been produced, for ease of installation. Demonstration scenarios are included.

- Participants: Pierre Senellart and Yann Ramusat
- Contact: Pierre Senellart
- Publications: Provenance and Probabilities in Relational Databases: From Theory to Practice - ProvSQL: Provenance and Probability Management in PostgreSQL
- URL: https://github.com/PierreSenellart/provsql

5.2. WAE

Web archive explorer
KEYWORDS: Information extraction - Web archives
FUNCTIONAL DESCRIPTION: The Web archive explorer is a system for extracting, fragmenting and exploring Web archives.

- Contact: Quentin Lobbe
- Publications: Archives, Web fragments and diasporas. For a disaggregated exploration of web archives related to online representations of diasporas - Where the dead blogs are. A Disaggregated Exploration of Web Archives to Reveal Extinct Online Collectives - Revealing Historical Events out of Web Archives
- URL: https://github.com/lobbeque/archive-miner

5.3. 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 FUNCTIONAL DESCRIPTION: Numerous bug fixes and robustness enhancements, link theorems to their repeated Versions, proper management of equations in repeated theorems
NEWS OF THE YEAR: Major 1.1.0 release adding several features (link theorems to their repeated versions, proper management of equations in repeated theorems), beyond this, bug fixes, robustness enhancements, better support for some document classes.

- Participant: Pierre Senellart
- Contact: Pierre Senellart
- URL: https://github.com/PierreSenellart/apxproof

5.4. Sgvizler2

KEYWORDS: SPARQL - Data visualization - JavaScript
FUNCTIONAL DESCRIPTION: This project is the reboot in Typescript of project Sgvizler of Martin G. Skjæveland.

- Partners: LRI - Laboratoire de Recherche en Informatique - BorderCloud
- Contact: Karima Rafes
- URL: https://github.com/BorderCloud/sgvizler2

5.5. SPARQL-PHP

KEYWORDS: SPARQL - PHP
FUNCTIONAL DESCRIPTION: Very simple SPARQL client for PHP.

- Partners: LRI - Laboratoire de Recherche en Informatique - BorderCloud
- Contact: Karima Rafes
- URL: https://github.com/BorderCloud/SPARQL

5.6. TFT

Tester for Triplestore
KEYWORDS: PHP - SPARQL
FUNCTIONAL DESCRIPTION: TFT (Tester for Triplestore) is a script PHP to pass tests through a SPARQL service.

- Partners: LRI - Laboratoire de Recherche en Informatique - BorderCloud
- Contact: Karima Rafes
- URL: https://github.com/BorderCloud/TFT

6. New Results

6.1. Query Enumeration

Query enumeration is the problem of enumerating the results of a query over a database one by one; the goal is to obtain, after some initial low preprocessing time (e.g., linear in the data), one solution after the other with low delay (e.g., constant-time) in between.

In a first work [26], we consider the enumeration of MSO queries over strings under updates. For each MSO query we build an index structure enjoying the following properties: The index structure can be constructed in linear time, it can be updated in logarithmic time and it allows for constant delay time enumeration. This improves from the previous known index structures allowing for constant delay enumeration that would need to be reconstructed from scratch, hence in linear time, in the presence of updates. We allow relabeling updates, insertion of individual labels and removal of individual labels.

In a second work [29], 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 treewidth, 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 $O(n^{1+\varepsilon})$, 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.

6.2. Provenance Circuits

We are interested in obtaining efficiently compact representation of the provenance of a query over a database.

In [28], we generalize three existing graph algorithms to compute the provenance of regular path queries over graph databases, in the framework of provenance semirings – algebraic structures that can capture different forms of provenance. Each algorithm yields a different trade-off between time complexity and generality, as each requires different properties over the semiring. Together, these algorithms cover a large class of semirings used for provenance (top-k, security, etc.). Experimental results suggest these approaches are complementary and practical for various kinds of provenance indications, even on a relatively large transport network.

In [16], we showcase ProvSQL, an open-source module for the PostgreSQL database management system that adds support for computation of provenance and probabilities of query results. A large range of provenance formalisms are supported, including all those captured by provenance semirings, provenance semirings with monus, as well as where-provenance. Probabilistic query evaluation is made possible through the use of knowledge compilation tools, in addition to standard approaches such as enumeration of possible worlds and Monte-Carlo sampling. ProvSQL supports a large subset of non-aggregate SQL queries.

Finally, in [20], [35], we focus on knowledge compilation, which can be used to obtain compact circuit-based representations of (Boolean) provenance. Some width parameters of the circuit, such as bounded treewidth or pathwidth, can be leveraged to convert the circuit to structured classes, e.g., deterministic structured NNFs (d-SDNNFs) or OBDDs. We show how to connect the width of circuits to the size of their structured representation, through upper and lower bounds. For the upper bound, we show how bounded-treewidth circuits can be converted to a d-SDNNF, in time linear in the circuit size. Our bound, unlike existing results, is constructive and only singly exponential in the treewidth. We show a related lower bound on monotone DNF or CNF formulas, assuming a constant bound on the arity (size of clauses) and degree (number of occurrences of each variable). Specifically, any d-SDNNF (resp., SDNNF) for such a DNF (resp., CNF) must be of exponential size in its treewidth; and the same holds for pathwidth when compiling to OBDDs. Our lower bounds, in contrast with most previous work, apply to any formula of this class, not just a well-chosen family. Hence, for our language of DNF and CNF, pathwidth and treewidth respectively characterize the efficiency of compiling to OBDDs and (d-)SDNNFs, that is, compilation is singly exponential in the width parameter.

6.3. Exploiting Content from the Web

One of our main domain of application is that of Web content. We investigate methods to acquire and exploit content from the Web.

In [30], we analyze form-based websites to discover sequences of actions and values that result in a valid form submission. Rather than looking at the text or DOM structure of the form, our method is driven by solving constraints involving the underlying client-side JavaScript code. In order to deal with the complexity of client-side code, we adapt a method from program analysis and testing, concolic testing, which mixes concrete code execution, symbolic code tracing, and constraint solving to find values that lead to new code paths. While concolic testing is commonly used for detecting bugs in stand-alone code with developer support, we show how it can be applied to the very different problem of filling Web forms. We evaluate our system on a benchmark of both real and synthetic Web forms.

In [21], we investigate *focused crawling*: collecting as many Web pages relevant to a target topic as possible while avoiding irrelevant pages, reflecting limited resources available to a Web crawler. We improve on the efficiency of focused crawling by proposing an approach based on reinforcement learning. Our algorithm evaluates hyperlinks most profitable to follow over the long run, and selects the most promising link based on this estimation. To properly model the crawling environment as a Markov decision process, we propose new representations of states and actions considering both content information and the link structure. The size of the state-action space is reduced by a generalization process. Based on this generalization, we use a linear-function approximation to update value functions. We investigate the trade-off between synchronous and asynchronous methods. In experiments, we compare the performance of a crawling task with and without learning; crawlers based on reinforcement learning show better performance for various target topics.

Finally, in [23], [24] we propose a framework to follow the dynamics of vanished Web communities, based on the exploration of corpora of Web archives. To achieve this goal, we define a new unit of analysis called Web fragment: a semantic and syntactic subset of a given Web page, designed to increase historical accuracy. This contribution has practical value for those who conduct large-scale archive exploration (in terms of time range and volume) or are interested in computational approaches to Web history and social science.

6.4. Knowledge Bases

Knowledge bases are collection of semantic facts (typically of the form subject–predicate–object) along with possible logical rules (e.g., in the form of existential rules) that apply to these facts. We investigate querying, data integration, and inference in such knowledge bases.

In [27], we focus on autocompletion of SPARQL queries over knowledge bases. We analyze several autocompletion features proposed by the main editors, highlighting the needs currently not taken into account while met by a user community we work with, scientists. Second, we introduce the first (to our knowledge) autocompletion approach able to consider snippets (fragments of SPARQL query) based on queries expressed by previous users, enriching the user experience. Third, we introduce a usable, open and concrete solution able to consider a large panel of SPARQL autocompletion features that we have implemented in an editor. Last but not least, we demonstrate the interest of our approach on real biomedical queries involving services offered by the Wikidata collaborative knowledge base.

In [25], we introduce a novel open-source framework for integrating the data of a user from different sources into a single knowledge base. Our framework integrates data of different kinds into a coherent whole, starting with email messages, calendar, contacts, and location history. We show how event periods in the user's location data can be detected and how they can be aligned with events from the calendar. This allows users to query their personal information within and across different dimensions, and to perform analytics over their emails, events, and locations. Our system models data using RDF, extending the schema.org vocabulary and providing a SPARQL interface.

Finally, in [22], [32], we view knowledge bases as composed of an instance that contains incomplete data and a set of existential rules, and investigate ontology-based query answering: answers to queries are logically entailed from the knowledge base. This brings to light the fundamental chase tool, and its different variants that have been proposed in the literature. It is well-known that the problem of determining, given a chase variant and a set of existential rules, whether the chase will halt on a given instance / on any instance, is undecidable. Hence, a crucial issue is whether it becomes decidable for known subclasses of existential rules. We consider

linear existential rules, a simple yet important subclass of existential rules. We study the decidability of the associated chase termination problem for different chase variants, with a novel approach based on a single graph and a single notion of forbidden pattern. Besides the theoretical interest of a unified approach, an original result is the decidability of the restricted chase termination for linear existential rules.

6.5. Transparency and Bias

In this last set of results, we investigate transparency and bias in data management.

Bias in online information has recently become a pressing issue, with search engines, social networks and recommendation services being accused of exhibiting some form of bias. In [15], we make the case for a systematic approach towards measuring bias. To this end, we discuss formal measures for quantifying the various types of bias, we outline the system components necessary for realizing them, and we highlight the related research challenges and open problems.

In [19], we pursue an investigation of data-driven collaborative work-flows. In the model, peers can access and update local data, causing side-effects on other peers' data. In this paper, we study means of explaining to a peer her local view of a global run, both at runtime and statically. We consider the notion of "scenario for a given peer" that is a subrun observationally equivalent to the original run for that peer. Because such a scenario can sometimes differ significantly from what happens in the actual run, thus providing a misleading explanation , we introduce and study a faithfulness requirement that ensures closer adherence to the global run. We show that there is a unique minimal faithful scenario, that explains what is happening in the global run by extracting only the portion relevant to the peer. With regard to static explanations, we consider the problem of synthesizing, for each peer, a "view program" whose runs generate exactly the peer's observations of the global runs. Assuming some conditions desirable in their own right, namely transparency and boundedness, we show that such a view program exists and can be synthesized. As an added benefit, the view program rules provide provenance information for the updates observed by the peer.

Finally, in two articles oriented towards applications and policy, we discuss bias and neutrality and their impact on regulation. In [18] we discuss the different forms of neutrality in the digital world, from the neutrality of networks to neutrality of content. In [17], we investigate the impact of bias and neutrality concerns on algorithms used by businesses.

7. Partnerships and Cooperations

7.1. Regional Initiatives

Michaël Thomazo has obtained a 6k€ budget from the Île-de-France region (DIM RFSI – *Réseau Francilien en Sciences Informatiques*) entitled *ISORE: Indexation sémantique d'ontologies, le cas des règles existentielles.* The grant was awarded when Michaël Thomazo was part of the Inria Saclay Cedar team, but the budget was transferred to the Valda team.

7.2. National Initiatives

7.2.1. ANR

Valda has been part of two ANR projects in 2018:

- HEADWORK (budget managed by Inria), together with IRISA (Druid, coordinator), Inria Lille (Links & Spirals), and Inria Rennes (Sumo), and two application partners: MNHN (Cesco) and FouleFactory. The topic is workflows for crowdsourcing. See http://headwork.gforge.inria.fr/.
- BioQOP (budget managed by ENS), with Idemia (coordinator) and GREYC, on the optimization of queries for privacy-aware biometric data management. See http://bioqop.di.ens.fr/.

In addition, two ANR projects were accepted in 2018 and will start early 2019:

- CQFD (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 (budget managed by Inria), LIGM (coordinator), IRIF, and LaBRI, on incomplete and inconsistent data.

7.3. International Initiatives

7.3.1. IIL projects

Valda has strong collaborations with the following international groups:

Univ. Edinburgh, United Kingdom: Peter Buneman and Leonid Libkin

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

TU Dresden, Germany: Markus Krötzsch and Sebastian Rudolph

Dortmund University, Germany: Thomas Schwentick

Warsaw University, Poland: Mikołaj Bojańczyk and Szymon Toruńczyk

Tel Aviv University, Israel: Daniel Deutch and Tova Milo

Drexel University, USA: Julia Stoyanovich

Univ. California San Diego, USA: Victor Vianu

National University of Singapore: Stéphane Bressan

7.4. International Research Visitors

7.4.1. Visits of International Scientists

Victor Vianu, Professor at UC San Diego and holder of an Inria international chair, spent 3 months within Valda, employed as an ENS invited professor.

7.4.2. Visits to International Teams

7.4.2.1. Research Stays Abroad

- Michaël Thomazo and Pierre Senellart have spent respectively two weeks and one week at TU Dresden, collaborating with Markus Krötzsch and Sebastian Rudolph.
- Pierre Senellart has spent a cumulated time of around three weeks at National University of Singapore, co-advising Debabrota Basu, PhD student working under the co-supervision of Stéphane Bressan, visiting Stéphane Bressan and other researchers at NUS, and participating in the French–Singapore workshop on AI, where Olivier Cappé represented CNRS.

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific Events Organisation

8.1.1.1. General Chair, Scientific Chair

- Luc Segoufin, chair of the steering committee of the conference series *Highlights of Logic, Games and Automata*

8.1.1.2. Member of the Organizing Committees

- Luc Segoufin and Pierre Senellart, co-organizers of École de Printemps en Informatique Théorique (EPIT) 2019
- Pierre Senellart, co-organizer of ACM-ICPC Southwestern Europe 2018 competition

8.1.2. Scientific Events Selection

8.1.2.1. Chair of Conference Program Committees

- Pierre Senellart, RoD (Reasoning on Data) workshop at The Web Conference 2018 (co-chair)

8.1.2.2. Member of the Conference Program Committees

- Camille Bourgaux, AAAI 2019
- Olivier Cappé, COLT 2018, ALT 2019
- Pierre Senellart, BDA 2018, PODS 2019
- Michaël Thomazo, IJCAI 2018, AAAI 2019

8.1.3. Journal

8.1.3.1. Member of the Editorial Boards

- Olivier Cappé, associate editor, *Annals of the Institute of Statistical Mathematics*

8.1.3.2. Reviewer - Reviewing Activities

- Pierre Senellart, *Transactions on Database Systems*, *VLDB Journal*

8.1.4. Invited Talks

- Pierre Senellart, keynote at Theory and Practice of Provenance (TaPP), London, United Kingdom
- Pierre Senellart, keynote at TempWeb workshop, The Web Conference, Lyon, France
- Pierre Senellart, LORIA Colloquium, Nancy, France

8.1.5. Leadership within the Scientific Community

- Serge Abiteboul is a member of the French Academy of Sciences, of the Academia Europa, and of the scientific council of the Société Informatique de France.
- Pierre Senellart is a member of the steering committee of BDA, the French scientific community on data management.

8.1.6. Scientific Expertise

- Pierre Senellart, FWO

8.1.7. Research Administration

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

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Licence: Pierre Senellart, *Databases*, 32 heqTD, L3, École normale supérieure
Licence: Pierre Senellart, *Algorithms*, 18 heqTD, L3, École normale supérieure
Licence: Michaël Thomazo, *Formal languages*, 22 heqTD, L3, Télécom ParisTech
Master: Serge Abiteboul & Pierre Senellart, *Web data management*, 36 heqTD, M2, MPRI

Pierre Senellart has various teaching responsibilities (L3 internships, M2 administration, tutoring, entrance competition) at ENS.

8.2.2. Supervision

PhD: Miyoung Han, *Reinforcement learning approaches in dynamic environments*, Télécom ParisTech, 19 July 2018, Pierre Senellart

PhD: Alexandre Vigny, *Query enumeration and nowhere dense graphs*, Université Paris-Diderot, 27 September 2018, Arnaud Durand & Luc Segoufin

PhD: Mikaël Monet, *Combined complexity of probabilistic query evaluation*, Université Paris-Saclay, 12 October 2018, Antoine Amarilli & Pierre Senellart

PhD: Quentin Lobbé, *Archives, fragments Web et diasporas. Pour une exploration désagrégée de corpus d'archives Web liées aux représentations en ligne des diasporas.* Université Paris-Saclay, 9 November 2018, Dana Diminescu & Pierre Senellart

PhD in progress: Julien Grange, *Graph properties: order and arithmetic in predicate logics*, started in September 2017, Luc Segoufin

PhD in progress: Karima Rafes, *Le Linked Data à l'université : la plateforme LinkedWiki*, defense planned in January 2019, Serge Abiteboul & Sarah Cohen-Boulakia

PhD in progress: Yann Ramusat, *Provenance-based routing in probabilistic graphs*, started in September 2018, Silviu Maniu & Pierre Senellart

PhD in progess: Yoan Russac, *Sequential methods for robust decision making*, started in December 2018, Olivier Cappé

8.2.3. Juries

- PhD Frederik Harwarth, June 2018, Humbold University, Luc Segoufin (reviewer)
- PhD Debabrota Basu, October 2018, National University of Singapore, Olivier Cappé (reviewer)
- PhD Govind, December 2018, Université Caen–Normandie, Pierre Senellart (president)
- PhD Ngurah Agus Sanjaya Er, December 2018, Télécom ParisTech, Pierre Senellart (president)

8.3. Popularization

8.3.1. Internal or external Inria responsibilities

Serge Abiteboul is the president of the strategic committee of the Blaise Pascal foundation for scientific mediation.

8.3.2. Articles and contents

Serge Abiteboul published *Le bot qui murmurait à l'oreille de la vieille dame* at the *Le Pommier* éditions, a collection of short stories on the digital world, accompanied with scientific and technical discussions.

Serge Abiteboul writes regular columns on popularization of computer science in La Recherche and Le Monde (Économie).

8.3.3. Education

Pierre Senellart participated to a week-long meeting of teachers in *classes préparatoires* in Lumini in May 2018 to discuss the future of computer science education and to give an introduction to database research.

9. Bibliography

Major publications by the team in recent years

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

[2] A. AMARILLI, M. MONET, P. SENELLART. *Connecting Width and Structure in Knowledge Compilation*, in "ICDT 2018 - 21st International Conference on Database Theory", Vienna, Austria, Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, March 2018, vol. 98, pp. 1-17 [*DOI :* 10.4230/LIPICS.ICDT.2018.6], https://hal.inria.fr/hal-01851564

[3] C. BOURGAUX, A. TURHAN. *Temporal Query Answering in DL-Lite over Inconsistent Data*, in "The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I", 2017, pp. 121–137, https://doi.org/10.1007/978-3-319-68288-4_8

[4] F. JACQUEMARD, L. SEGOUFIN, J. DIMINO. *FO2(<, +1, ~) on data trees, data tree automata and branching vector addition systems*, in "Logical Methods in Computer Science", 2016, vol. 12, n° 2, https://doi.org/10.2168/LMCS-12(2:3)2016

[5] P. LAGRÉE, O. CAPPÉ, B. CAUTIS, S. MANIU. *Effective Large-Scale Online Influence Maximization*, in "2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017", 2017, pp. 937–942, https://doi.org/10.1109/ICDM.2017.118

[6] M. LECLÈRE, M.-L. MUGNIER, M. THOMAZO, F. ULLIANA. *A Single Approach to Decide Chase Termination on Linear Existential Rules*, in "DL 2018 - Description Logics", Tempe, United States, October 2018, https://arxiv.org/abs/1810.02132 , https://hal-lirmm.ccsd.cnrs.fr/lirmm-01892353

[7] S. MANIU, R. CHENG, P. SENELLART. *An Indexing Framework for Queries on Probabilistic Graphs*, in "ACM Trans. Datab. Syst", 2017, https://hal.inria.fr/hal-01437580

[8] D. MONTOYA, S. ABITEBOUL, P. SENELLART. *Hup-me: inferring and reconciling a timeline of user activity from rich smartphone data*, in "Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Bellevue, WA, USA, November 3-6, 2015", 2015, pp. 62:1–62:4, http://doi.acm.org/10.1145/2820783.2820852

[9] N. SCHWEIKARDT, L. SEGOUFIN, A. VIGNY. *Enumeration for FO Queries over Nowhere Dense Graphs*, in "PODS 2018 - Principles Of Database Systems", Houston, United States, June 2018, https://hal.inria.fr/hal-01895786

[10] P. SENELLART, L. JACHIET, S. MANIU, Y. RAMUSAT. *ProvSQL: Provenance and Probability Management in PostgreSQL*, in "Proceedings of the VLDB Endowment (PVLDB)", August 2018, vol. 11, n° 12, pp. 2034-2037 [*DOI :* 10.14778/3229863.3236253], https://hal.inria.fr/hal-01851538

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] M. HAN. *Reinforcement Learning Approaches in Dynamic Environments*, Télécom ParisTech, July 2018, https://hal.inria.fr/tel-01891805

[12] Q. LOBBÉ. *Archives, Web fragments and diasporas. For a disaggregated exploration of web archives related to online representations of diasporas*, Université Paris-Saclay, November 2018, https://hal.inria.fr/tel-01963548

[13] M. MONET. *Combined Complexity of Probabilistic Query Evaluation*, Université Paris-Saclay, October 2018, https://hal.inria.fr/tel-01963559

[14] A. VIGNY. *Query enumeration and nowhere dense graphs*, Université Paris-Diderot, September 2018, https://hal.inria.fr/tel-01963540

Articles in International Peer-Reviewed Journals

[15] E. PITOURA, P. TSAPARAS, G. FLOURIS, I. FUNDULAKI, P. PAPADAKOS, S. ABITEBOUL, G. WEIKUM. *On Measuring Bias in Online Information*, in "SIGMOD record", 2018, pp. 1-6, https://hal.inria.fr/hal-01638069

[16] P. SENELLART, L. JACHIET, S. MANIU, Y. RAMUSAT. *ProvSQL: Provenance and Probability Management in PostgreSQL*, in "Proceedings of the VLDB Endowment (PVLDB)", August 2018, vol. 11, n^{o} 12, pp. 2034-2037 [*DOI :* 10.14778/3229863.3236253], https://hal.inria.fr/hal-01851538

Articles in National Peer-Reviewed Journals

[17] S. ABITEBOUL. *Les algorithmes du commerce*, in "Concurrences - revue des droits de la concurrence", 2018, https://hal.inria.fr/hal-01744283

[18] S. ABITEBOUL. *Les déclinaisons de la neutralité*, in "ANNALES DES MINES", December 2018, https://hal.inria.fr/hal-01963510

International Conferences with Proceedings

[19] S. ABITEBOUL, P. BOURHIS, V. VIANU. *Explanations and Transparency in Collaborative Workflows*, in "PODS 2018 - 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles Of Database Systems", Houston, Texas, United States, June 2018, https://hal.inria.fr/hal-01744978

[20] A. AMARILLI, M. MONET, P. SENELLART. *Connecting Width and Structure in Knowledge Compilation*, in "ICDT 2018 - 21st International Conference on Database Theory", Vienna, Austria, Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, March 2018, vol. 98, pp. 1-17 [*DOI :* 10.4230/LIPICS.ICDT.2018.6], https://hal.inria.fr/hal-01851564

[21] M. HAN, P.-H. WUILLEMIN, P. SENELLART. *Focused Crawling through Reinforcement Learning*, in "18th International Conference on Web Engineering (ICWE 2018)", Cáceres, Spain, T. MIKKONE, R. KLAMMA, J. HERNÁNDEZ (editors), Lecture Notes in Computer Science, Springer, June 2018, vol. 10845, pp. 261-278 [*DOI :* 10.1007/978-3-319-91662-0_20], https://hal.inria.fr/hal-01851547

[22] M. LECLÈRE, M.-L. MUGNIER, M. THOMAZO, F. ULLIANA. *A Single Approach to Decide Chase Termination on Linear Existential Rules*, in "DL 2018 - Description Logics", Tempe, United States, October 2018, https://arxiv.org/abs/1810.02132 , https://hal-lirmm.ccsd.cnrs.fr/lirmm-01892353

[23] Q. LOBBÉ. *Revealing Historical Events out of Web Archives*, in "22nd International Conference on Theory and Practice of Digital Libraries (TPDL 2018)", Porto, Portugal, September 2018, https://hal.archives-ouvertes.fr/hal-01895951

[24] Q. LOBBÉ. *Where the dead blogs are. A Disaggregated Exploration of Web Archives to Reveal Extinct Online Collectives*, in "ICADL 2018 - 20th International Conference on Asia-Pacific Digital Libraries", Hamilton, New Zealand, November 2018, pp. 1-12, https://hal.archives-ouvertes.fr/hal-01895955

[25] D. MONTOYA, T. P. TANON, S. ABITEBOUL, P. SENELLART, F. M. SUCHANEK. *A Knowledge Base for Personal Information Management*, in "LDOW2018 - Linked Open Data Workshop at the World Wide Web Conference", Lyon, France, April 2018, https://hal-imt.archives-ouvertes.fr/hal-01719312

[26] M. NIEWERTH, L. SEGOUFIN. *Enumeration of MSO Queries on Strings with Constant Delay and Logarithmic Updates*, in "Principles of Databse Systems, PODS'18", Houston, United States, ACM Press, June 2018 [*DOI :* 10.1145/3196959.3196961], https://hal.inria.fr/hal-01895796

[27] K. RAFES, S. ABITEBOUL, S. COHEN-BOULAKIA, B. RANCE. *Designing scientific SPARQL queries using autocompletion by snippets*, in "14th IEEE International Conference on eScience", Amsterdam, Netherlands, October 2018, https://hal.archives-ouvertes.fr/hal-01874780

[28] Y. RAMUSAT, S. MANIU, P. SENELLART. *Semiring Provenance over Graph Databases*, in "10th USENIX Workshop on the Theory and Practice of Provenance (TaPP 2018)", London, United Kingdom, July 2018, https://hal.inria.fr/hal-01850510

[29] N. SCHWEIKARDT, L. SEGOUFIN, A. VIGNY. *Enumeration for FO Queries over Nowhere Dense Graphs*, in "PODS 2018 - Principles Of Database Systems", Houston, United States, June 2018, https://hal.inria.fr/hal-01895786

[30] B. SPENCER, M. BENEDIKT, P. SENELLART. *Form Filling based on Constraint Solving*, in "18th International Conference on Web Engineering (ICWE 2018)", Cáceres, Spain, T. MIKKONE, R. KLAMMA, J. HERNÁNDEZ (editors), LNCS - Lecture Notes in Computer Science, Springer, June 2018, vol. 10845 [*DOI :* 10.1007/978-3-319-91662-0_20], https://hal.inria.fr/hal-01851555

Scientific Books (or Scientific Book chapters)

[31] S. ABITEBOUL. *The Digital Shoebox*, in "Memory, edited by Philippe Tortell, Mark Turin, and Margot Young", UBC Press, October 2018, https://hal.inria.fr/hal-01875161

Research Reports

[32] M. LECLÈRE, M.-L. MUGNIER, M. THOMAZO, F. ULLIANA. *A Single Approach to Decide Chase Termination on Linear Existential Rules*, arXiv:1810.02132, October 2018, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01892375

Other Publications

[33] A. AMARILLI, M. L. BA, D. DEUTCH, P. SENELLART. *Computing Possible and Certain Answers over Order-Incomplete Data*, October 2018, https://arxiv.org/abs/1801.06396 - 63 pages, 48 references. Submitted. Extended journal version of arXiv:1707.07222, https://hal.inria.fr/hal-01891811

[34] A. AMARILLI, P. BOURHIS, M. MONET, P. SENELLART. *Evaluating Datalog via Tree Automata and Cycluits*, October 2018, https://arxiv.org/abs/1808.04663 - 53 pages, 61 references. Journal version of "Combined Tractability of Query Evaluation via Tree Automata and Cycluits (Extended Version)" at arXiv:1612.04203. To appear in Theory of Computing Systems, https://hal.inria.fr/hal-01891814

[35] A. AMARILLI, M. MONET, P. SENELLART. *Connecting Width and Structure in Knowledge Compilation (Extended Version)*, May 2018, https://arxiv.org/abs/1709.06188 - 33 pages, no figures, 40 references. This is the full version with proofs of the corresponding ICDT'18 publication, and it integrates all reviewer feedback.

Except for the additional appendices, and except for formatting differences and inessential changes, the contents are the same as in the conference version [*DOI :* 10.4230/LIPIcs.ICDT.2018.6], https://hal.inria.fr/hal-01614551

[36] D. Basu, P. Senellart, S. Bressan. *BelMan: Bayesian Bandits on the Belief–Reward Manifold*, October 2018, https://arxiv.org/abs/1805.01627 - working paper or preprint, https://hal.inria.fr/hal-01891813

[37] W. Kazana, L. Segoufin. *First-order queries on classes of structures with bounded expansion*, February 2018, working paper or preprint, https://hal.inria.fr/hal-01706665

References in notes

[38] S. Abiteboul, B. André, D. Kaplan. *Managing your digital life*, in "Commun. ACM", 2015, vol. 58, n^{o} 5, pp. 32–35, http://doi.acm.org/10.1145/2670528

[39] S. Abiteboul, P. Bourhis, V. Vianu. *Comparing workflow specification languages: A matter of views*, in "ACM Trans. Database Syst.", 2012, vol. 37, n^{o} 2, pp. 10:1–10:59, http://doi.acm.org/10.1145/2188349.2188352

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

[41] S. Abiteboul, D. Deutch, V. Vianu. *Deduction with Contradictions in Datalog*, in "Proc. 17th International Conference on Database Theory (ICDT), Athens, Greece, March 24-28, 2014", N. Schweikardt, V. Christophides, V. Leroy (editors), OpenProceedings.org, 2014, pp. 143–154, https://doi.org/10.5441/002/icdt.2014.17

[42] S. Abiteboul, L. Herr, 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", R. Hull (editor), ACM Press, 1996, pp. 49–57, http://doi.acm.org/10.1145/237661.237674

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

[44] S. Abiteboul, B. Kimelfeld, Y. Sagiv, P. Senellart. *On the expressiveness of probabilistic XML models*, in "VLDB J.", 2009, vol. 18, n^{o} 5, pp. 1041–1064, https://doi.org/10.1007/s00778-009-0146-1

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

[46] S. Abiteboul, L. Segoufin, V. Vianu. *Representing and querying XML with incomplete information*, in "ACM Trans. Database Syst.", 2006, vol. 31, n^{o} 1, pp. 208–254, http://doi.acm.org/10.1145/1132863.1132869

[47] A. Amarilli, P. Bourhis, 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, https://doi.org/10.1007/978-3-662-47666-6_5

[48] A. AMARILLI, P. BOURHIS, 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", T. MILO, W. TAN (editors), ACM, 2016, pp. 355–370, http://doi.acm.org/10.1145/2902251.2902301

[49] Y. AMSTERDAMER, D. DEUTCH, V. TANNEN. *Provenance for aggregate queries*, in "Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2011, June 12-16, 2011, Athens, Greece", M. LENZERINI, T. SCHWENTICK (editors), ACM, 2011, pp. 153–164, http://doi.acm.org/10.1145/1989284.1989302

[50] Y. AMSTERDAMER, Y. GROSSMAN, T. MILO, P. SENELLART. *CrowdMiner: Mining association rules from the crowd*, in "PVLDB", 2013, vol. 6, n^{o} 12, pp. 1250–1253, http://www.vldb.org/pvldb/vol6/p1250-amsterdamer.pdf

[51] 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", R. HULL, W. FAN (editors), ACM, 2013, pp. 175–188, http://doi.acm.org/10.1145/2463664.2465216

[52] D. BARBARÁ, H. GARCIA-MOLINA, D. PORTER. *The Management of Probabilistic Data*, in "IEEE Trans. Knowl. Data Eng.", 1992, vol. 4, n^{o} 5, pp. 487–502, https://doi.org/10.1109/69.166990

[53] D. BASU, Q. LIN, W. CHEN, H. T. VO, Z. YUAN, P. SENELLART, S. BRESSAN. *Regularized Cost-Model Oblivious Database Tuning with Reinforcement Learning*, in "T. Large-Scale Data- and Knowledge-Centered Systems", 2016, vol. 28, pp. 96–132, https://doi.org/10.1007/978-3-662-53455-7_5

[54] M. BENEDIKT, G. GOTTLOB, 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", M. LENZERINI, T. SCHWENTICK (editors), ACM, 2011, pp. 211–222, http://doi.acm.org/10.1145/1989284.1989309

[55] M. BENEDIKT, P. SENELLART. *Databases*, in "Computer Science, The Hardware, Software and Heart of It", Springer, 2011, pp. 169–229, https://doi.org/10.1007/978-1-4614-1168-0_10

[56] M. BIENVENU, D. DEUTCH, D. MARTINENGHI, P. SENELLART, 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", M. BRAMBILLA, S. CERI, T. FURCHE, G. GOTTLOB (editors), CEUR Workshop Proceedings, CEUR-WS.org, 2012, vol. 884, pp. 21–24, http://ceur-ws.org/Vol-884/VLDS2012_p21_Bienvenu.pdf

[57] M. BOJAŃCZYK, L. SEGOUFIN, 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", R. HULL, W. FAN (editors), ACM, 2013, pp. 63–74, http://doi.acm.org/10.1145/2463664.2465228

[58] P. BUNEMAN, S. KHANNA, 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", J. VAN DEN BUSSCHE, V. VIANU (editors), Lecture Notes in Computer Science, Springer, 2001, vol. 1973, pp. 316–330, https://doi.org/10.1007/3-540-44503-X_20

[59] B. COURCELLE. *The Monadic Second-Order Logic of Graphs. I. Recognizable Sets of Finite Graphs*, in "Inf. Comput.", 1990, vol. 85, n° 1, pp. 12–75, https://doi.org/10.1016/0890-5401(90)90043-H

[60] N. N. DALVI, D. SUCIU. *The dichotomy of probabilistic inference for unions of conjunctive queries*, in "J. ACM", 2012, vol. 59, n° 6, pp. 30:1–30:87, http://doi.acm.org/10.1145/2395116.2395119

[61] A. DESHPANDE, Z. G. IVES, V. RAMAN. *Adaptive Query Processing*, in "Foundations and Trends in Databases", 2007, vol. 1, n° 1, pp. 1–140, https://doi.org/10.1561/1900000001

[62] P. DONMEZ, 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", J. G. SHANAHAN, S. AMER-YAHIA, I. MANOLESCU, Y. ZHANG, D. A. EVANS, A. KOLCZ, K. CHOI, A. CHOWDHURY (editors), ACM, 2008, pp. 619–628, http://doi.acm.org/10.1145/1458082.1458165

[63] M. FAHEEM, 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", R. B. ALLEN, J. HUNTER, M. L. ZENG (editors), Lecture Notes in Computer Science, Springer, 2015, vol. 9469, pp. 39–51, https://doi.org/10.1007/978-3-319-27974-9_5

[64] N. FRANCIS, L. SEGOUFIN, C. SIRANGELO. *Datalog Rewritings of Regular Path Queries using Views*, in "Logical Methods in Computer Science", 2015, vol. 11, n° 4, https://doi.org/10.2168/LMCS-11(4:14)2015

[65] A. GALLAND, S. ABITEBOUL, A. MARIAN, P. SENELLART. *Corroborating information from disagreeing views*, in "Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010", B. D. DAVISON, T. SUEL, N. CRASWELL, B. LIU (editors), ACM, 2010, pp. 131–140, http://doi.acm.org/10.1145/1718487.1718504

[66] F. GEERTS, A. POGGI. *On database query languages for K-relations*, in "J. Applied Logic", 2010, vol. 8, n° 2, pp. 173–185, https://doi.org/10.1016/j.jal.2009.09.001

[67] L. GETOOR. *Introduction to statistical relational learning*, MIT Press, 2007

[68] G. GOURITEN, S. MANIU, 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", L. FERRES, G. ROSSI, V. A. F. ALMEIDA, E. HERDER (editors), ACM, 2014, pp. 35–45, http://doi.acm.org/10.1145/2631775.2631795

[69] T. J. GREEN, G. KARVOUNARAKIS, 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", L. LIBKIN (editor), ACM, 2007, pp. 31–40, http://doi.acm.org/10.1145/1265530.1265535

[70] T. J. GREEN, V. TANNEN. *Models for Incomplete and Probabilistic Information*, in "IEEE Data Eng. Bull.", 2006, vol. 29, n° 1, pp. 17–24, http://sites.computer.org/debull/A06mar/green.ps

[71] A. Y. HALEVY. *Answering queries using views: A survey*, in "VLDB J.", 2001, vol. 10, n° 4, pp. 270–294, https://doi.org/10.1007/s007780100054

[72] M. A. HEARST, S. T. DUMAIS, E. OSUNA, J. PLATT, B. SCHOLKOPF. *Support vector machines*, in "IEEE Intelligent Systems", 1998, vol. 13, n^{o} 4, pp. 18–28, https://doi.org/10.1109/5254.708428

[73] T. IMIELINSKI, W. LIPSKI JR.. *Incomplete Information in Relational Databases*, in "J. ACM", 1984, vol. 31, n^{o} 4, pp. 761–791, http://doi.acm.org/10.1145/1634.1886

[74] W. KAZANA, L. SEGOUFIN. *Enumeration of first-order queries on classes of structures with bounded expansion*, 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", R. HULL, W. FAN (editors), ACM, 2013, pp. 297–308, http://doi.acm.org/10.1145/2463664.2463667

[75] B. KIMELFELD, P. SENELLART. *Probabilistic XML: Models and Complexity*, in "Advances in Probabilistic Databases for Uncertain Information Management", Z. MA, L. YAN (editors), Studies in Fuzziness and Soft Computing, Springer, 2013, vol. 304, pp. 39–66, https://doi.org/10.1007/978-3-642-37509-5_3

[76] A. C. KLUG. *Equivalence of Relational Algebra and Relational Calculus Query Languages Having Aggregate Functions*, in "J. ACM", 1982, vol. 29, n^{o} 3, pp. 699–717, http://doi.acm.org/10.1145/322326.322332

[77] D. KOSSMANN. *The State of the art in distributed query processing*, in "ACM Comput. Surv.", 2000, vol. 32, n^{o} 4, pp. 422–469, http://doi.acm.org/10.1145/371578.371598

[78] J. D. LAFFERTY, A. MCCALLUM, 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", C. E. BRODLEY, A. P. DANYLUK (editors), Morgan Kaufmann, 2001, pp. 282–289

[79] S. LEI, S. MANIU, L. MO, R. CHENG, 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, http://doi.acm.org/10.1145/2783258.2783271

[80] M. MOHRI. *Semiring Frameworks and Algorithms for Shortest-Distance Problems*, in "Journal of Automata, Languages and Combinatorics", 2002, vol. 7, n^{o} 3, pp. 321–350

[81] F. NEVEN. *Automata Theory for XML Researchers*, in "SIGMOD Record", 2002, vol. 31, n^{o} 3, pp. 39–46, http://doi.acm.org/10.1145/601858.601869

[82] L. SEGOUFIN. *A glimpse on constant delay enumeration (Invited Talk)*, in "31st International Symposium on Theoretical Aspects of Computer Science (STACS 2014), STACS 2014, March 5-8, 2014, Lyon, France", E. W. MAYR, N. PORTIER (editors), LIPIcs, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2014, vol. 25, pp. 13–27, https://doi.org/10.4230/LIPIcs.STACS.2014.13

[83] P. SENELLART, A. MITTAL, D. MUSCHICK, R. GILLERON, 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", C. Y. CHAN, N. POLYZOTIS (editors), ACM, 2008, pp. 9–16, http://doi.acm.org/10.1145/1458502.1458505

[84] B. SETTLES, M. CRAVEN, L. FRIEDLAND. *Active learning with real annotation costs*, in "NIPS 2008 Workshop on Cost-Sensitive Learning", 2008, http://burrsettles.com/pub/settles.nips08ws.pdf

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

[86] F. M. SUCHANEK, S. ABITEBOUL, P. SENELLART. *PARIS: Probabilistic Alignment of Relations, Instances, and Schema*, in "PVLDB", 2011, vol. 5, n^{o} 3, pp. 157–168, http://www.vldb.org/pvldb/vol5/p157_fabianmsuchanek_vldb2012.pdf

[87] D. SUCIU, D. OLTEANU, C. RÉ, C. KOCH. *Probabilistic Databases*, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2011, https://doi.org/10.2200/S00362ED1V01Y201105DTM016

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

[89] 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", H. R. LEWIS, B. B. SIMONS, W. A. BURKHARD, L. H. LANDWEBER (editors), ACM, 1982, pp. 137–146, http://doi.acm.org/10.1145/800070.802186

[90] K. ZHOU, M. LALMAS, T. SAKAI, R. CUMMINS, 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", Q. HE, A. IYENGAR, W. NEJDL, J. PEI, R. RASTOGI (editors), ACM, 2013, pp. 689–698, http://doi.acm.org/10.1145/2505515.2505691

[91] M. T. ÖZSU, P. VALDURIEZ. *Principles of Distributed Database Systems, Third Edition*, Springer, 2011, https://doi.org/10.1007/978-1-4419-8834-8