Activity Report 2017

Team Valda

Value from Data

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

RESEARCH CENTER
Paris

THEME
Data and Knowledge Representation and Processing

Table of contents

Team Valda

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

VALDA has integrated members of the Inria DAHU project-team in 2017. Their relevant activity has been integrated in this activity report.

Keywords:

Computer Science and Digital Science:

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
A3.1.9. - Database
A3.2.2. - Knowledge extraction, cleaning
A3.2.3. - Inference
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
A9.1. - Knowledge

Other Research Topics and Application Domains:

B6.3.1. - Web
B6.3.4. - Social Networks
B6.5. - Information systems
B9.5.5. - Sociology
B9.5.10. - Digital humanities
B9.7.2. - Open data
B9.8. - Privacy
B9.10. - Ethics

1. Personnel

Research Scientists

Serge Abiteboul [Inria, Senior Researcher, HDR]
Luc Segoufin [Inria, Senior Researcher, from Sep 2017, HDR]

Faculty Member

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

PhD Students

Julien Grange [École normale supérieure, from Sep 2017]
Miyoung Han [Institut Telecom ex GET Groupe des Ecoles des Télécommunications]
Quentin Lobbe [Institut Telecom ex GET Groupe des Ecoles des Télécommunications]

Mikael Monet [Institut Telecom ex GET Groupe des Ecoles des Télécommunications]
David Montoya [ENGIE, until March 2017]
Karima Rafes [BorderCloud]
Alexandre Vigny [Univ Denis Diderot, from Sep 2017]
Su Yang [Ecole Nationale Supérieure des Mines de Paris, until Apr 2017]

Intern

Yann Ramusat [Inria, from Mar 2017 until Jul 2017]

Administrative Assistants

Linday Polienor [until June 2017]
Sandrine Vergès [from September 2017]

Visiting Scientist

Victor Vianu [UCSD & École normale supérieure, from Jul 2017]

External Collaborator

Yann Ramusat [Student at École normale supérieure on a long-term project, from Sep 2017]

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 [2], [5] 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 [33][3], and possibly distributed [78] 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.

[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*.

- *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* [59] or *distributed query optimization* [65].

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 [50] or aggregated search [77]. 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 [75] (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.

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 [2]. 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 [5]. 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 [76] 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 [35], queries, expressed in temporal logics, can often by compiled to automata; in graph databases [41], 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 [68], or for more complex languages to data tree automata[7]. Another reason of the importance of automata theory, and tree automata in particular, comes from Courcelle's results [48] 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* [2]. 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 [46].

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 [32].

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 [2]. 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* [61] 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 [57], and to a probabilistic formalism by assigning independent probabilities to the Boolean events [58]. These two extensions form the

basis of modern provenance and probability management, subsuming in a large way previous works [47], [42]. 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 [74] [63], [38].

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 [70], crowdsourcing [40], focused crawling [56], or automatic database tuning [43] critically rely on machine learning techniques, such as classification [60], probabilistic models [55], or reinforcement learning [75].

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 [71] consider more realistic costs. Also, oracles are usually assumed to be perfect with only a few exceptions [51]. 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, 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.

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

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 [46], [62],[6], 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 [46], or query rewriting using views [6].

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 [69]. We have contributed a lot to this fields by providing several such indexes [62] 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.

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; 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 [63], [36][12], provenance management [4], incomplete information [37], and uncertainty analysis and propagation in conflicting datasets [53], [34]. 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 [57], extensions to queries with negation [54], aggregation [39], or recursion [67]?

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 [49] on probabilistic query evaluation with the instance-based one of [4] [38].

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, [1] 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) [1].

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 [70].

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 [73] 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 [3] in a broad sense: semi-structured, structured, or unstructured content extracted from Web databases [70]; knowledge bases from the Semantic Web [73]; social networks [9]; Web archives and Web crawls [52]; Web applications and deep Web databases [45]; crowdsourcing platforms [40]. We intend to continue using Web data as a natural application domain for the research within Valda when relevant. For instance [44], 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: ProvSQL becomes usable for a large range of queries. Support for semirings and m-semirings is present, support for probability computation has been added through a variety of techniques, including knowledge compilation, support for where-provenance is currently being implemented.

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

5.2. Thymeflow

KEYWORD: Personal information
FUNCTIONAL DESCRIPTION: ThymeFlow allows in particular the development of plugins for both interacting with existing Web sources and presenting users with rich interfaces and query facilities over their personal information. A preliminary version of ThymeFlow tools has also been deployed on the Cozy Cloud personal cloud system. The model allows the open-source community to contribute individual plugins while we focus on providing users with useful ways to exploit their personal information.
NEWS OF THE YEAR: Minor maintenance.

- Participants: David Montoya, Pierre Senellart, Serge Abiteboul and Su Yang
- Partner: ENGIE
- Contact: Pierre Senellart
- Publication: Personal Knowledge Base Systems
- URL: https://github.com/thymeflow/thymeflow/

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: Ability to specify a sectioning counter, Compilation fix of proofs-ketch in inline mode
NEWS OF THE YEAR: Overall software maintenance. Support for more document classes. Some new features.

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

6. New Results

6.1. Enumeration of Query Results

In many applications the output of a query may have a huge size and computing all the answers may already consume too many of the allowed resources. In this case it may be appropriate to first output a small subset of the answers and then, on demand, output a subsequent small numbers of answers and so on until all possible answers have been exhausted. To make this even more attractive it is preferable to be able to minimize the time necessary to output the first answers and, from a given set of answers, also minimize the time necessary to output the next set of answers - this second time interval is known as the *delay*. We have shown that this was doable with a almost linear preprocessing time and constant enumeration delay for first-order queries over structures having local bounded expansion [22].

6.2. Ethical Data Management

Issues of responsible data analysis and use are coming to the forefront of the discourse in data science research and practice [14]. The research has been focused on analyzing the fairness, accountability and transparency (FAT) properties of specific algorithms and their outputs. Although these issues are most apparent in the social sciences where fairness is interpreted in terms of the distribution of resources across protected groups, management of bias in source data affects a variety of fields. Consider climate change studies that require representative data from geographically diverse regions, or supply chain analyses that require data that represents the diversity of products and customers. In a paper [23], we argue that FAT properties must be considered as database system issues, further upstream in the data science lifecycle: bias in source data goes unnoticed, and bias may be introduced during pre-processing (fairness), spurious correlations lead to reproducibility problems (accountability), and assumptions made during pre-processing have invisible but significant effects on decisions (transparency). As machine learning methods continue to be applied broadly by non-experts, the potential for misuse increases. There is a need for a data sharing and collaborative analytics platform with features to encourage (and in some cases, enforce) best practices at all stages of the data science lifecycle. We describe features of such a platform, which we term Fides, in the context of of urban analytics, outlining a systems research agenda in responsible data science.

6.3. Structure and Tractability of Uncertain Data

A major part of the work conducted in Valda has been to study the connections between tractability and structure in databases, in particular uncertain databases.

In a first line of work, we have investigated incompleteness related to order. In [18], we have introduced a query language for order-incomplete data, based on the positive relational algebra with order-aware accumulation. We have used partial orders to represent order-incomplete data, and studied possible and certain answers for queries in this context, showing these problems are respectively NP-complete and coNP-complete, but identifying tractable cases depending on query operators and the structure of input partial orders. In [16], we consider a different setting where some partial order is known, but actual values are unknown. Our work is the first to propose a principled scheme to derive the value distributions and expected values of unknown items in this setting, with the goal of computing estimated top-k results by interpolating the unknown values from the known ones. We have studied the complexity of this general task, and show tight complexity bounds, proving that the problem is intractable, but can be tractably approximated. We have also isolated structure-based restrictions that allow for a PTIME solution.

In [17], we have investigated parameterizations of both database instances and queries that make query evaluation fixed-parameter tractable in combined complexity, first in a setting without uncertainty. For this, we have introduced a new Datalog fragment with stratified negation, intensional-clique-guarded Datalog (ICG-Datalog), with linear-time evaluation on structures of bounded treewidth for programs of bounded rule size. Our result is shown by compiling to alternating two-way automata, whose semantics is defined via cyclic provenance circuits (cycluits) that can be tractably evaluated. Finally, we move to the probabilistic setting and have shown that probabilistic query evaluation remains intractable in combined complexity under this parameterization.

Finally, a last line of work concerns efficient queries over probabilistic graphs. In a first theoretical work [19], we have studied the combined complexity of conjunctive query evaluation on probabilistic graphs, which can be alternatively phrased as a probabilistic version of the graph homomorphism problem. We have shown that the complexity landscape is surprisingly rich, using a variety of technical tools. In a more practical work [12], we have proposed indexing techniques and algorithms to evaluate source-to-target queries in probabilistic graphs, by exploiting their structure. We have shown that these significantly enhance the accuracy and efficiency of existing query evaluation approaches on probabilistic graphs.

7. Partnerships and Cooperations

7.1. Regional Initiatives

Valda has obtained a 10k€ budget from ENS in 2017, as a start-up grant from the team (*Action Concertée Incitative*).

Inria established a bilateral contract with the Centre – Val de Loire region, for the expertise and audit of a research project by Pierre Senellart. Because of delays due to the company being audited, the expertise is still in progress.

7.2. National Initiatives

7.2.1. ANR

Valda has been part of one ANR project in 2017 (Headwork, budget managed by Inria), together with IRISA (DRUID team, coordinator), Inria Lille (LINKS & SPIRAL), and Inria Rennes (SUMO), and two application partners: MNHN (Cesco) and FouleFactory. The topic is workflows for crowdsourcing. See http://headwork.gforge.inria.fr/.

In addition, another project (BioQOP, budget managed by ENS) will start in January 2018, with Morpho and GREYC, on the optimization of queries for privacy-aware biometric data management

7.3. International Initiatives

7.3.1. Informal International Partners

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, and Georg Gottlob

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 6 months within Valda: three months employed by Inria and three months as an ENS invited professor.

7.4.1.1. Internships

Deabrota Basu, PhD student at National University of Singapore, stayed 2.5 months within Valda, to work with Pierre Senellart.

7.4.2. Visits to International Teams

7.4.2.1. Research Stays Abroad

- Pierre Senellart has spent around two months at the University of Edinburgh, collaborating with Peter Buneman and Leonid Libkin.
- Pierre Senellart has spent a cumulated time of more than one month at National University of Singapore, co-advising Debabrota Basu, PhD student working under the co-supervision of Stéphane Bressan.

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific Events Organisation

8.1.1.1. Member of the Organizing Committees

- Serge Abiteboul, organization of Personal Analytics & Privacy workshop, joint with ECML-PKDD 2017, Skopje, Macedonia
- Serge Abiteboul, scientific organization of colloquium on *La communauté scientifique face au renseignement*, École militaire, Parsi, France
- Serge Abiteboul, organization of colloquium on *Les enjeux scientifiques de l'éthique du numérique*, Académie des sciences
- Pierre Senellart, organization of ParisBD 2017, Télécom ParisTech, Paris, France
- Pierre Senellart, co-organizer of ACM-ICPC Southwestern Europe 2017 competition

8.1.2. Scientific Events Selection

8.1.2.1. Chair of Conference Program Committees

- Pierre Senellart, BDA 2017 (French conference on data management)
- Pierre Senellart, WebDB workshop, joint with SIGMOD 2017

8.1.2.2. Member of the Conference Program Committees

- Pierre Senellart, *Gems of PODS 2017* committee
- Pierre Senellart, SIGMOD 2017 (*distinguished PC member*), ICDT 2017, EDBT 2018

8.1.3. Journal

8.1.3.1. Reviewer - Reviewing Activities

- Pierre Senellart, *Journal of the ACM*, *VLDB Journal*, *Artificial Intelligence*

8.1.4. Invited Talks

- Serge Abiteboul, keynote at PPDP-LOPSTR, Namur, Belgium
- Serge Abiteboul, keynote at ETAPS, Uppsala, Sweden
- Serge Abiteboul, keynote at Law & Big Data Conference, Paris, 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.

8.1.6. Scientific Expertise

- Pierre Senellart, ANR, NSF

8.1.7. Research Administration

- Serge Abiteboul was the president of the Dune jury (*Développement d'universités numériques expérimentales*)
- Serge Abiteboul participated in the NCU jury (*nouveaux cursus à l'université*)
- Serge Abiteboul contributed to the report on *Éthique de la recherche en apprentissage machine* of Cerna-Allistene
- Serge Abiteboul is co-chair of the "Committee on Gender Equality and Equal Opportunities" of Inria.
- 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 vice-director of the DI ENS laboratory, joint between ENS, CNRS, and Inria.

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Licence: Pierre Senellart, *Databases*, 54 heqTD, L3, École normale supérieure

Licence: Pierre Senellart, *Algorithms*, 18 heqTD, L3, École normale supérieure

Master: Serge Abiteboul & Pierre Senellart, *Web data management*, 36 heqTD, M2, MPRI

Master: Luc Segoufin, *Logic, descriptive complexity and database theory*, 36 heqTD, M2, MPRI

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

Serge Abiteboul proposed with Benjamin Nguyen and Philippe Rigaux a second session of the Mooc "Bases de données relationnelles: comprendre pour maîtriser" (FUN). He proposed with Julia Stoyanovich a course on "Ethical data management" at the EDBT Summer School, Genova, 2017.

8.2.2. Supervision

PhD : David Montoya, *Une base de connaissance personnelle intégrant les données d'un utilisateur et une chronologie de ses activités*, Université Paris-Saclay, 6 March 2017, Serge Abiteboul & Pierre Senellart

PhD in progress: Debabrota Basu, *Reinforcement learning applications to data management problems*, started in 2015, Stéphane Bressan & Pierre Senellart

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

PhD in progress: Miyoung Han, *Learning approaches to dynamic data management*, started in 2015, Pierre Senellart

PhD in progress: Quentin Lobbé, *Diachronic analysis of diaspora communities through web archives enrichment*, started in 2015, Pierre Senellart & Dana Diminescu

PhD in progress: Mikaël Monet, *Efficient querying of large uncertain graphs by exploiting their structure*, started in 2015, Pierre Senellart & Antoine Amarilli

PhD in progress: Karima Rafes, *Security and management of personal data in the Web of things*, started in 2015, Serge Abiteboul & Sarah Cohen-Boulakia

PhD in progress: Alexandre Vigny, *Query enumeration on nowhere-dense graphs*, started in 2015, Luc Segoufin & Arnaud Durand

8.2.3. Juries

- PhD Paul Lagrée, October 2017, Université Paris-Saclay, Pierre Senellart

8.3. Popularization

Serge Abiteboul is involved in several popular science activities. He founded and animates the blog http://binaire.blog.lemonde.fr/ on computer science. He was the scientific curator (*commissaire scientifique*) of the exhibition "Terra Data" at the Cité des Sciences. He published two scientific popularization books in 2017, "Le temps des algorithmes" [26], with Gilles Dowek, and "Terra data" [27], with Valéeie Peugeot.

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

9. Bibliography

Major publications by the team in recent years

[1] 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

[2] S. ABITEBOUL, R. HULL, V. VIANU. *Foundations of Databases*, Addison-Wesley, 1995, http://webdam.inria.fr/Alice/

[3] S. ABITEBOUL, I. MANOLESCU, P. RIGAUX, M. ROUSSET, P. SENELLART. *Web Data Management*, Cambridge University Press, 2011, http://webdam.inria.fr/Jorge

[4] 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

[5] 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

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

[7] 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^{o} 2, https://doi.org/10.2168/LMCS-12(2:3)2016

[8] W. KAZANA, L. SEGOUFIN. *Enumeration of monadic second-order queries on trees*, in "ACM Trans. Comput. Log.", 2013, vol. 14, n^{o} 4, pp. 25:1–25:12, http://doi.acm.org/10.1145/2528928

[9] 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

[10] 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

Publications of the year

Articles in International Peer-Reviewed Journals

[11] D. FIGUEIRA, L. SEGOUFIN. *Bottom-up automata on data trees and vertical XPath*, in "Logical Methods in Computer Science", 2017, pp. 1-40, https://arxiv.org/abs/1710.08748 [*DOI :* 10.08748], https://hal.inria.fr/hal-01631219

[12] 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

[13] P. SENELLART. *Provenance and Probabilities in Relational Databases: From Theory to Practice*, in "SIGMOD record", December 2017, pp. 1-11, https://hal.inria.fr/hal-01672566

Invited Conferences

[14] S. ABITEBOUL. *Issues in Ethical Data Management - Extended Abstract*, in "PPDP 2017 - 19th International Symposium on Principles and Practice of Declarative Programming", Namur, Belgium, October 2017, https://hal.inria.fr/hal-01621687

[15] S. ABITEBOUL, D. MONTOYA. *Personal Knowledge Base Systems*, in "PAP 2017, Personal analytics and privacy", Skopje, Macedonia, September 2017, https://hal.inria.fr/hal-01592601

International Conferences with Proceedings

[16] A. AMARILLI, Y. AMSTERDAMER, T. MILO, P. SENELLART. *Top-k Querying of Unknown Values under Order Constraints*, in "ICDT 2017 - International Conference on Database Theory", Venice, Italy, March 2017 [*DOI :* 10.4230/LIPICS.ICDT.2017.5], https://hal.inria.fr/hal-01439295

[17] A. AMARILLI, P. BOURHIS, M. MONET, P. SENELLART. *Combined Tractability of Query Evaluation via Tree Automata and Cycluits*, in "ICDT 2017 - International Conference on Database Theory", Venice, Italy, March 2017 [*DOI :* 10.4230/LIPICS.ICDT.2017.6], https://hal.inria.fr/hal-01439294

[18] A. AMARILLI, M. LAMINE BA, D. DEUTCH, P. SENELLART. *Possible and Certain Answers for Queries over Order-Incomplete Data*, in "24th International Symposium on Temporal Representation and Reasoning (TIME 2017)", Mons, Belgium, S. SCHEWE, T. SCHNEIDER, J. WIJSEN (editors), Schloss Dagstuhl, October 2017, vol. 90, pp. 4:1-4:19, https://arxiv.org/abs/1707.07222 [*DOI :* 10.4230/LIPICS.TIME.2017.4], https://hal.inria.fr/hal-01570603

[19] A. AMARILLI, M. MONET, P. SENELLART. *Conjunctive Queries on Probabilistic Graphs: Combined Complexity*, in "Principles of Database Systems (PODS)", Chicago, United States, May 2017, https://arxiv.org/abs/1703.03201 [*DOI :* 10.1145/3034786.3056121], https://hal.inria.fr/hal-01486634

[20] M. CROCHEMORE, A. HELIOU, G. KUCHEROV, L. MOUCHARD, S. P. PISSIS, Y. RAMUSAT. *Minimal absent words in a sliding window & applications to on-line pattern matching*, in "FCT 2017", Bordeaux,

France, Lecture Notes in Computer Science, Springer, September 2017, forthcoming, https://hal.archives-ouvertes.fr/hal-01569264

[21] O. SAVKOVIĆ, E. KHARLAMOV, W. NUTT, P. SENELLART. *Towards Approximating Incomplete Queries over Partially Complete Databases (Extended Abstract)*, in "AMW", Montevideo, Uruguay, AMW 2017 - 11th Alberto Mendelzon International Workshop on Foundations of Data Management Montevideo, Uruguay June 5 – 9, 2017, June 2017, https://hal.inria.fr/hal-01586884

[22] L. SEGOUFIN, A. VIGNY. *Constant Delay Enumeration for FO Queries over Databases with Local Bounded Expansion*, in "ICDT", Venise, Italy, March 2017, https://hal.inria.fr/hal-01589303

[23] J. STOYANOVICH, B. HOWE, S. ABITEBOUL, G. MIKLAU, A. SAHUGUET, G. WEIKUM. *Fides: Towards a Platform for Responsible Data Science*, in "SSDBM'17 - 29th International Conference on Scientific and Statistical Database Management", Chicago, United States, June 2017 [*DOI :* 10.1145/3085504.3085530], https://hal.inria.fr/hal-01522418

National Conferences with Proceedings

[24] K. RAFES, S. COHEN-BOULAKIA, S. ABITEBOUL. *Une autocomplétion générique de SPARQL dans un contexte multi-services*, in "BDA 2017 - 33ème conférence sur la «Gestion de Données — Principes, Technologies et Applications»", Nancy, France, November 2017, https://hal.inria.fr/hal-01627760

Books or Proceedings Editing

[25] A. MELIOU, P. SENELLART (editors). *Proceedings of the 20th International Workshop on the Web and Databases, WebDB 2017*, May 2017, https://hal.inria.fr/hal-01523772

Scientific Popularization

[26] S. ABITEBOUL, G. DOWEK. *Le temps des algorithmes*, Editions Le Pommier, 2017, 192 p. , https://hal.inria.fr/hal-01502505

[27] S. ABITEBOUL, V. PEUGEOT. *Terra Data : Qu'allons-nous faire des données numériques ?*, Editions Le Pommier, 2017, 320 p. , https://hal.inria.fr/hal-01502512

[28] P. SENELLART. *Archivage du Web*, in "Les Big Data à découvert", CNRS Éditions, March 2017, https://hal.inria.fr/hal-01497800

Other Publications

[29] A. AMARILLI, Y. AMSTERDAMER, T. MILO, P. SENELLART. *Top-k Querying of Unknown Values under Order Constraints (Extended Version)*, January 2017, https://arxiv.org/abs/1701.02634 - 32 pages, 1 figure, 1 algorithm, 51 references. Extended version of paper at ICDT'17, https://hal.inria.fr/hal-01439310

[30] A. AMARILLI, M. L. BA, D. DEUTCH, P. SENELLART. *Possible and Certain Answers for Queries over Order-Incomplete Data*, October 2017, https://arxiv.org/abs/1707.07222 - 55 pages, 5 figures, 1 table, 44 references. Accepted at TIME'17. This paper is the full version with appendices of the article in the TIME proceedings. The main text of this full version is the same as the TIME proceedings version, except some superficial changes (to fit the proceedings version to 15 pages, and to obey LIPIcs-specific formatting requirements) [*DOI :* 10.4230/LIPICS.TIME.2017.4], https://hal.inria.fr/hal-01614571

[31] P. SENELLART, A. AMARILLI, M. MONET. *Connecting Width and Structure in Knowledge Compilation*, October 2017, https://arxiv.org/abs/1709.06188 - 32 pages, no figures, 39 references. Submitted, https://hal.inria.fr/hal-01614551

References in notes

[32] S. ABITEBOUL, P. BOURHIS, V. VIANU. *Comparing workflow specification languages: A matter of views*, in "ACM Trans. Database Syst.", 2012, vol. 37, n° 2, pp. 10:1–10:59, http://doi.acm.org/10.1145/2188349.2188352

[33] S. ABITEBOUL, P. BUNEMAN, D. SUCIU. *Data on the Web: From Relations to Semistructured Data and XML*, Morgan Kaufmann, 1999

[34] 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

[35] S. ABITEBOUL, L. HERR, J. V. 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

[36] S. ABITEBOUL, B. KIMELFELD, Y. SAGIV, P. SENELLART. *On the expressiveness of probabilistic XML models*, in "VLDB J.", 2009, vol. 18, n° 5, pp. 1041–1064, https://doi.org/10.1007/s00778-009-0146-1

[37] S. ABITEBOUL, L. SEGOUFIN, V. VIANU. *Representing and querying XML with incomplete information*, in "ACM Trans. Database Syst.", 2006, vol. 31, n° 1, pp. 208–254, http://doi.acm.org/10.1145/1132863.1132869

[38] 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

[39] 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

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

[41] 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

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

[43] 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

[44] 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

[45] 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

[46] 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

[47] 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. V. 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

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

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

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

[51] 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

[52] 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

[53] 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

[54] 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

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

[56] 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

[57] 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

[58] 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

[59] 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

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

[61] T. IMIELINSKI, W. L. JR.. *Incomplete Information in Relational Databases*, in "J. ACM", 1984, vol. 31, n° 4, pp. 761–791, http://doi.acm.org/10.1145/1634.1886

[62] 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

[63] 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

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

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

[66] 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

[67] 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

[68] 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

[69] 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

[70] 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

[71] 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

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

[73] 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

[74] 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

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

[76] 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

[77] 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

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