Rapport d'évaluation à mi-parcours

Uncertainty over Structured and Intensional Data

Antoine Amarilli

sous la supervision de Pierre Senellart, Télécom ParisTech, Institut Mines–Télécom

General summary

I am doing a PhD in theoretical computer science, in the DBWeb team of the INFRES department, in the Télécom ParisTech engineering school. I have started in September 2013, after completing my studies at the École normale supérieure (ENS) in Paris, and am funded for a duration of three years by an Allocation spécifique normalien (grant for ENS students from the French Ministry of Higher Education and Research). My supervisor is Prof. Pierre Senellart.

This midterm report is a summary of the current state of my PhD research. I first present my research topic (Section 1) with a general description, envisioned use cases, and an abstract overview of the proposed solution. I pursue with a summary of my research results so far and my ongoing collaborations (Section 2), on various independent lines of work which focus on more specific aspects of the overall picture. I conclude by a summary of my other commitments (Section 3).

1 Research topic

The overall field of my PhD research is the theory of data management systems, or *databases*. Its overall goal is to evaluate the interaction of three aspects of data which are not adequately managed by state-of-the-art systems:

Uncertainty. Data is often noisy, biased, or incomplete, whether it was initially erroneous or has progressively become stale. In particular, data is often produced by automated systems that can make mistakes: e.g., it may be automatically extracted from Web sources, or obtained by parsing natural language.

The uncertainty on data cannot be managed as an afterthought: it must be modeled and maintained throughout the whole computation to estimate in a principled manner our overall confidence in final query answers. Yet, it is challenging to represent incomplete or probabilistic data in a way that is compact (i.e., does not explicitly enumerate all probabilities), compositional (i.e., the result of operations on data can still be represented in the same framework) and yet operational (i.e., query results can be computed efficiently on the representation).

Intensionality. We cannot realistically assume that all of the data resides on the machine where the computation is performed. We often need data which is only available on remote services (e.g., from the Web) and must be accessed through targeted queries, because bulk downloading of the data would not be practical. Some data may also be available only by extracting it from the real world, e.g., by asking questions to humans. Last, part of the data may not be materialized, but implicit: for instance, we may want to reason about the implied consequences of the existing data under logical rules. We call this the *intensionality* of data: as opposed to extensional data which is explicitly given and can be accessed directly, intensional data is not immediately present and must be used sparingly.

Intensionality must be taken into account when evaluating a query, by deciding which accesses are likely to be the most relevant for that query, and estimating their cost, to achieve the

best trade-off. This requires us to design a dynamic query evaluation plan, making decisions at every step depending on the results that we have obtained so far.

Structure. Data often has a very heterogeneous structure. Web pages are XML or HTML trees; the links between Web pages have a graph structure when we crawl them; the relational tables managed by standard database systems are yet another way to represent data.

We aim at developing an approach that can be applied to different sources with diverse representations, taking into account their individual specificities. In other words, our approach should not be limited to a single model such as relational data or XML data.

My PhD research deals with the foundational aspects of data management following these three axes: while practical implementations or experiments are not ruled out, I have no immediate plans to work on them. My current focus is on the many fundamental questions which arise in the theoretical design of a system that would manage data according to these dimensions.

I now give examples of practical scenarios where our approach would be needed, and then present the general design that we envision for systems that could perform such tasks.

1.1 Example applications

Structured extraction on news stories. Consider the problem of harvesting a structured collection of facts about an event or topic of interest from Web sources, to answer a specific complex query. For instance, I am interested in the global surveillance disclosures about the NSA, and want a timeline of the announced leaks, or a map of the route followed by Edward Snowden when fleeing the USA. Or I am following the debate about the recently proposed anti-terrorism law in France, and I want to know the affiliations of the politicians who support or oppose it.

The information required to answer the query must be extracted from a wide collection of sources: news articles in an open-ended collection of websites, but also structured open data (to know, e.g., which politician is affiliated to which party) or semantic Web sources such as Wikidata (Vrandecic and Krötzsch, 2014) for background information (e.g., what are the geographical coordinates of cities). The *structures* of these sources are heterogeneous: the news articles that would be crawled on the Web are structured like a graph, but their content is text; open data would typically be relational tables or spreadsheets, and the semantic Web facts are RDF triples.

Of course, we cannot crawl all news articles and siphon all sources, and need an *intensional* approach: locate promising articles, e.g., using keywords, and then run more costly natural language processing on them (e.g., named entity recognition, triple extraction, sentiment analysis, etc.); as for the semantic Web and open data sources, we should retrieve selectively the facts that we need from them. Of course, *uncertainty* also comes into play, because the output of the extraction may be noisy: we need to remember which facts are more likely, depending on whether the extractor was confident or not, and whether the fact occurred in many sources or in just one source. We can then use background constraints to prune the facts (e.g., Snowden can only be in one place at a time), relying on our estimation of their individual likelihood.

If the uncertainty is too high, we can also use other mechanisms, such as human annotators or crowd workers, to clean the data (e.g., confirm that a fact was correctly extracted from a sentence, or that a name was correctly disambiguated). However, this is costly in terms of money and latency, so it should be used sparingly. Alternatively, we can just crawl more articles and extract more information to try to confirm or refute dubious claims. Our goal is to achieve a good compromise between our confidence in the final answer, and the cost paid to obtain it.

Personal information management systems. Personal data is usually stored in many separate places: email services, calendar providers, contacts in an address book or on a social network, etc.,

all of which have their own *structure*. Imagine now that we want to evaluate a user query that involves all of this data: "find out who I need to warn about my upcoming trips".

The orchestration of such a query is complex: we need to find when the upcoming trips are, using, e.g., emails from travel agencies; find the conflicting meetings in the calendar; determine who takes part to the meetings, which can be explicitly indicated or may need to be implicitly inferred (based on email exchanges, or deductions guided by the past attendees to similar meetings); last, find contact information for all the people involved, using, e.g., an address book.

Now, if the data is stored on remote servers, we cannot download all of it to evaluate the query. Rather, we need to retrieve *intensionally* the information that seems to be relevant (e.g., emails that look like an airline ticket) and process it. Of course, as the extraction will again be noisy, *uncertainty* needs to be managed as well.

1.2 Overall approach

From the example applications, one can imagine the general design of a system designed to answer complex queries on intensional and uncertain data of heterogeneous structure:

1. The system has a representation of its current knowledge about the world, which accommodates the structure of all sources, and represents our uncertainty on possible world states.

2. It also has a representation of all sources, and of the accesses that can be performed on them, including information about their cost, expected results, access limitations, and the logical constraints that hold about what they can return.

3. At every step, the system decides interactively which access to perform to make progress towards the user goal. It attempts to minimize the cost, under multiple dimensions: time, monetary cost, bandwidth, usage for rate limitations policies, etc.

4. Once an access has been performed, the system revises its knowledge of the world based on the new information, computing a new probabilistic representation from its past knowledge and from the observed evidence.

5. The process continues until the system has sufficient confidence about how to answer the user query. It then computes its final answer from the current knowledge, and includes probabilistic annotations that indicate its degree of certainty.

My advisor and I have elaborated in more detail our overall vision for such a system, in a yet-unpublished paper (Amarilli and Senellart, 2014a). We have also written a tutorial proposal about the various lines of related work on this topic (Amarilli and Senellart, 2014b), which we have submitted to the EDBT/ICDT 2015 conference.

2 Ongoing research

As my PhD topic is very broad, most of my research effort so far has been to study different subproblems of the general topic, under various angles. I now review the results that I have obtained, and my ongoing collaborations.

2.1 Tractability for probabilistic data and rules

The most important research effort of the first half of my PhD has been a study of conditions that can be imposed on incomplete and probabilistic data to ensure that query evaluation is computationally tractable. This work was started in January 2014 with my advisor and Pierre Bourhis

(CNRS LIFL), and has been conducted since then, including two four-day visits to Lille in July and August 2014 to work with Pierre Bourhis, and a one-month visit to Singapore in September–October 2014 to work with my advisor.

General presentation. In general, it is known that evaluating simple queries on probabilistic data with simple correlations is generally intractable (#P-hard) in the input database, which is an important practical obstacle to probabilistic data management. Our idea is that all may not be lost, because the input databases are not arbitrary in practice, so it could be interesting to identify ways to restrict the input data and make the problem tractable. For instance, existing results for non-probabilistic query evaluation and counting (Courcelle, 1990; Arnborg et al., 1991) show that such problems become tractable if the input data is assumed to be close to a tree (formally, if its treewidth is bounded by a constant).

Our work is to develop a general model for probabilistic data, and a notion of tree-likeness for this model; we then show that query evaluation is tractable in the input database instance if the treewidth is assumed to be bounded. Our result covers many expressive query languages that satisfy a general condition of being rewritable to tree automata for fixed treewidth. It also covers multiple existing probabilistic models for both XML and relational data, so it is not tied to one specific structure. This genericity of the model is reminiscent of problems that my advisor and I studied before my PhD (Amarilli and Senellart, 2013), about ways to connect tractability results for the relational and XML probabilistic formalisms.

Our general proof approach is as follows. Similarly to existing work, we decompose bounded-treewidth data into a binary tree representation on a finite alphabet, called a tree encoding, and compile the queries to bottom-up tree automata on such representations. The main technical innovation is a natural construction that instruments tree automata runs on an uncertain tree encoding obtained from the probabilistic instance (describing all the tree encodings of its possible worlds): we compile the automaton to a circuit and stitch it to a circuit representation of the probabilistic annotations of the instance, yielding a circuit lineage for the query result. We show that known message-passing algorithms for probability computation can then be run effectively on the circuit, because its treewidth is bounded.

Our result also has interesting connections to semiring provenance (for monotone queries and absorptive semirings), although this is less directly relevant to my PhD topic.

Publication. I have informally presented our results as a lightning talk at the AMW School in June 2014 and at the Highlights 2014 workshop in September. We have submitted them in October as a conference paper to PODS 2015 which is currently under review (Amarilli et al., 2014d). I plan to give a more complete presentation of the results in an upcoming visit in Lille.

Ongoing work. We intend to continue investigating these topics. Our main goal is the question of reasoning under probabilistic rules, which was in fact our initial motivation for this line of work.

A possible application for uncertain rule reasoning is the recent result by Luís Galárraga (DBWeb) et al. (Galárraga et al., 2013) about mining probabilistic trends in knowledge bases. It would probably be very useful to extrapolate such rules to deduce statistically likely facts which are missing from the data. However, it would not be feasible to materialize all of those consequences, so we would want to consider them *intensionally*, when answering a specific query. Yet, as the rules are probabilistic, this is a complicated task, which may be intractable.

We have formally defined a semantics for this problem, for expressive rule languages that may assert the existence of new objects (so that the implied instance may be of unbounded size). We also obtained results about the tractability of reasoning under such rules, by reducing to our previous work, if the rule language is restricted. We plan to write this up and publish it, either as a separate

work or as part of an extend journal version of our conference submission (Amarilli et al., 2014d), depending on its fate.

We would also like to evaluate the practical applicability of our work. In this context, in addition to the possible connection to (Galárraga et al., 2013), I have been involved in discussions with Michael Benedikt (University of Oxford), Silviu Maniu (Hong Kong University), and my advisor, about developing practical techniques to evaluate queries on probabilistic graphs (intended as a follow-up to (Maniu et al., 2014)). My advisor and I also intend to submit an offer to the Master parisien de recherche en informatique (MPRI) for a Master's internship in 2015, about a practical implementation of our results: this would lead to interesting theoretical questions about reducing the size of the circuits for practical query instrumentation.

2.2 Open-world query answering

During a five-month internship in 2013 before the beginning of my PhD, I have been visiting the University of Oxford and worked with Michael Benedikt about the problem of open-world query answering. I am still involved in this research.

General presentation. We consider a relational database instance, and we see it according to the open-world semantics: the true state of the world is uncertain, but includes at least the facts of the instance. In addition, we have some background logical constraints which we know about the world, expressed in languages either inspired by traditional data management (e.g., functional dependencies, inclusion dependencies) or by mathematical logic (e.g., the guarded two-variable fragment with counting quantifiers (Pratt-Hartmann, 2009)); so we are reasoning about the *intensional* consequences of those rules. We want to evaluate a query on the unknown world according to the certain answers semantics: we want to know if it necessarily holds (i.e., is implied by the instance and constraints), or if there is some possible state of the world in which it does not hold. We call this the open-world query answering problem.

The problem is of course undecidable if the constraints are purely arbitrary, as it is at least as hard as deciding the satisfiability of the constraints. Yet, more surprisingly, this problem is often undecidable, even for comparatively inexpressive logic fragments, as soon as we have number restriction constraints, such as functional dependencies or equality-generating dependencies (Calì et al., 2003): intuitively, number restrictions limit the number of elements that may be related to another element. Our first contribution is to show the decidability of a new class of constraints with number restrictions that intuitively combines two settings that were previously separate: expressive two-variable logics (Pratt-Hartmann, 2009), and weaker integrity constraints on the arbitrary-arity predicates. This construction proceeds by a rewriting to arity-two and an unraveling argument.

Another surprising thing is that known open-world query answering techniques that can deal with number restrictions usually do not apply if we impose that the world is finite. This reasonable assumption seems mild, but it makes an important difference, because it prohibits infinite chains of facts and can force them to "loop back", which is not innocuous because of number restrictions. Existing work on finite query answering under number restrictions has either focused only on the arity-two case (Pratt-Hartmann, 2009; Ibáñez-García et al., 2014), which is unsatisfactory for traditional database management, or studied cases where the interaction between number restrictions and other constraints was severely limited (Rosati, 2006), and one could prove that the finiteness assumption did not change query answers.

Our second, deeper contribution, is to show that the finite open-world query answering can be decided outside of this context, namely, under functional dependencies and unary inclusion dependencies, with no restriction on their interaction, and arbitrary arity signatures. Following (Rosati, 2008; Ibáñez-García et al., 2014) we close the dependencies under implication on finite structures using an existing construction (Cosmadakis et al., 1990); and we show that this approach is sound

in our context. This is done by a very intricate construction of a finite universal model, reminiscent of (Ibáñez-García et al., 2014) but independently developed, with additional work required to handle the higher arity facts and functional dependencies.

Ongoing work. I have written up these results in October–December 2013 as a PODS 2014 submission (Amarilli, 2014a), which has been unfortunately rejected. My main focus until January 2015, following a two-week visit to Oxford in November 2014, is to work with Michael Benedikt on writing them up again. The goal is to submit the finite query answering contribution to LICS 2015, and hopefully extend and submit the other results to a different venue.

In the meantime, I have presented this work at the Dahu working group at ENS Cachan in January 2014, and at the Dagstuhl seminar "Querying and Reasoning under Expressive Constraints", where I have been able to discuss with the authors of (Ibáñez-García et al., 2014).

2.3 Crowdsourcing

I have visited Tel Aviv in 2012–2013 for a three-month research internship supervised by Tova Milo, again before the start of my PhD, and worked there with Yael Amsterdamer and Tova on extensions of their crowd data mining work (Amsterdamer et al., 2013).

General presentation. Crowd data mining aims at inferring interesting trends about data for which there is no centralized database, but that can be accessed *intensionally*, by posing questions to humans. Possible application domains are the study of folk medicine practices (i.e., figure out what people usually do when they are sick), and the question of finding frequent activity combinations (e.g., to plan a day out in a foreign city). To access the knowledge of people, we use crowdsourcing platforms: they are a system to pose human intelligence queries to a crowd of general users, paying them a small amount of money for each answer.

Our work studies frequent itemset identification, a standard data mining task, under a taxonomy on the items (Srikant and Agrawal, 1995). We want to identify which itemsets are frequent, and do so intensionally by asking crowd users whether individual itemsets are frequent. We study strategies to choose interactively the next question to ask, following a tradeoff between the number of queries to ask, and the computational effort invested to choose them. Outside of the crowd context, this work also relates to data mining through oracle calls (Mannila and Toivonen, 1997).

Publication. Our work was submitted in August 2013 and presented in March 2014 to the ICDT conference (Amarilli et al., 2014a). I have rehearsed this presentation as a seminar in Tel Aviv University, and gave the presentation again during my July visit in Lille. Discussing with Radu Ciucanu and Angela Bonifati at ICDT and in Lille, we were able to identify a connection between our approach and their research on the identification of join patterns in a query-by-example context (Bonifati et al., 2014).

Ongoing work. Our understanding of the problem so far does not satisfactorily account for the *uncertainty* on crowd answers, even though this would be desirable as crowd workers are typically very unreliable.

We have published an initial vision of an extension of our approach to the UnCrowd 2014 workshop (Amarilli et al., 2014b), which I presented in April 2014, and we have continued this research with my coauthors and my supervisor, including a two-week visit in Tel Aviv in March 2014.

Our initial motivation was to mine top-k frequent itemsets from the crowd, but we focused more on the problem of consolidating and extrapolating uncertain numerical crowd answers on correlated values (e.g., under the constraints on support values imposed by monotonicity). We identified an

interesting connection to volume computation and sampling in convex polytopes, which yields hardness results and tractable approximations, and motivates a study of how linear interpolation can generalize to general posets. Our research is connected by the recent work of our co-authors to improve the design of crowd mining systems (Amsterdamer et al., 2014).

2.4 Sampling and pricing

I have worked with Ruiming Tang (National University of Singapore) during his two-month visit at Télécom ParisTech in 2013–2014, on the topic of pricing for XML data; my supervisor and Stéphane Bressan (NUS) are also involved in this work.

General presentation. Data pricing is the problem of intensionality studied from the other side: understand how data providers should set the user cost of data items and queries. The main problem is to ensure desirable properties about the pricing scheme, e.g., arbitrage-freeness: a query should never be more expensive to run than a set of queries that indirectly yields the same information. We focused on the pricing of XML documents, under a simple scheme where the users buys a random subtree of the document whose size will depend on the amount that they paid. Our main technical focus was on how to sample subtrees of XML documents uniformly at random, which is incidentally connected to questions about probabilistic XML (Cohen et al., 2009).

Publication. Our work was presented at the DEXA 2014 conference (Tang et al., 2014) and was invited for a TLKDS special issue, for which we are preparing an extended version. In the meantime, following discussions initiated at the DASFAA 2014 conference and continued during my visit in Singapore, we are looking at the general problem of how to sample efficiently from probabilistic distributions on words and trees, under a more expressive language of constraints. In particular, we introduced a formal language class on words, the p-regular languages, for which we can show that sampling is tractable.

We also plan to submit a challenge paper to JDIQ on the more general topic of *data conditioning*, the question of revising probabilistic representations to integrate new evidence; note that this is one of the needed steps for our overall vision of the system (Amarilli and Senellart, 2014a).

2.5 Uncertain ordered data

I am involved in research with M. Lamine Ba (PhD student in DBWeb), Daniel Deutch (Tel Aviv University), and my advisor, about the representation of *uncertainty* about possible orderings of a relational database.

Our initial motivation came from Lamine's own research about the representation of versioned XML documents using probabilistic XML techniques (Ba et al., 2013a,b). A question left open by these works was that of the order on the nodes of the XML document. Indeed, order is usually crucial in real-world XML documents; however, it was not obvious how this information could be versioned. We accordingly embarked on a study of how to represent the possible orderings in a compact fashion, moving to the relational database setting. We also leveraged Daniel's expertise on semiring provenance to represent the possible choices that could be made when deciding on the order.

Our initial results in this direction were submitted to PODS 2014 (Amarilli et al., 2014c) but have not been accepted. We resumed our study of this problem during my research visit in Tel Aviv and are still working on it, with the plan of submitting a new version to PODS 2015. Our current approach is to extend the relational algebra to uncertain ordered relations, in a way analogous to (Grumbach and Milo, 1995), justifying the naturalness of our semantics by axioms, and studying the expressiveness and complexity questions that arise from it.

Inspired by our work on this topic, I have independently studied the question of possibility for *probabilistic* XML documents, which I believe to be a core question for any probabilistic representation: decide whether an input certain document is a possible world of an input probabilistic document, i.e., whether it can be obtained in some probabilistic outcome. I have studied the complexity of this task for various probabilistic XML languages and problem phrasings (ordered or unordered trees, local or global choices, computing the probability or deciding whether it is > 0). I have presented these results at the AMW workshop (Amarilli, 2014b), in June 2014, and presented an extended version with proofs at BDA in October 2014.

2.6 Knowledge bases

I have worked on knowledge bases during my Master's internship, and more specifically on the PARIS ontology alignment system (Suchanek et al., 2011). Part of my results was published at the VLDS workshop of VLDB 2012 (Oita et al., 2012). Since then, I have been involved in writing up an invited paper to APWeb 2014 (Amarilli et al., 2014e) about the YAGO knowledge base (Suchanek et al., 2007), and an article submitted to WWW 2015 about the extraction of Web entities using unique identifiers (Talaika et al., 2014).

3 Other commitments

Teaching. My contract includes a yearly teaching mission of 64 hours, which also takes place in Télécom ParisTech. The classes I have taught during my first year in 2013–2014, and will teach again during my second year in 2014–2015, are the following:

Formal languages. I am responsible for one of six student groups of the INF105 formal language class (a mandatory class for first year Télécom students), covering the basics of automata theory, regular expressions, and context-free grammars. I teach all classes and exercise sessions for my group. Course evaluations by my students in 2013–2014 have been positive.

Web technologies. I teach all sessions of an M1 course of the COMASIC master, which covers the full stack of Web technologies, client-side and server-side. I wrote my own course material (over 300 slides) as well as the lab assignment.

Programming contest training. With my supervisor, I teach a master-level course (INF280) which trains Télécom students for programming contests such as ACM-ICPC. I developed part of the course material and wrote all assignment solutions. I have additionally helped run Télécom's internal programming contest in 2014 to select the SWERC contestants for Télécom, and will be the Télécom coach at the SWERC in Porto in November 2014.

I also intend to be involved this year (as a proofreader) in the preparation of the computer science exam for the Concours commun Mines–Ponts 2015.

Conferences. Since the beginning of my PhD, I have attended the WebDone workshop, BDA 2013, ICDT 2014, DASFAA 2014 and the UnCrowd 2014 workshop, the AMW 2014 workshop and AMW School, the Dagstuhl seminar "Querying and Reasoning under Expressive Constraints", the Highlights 2014 workshop, and BDA 2014. Except for BDA 2013 and DASFAA 2014, I was a presenter in all these venues.

Peer review. I have been an external reviewer for a special issue of *Distributed and Parallel Databases*, and have informally reviewed submissions to SIGMOD 2014, LATIN 2014 and ICDE 2015.

References

Amarilli, A. (2014a). Open-world query answering under number restrictions. Preprint: `http://a3nm.net/publications/amarilli2014open.pdf`.

Amarilli, A. (2014b). The possibility problem for probabilistic XML. In *Proc. AMW*, Cartagena, Colombia.

Amarilli, A., Amsterdamer, Y., and Milo, T. (2014a). On the complexity of mining itemsets from the crowd using taxonomies. In *Proc. ICDT*, Athens, Greece.

Amarilli, A., Amsterdamer, Y., and Milo, T. (2014b). Uncertainty in crowd data sourcing under structural constraints. In *Proc. UnCrowd*, Denpasar, Indonesia.

Amarilli, A., Ba, L. M., Deutch, D., and Senellart, P. (2014c). Provenance for nondeterministic order-aware queries. Preprint: `http://a3nm.net/publications/amarilli2014provenance.pdf`.

Amarilli, A., Bourhis, P., and Senellart, P. (2014d). Probabilities and provenance via tree decompositions. Preprint: `http://a3nm.net/publications/amarilli2015probabilities.pdf`. Submitted to PODS 2015.

Amarilli, A., Galárraga, L., Preda, N., and Suchanek, F. M. (2014e). Recent topics of research around the YAGO knowledge base. In *APWEB*.

Amarilli, A. and Senellart, P. (2013). On the connections between relational and XML probabilistic data models. In *Proc. BNCOD*, Oxford, United Kingdom.

Amarilli, A. and Senellart, P. (2014a). UnSAID: Uncertainty and structure in the access to intensional data. Preprint: `http://a3nm.net/publications/amarilli2014unsaid.pdf`. Vision article.

Amarilli, A. and Senellart, P. (2014b). What is the best thing to do next?: A tutorial on intensional data management. Preprint: `http://a3nm.net/publications/amarilli2015what.pdf`. Tutorial proposal. Submitted to EDBT/ICDT 2015.

Amsterdamer, Y., Davidson, S. B., Milo, T., Novgorodov, S., and Somech, A. (2014). OASSIS: query driven crowd mining. In *Proc. SIGMOD*, Snowbird, USA.

Amsterdamer, Y., Grossman, Y., Milo, T., and Senellart, P. (2013). Crowd mining. In *Proc. SIGMOD*, New York, USA.

Arnborg, S., Lagergren, J., and Seese, D. (1991). Easy problems for tree-decomposable graphs. *J. Algorithms*, 12(2).

Ba, M. L., Abdessalem, T., and Senellart, P. (2013a). Merging uncertain multi-version XML documents. In *Proc. DChanges*, Florence, Italy.

Ba, M. L., Abdessalem, T., and Senellart, P. (2013b). Uncertain version control in open collaborative editing of tree-structured documents. In *Proc. DocEng*, Florence, Italy.

Bonifati, A., Ciucanu, R., and Staworko, S. (2014). Interactive inference of join queries. In *Proc. EDBT*, Athens, Greece.

Calì, A., Lembo, D., and Rosati, R. (2003). On the decidability and complexity of query answering over inconsistent and incomplete databases. In *PODS*.

Cohen, S., Kimelfeld, B., and Sagiv, Y. (2009). Running tree automata on probabilistic XML. In *Proc. PODS*, Providence, USA.

Cosmadakis, S. S., Kanellakis, P. C., and Vardi, M. Y. (1990). Polynomial-time implication problems for unary inclusion dependencies. *JACM*, 37(1).

Courcelle, B. (1990). Graph rewriting: An algebraic and logic approach. In *Handbook of Theoretical Computer Science*. Elsevier.

Galárraga, L., Teflioudi, C., Hose, K., and Suchanek, F. M. (2013). AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *Proc. WWW*, Rio de Janeiro, Brazil.

Grumbach, S. and Milo, T. (1995). An algebra for pomsets. In *Proc. ICDT*, Prague, Czech Republic.

Ibáñez-García, Y., Lutz, C., and Schneider, T. (2014). Finite model reasoning in horn description logics. In *Proc. KR*, Vienna, Austria.

Maniu, S., Cheng, R., and Senellart, P. (2014). ProbTree: A query-efficient representation of probabilistic graphs. In *Proc. BUDA*, Snowbird, USA.

Mannila, H. and Toivonen, H. (1997). Levelwise search and borders of theories in knowledge discovery. *Data mining and knowledge discovery*, 1(3).

Oita, M., Amarilli, A., and Senellart, P. (2012). Cross-fertilizing deep Web analysis and ontology enrichment. In *Proc. VLDS*, Istanbul, Turkey. Vision article.

Pratt-Hartmann, I. (2009). Data-complexity of the two-variable fragment with counting quantifiers. *Inf. Comput.*, 207(8).

Rosati, R. (2006). On the decidability and finite controllability of query processing in databases with incomplete information. In *Proc. PODS*, Chicago, USA.

Rosati, R. (2008). Finite model reasoning in DL-Lite. In *Proc. ESWC*, Tenerife, Spain.

Srikant, R. and Agrawal, R. (1995). Mining generalized association rules. In *Proc. VLDB*, Zurich, Switzerland.

Suchanek, F. M., Abiteboul, S., and Senellart, P. (2011). Paris: Probabilistic alignment of relations, instances, and schema. *Proc. VLDB*, 5(3).

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Talaika, A., Biega, J., Amarilli, A., and Suchanek, F. M. (2014). Harvesting entities from the web using unique identifiers. Preprint: `http://a3nm.net/publications/talaika2015harvesting.pdf`. Submitted to WWW 2015.

Tang, R., Amarilli, A., Senellart, P., and Bressan, S. (2014). Get a sample for a discount: Sampling-based XML data pricing. In *Proc. DEXA*, Munich, Germany.

Vrandecic, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10).