Project-Team SHAMAN

A Symbolic and Human-Centric View of Data Management

Lannion

Activity Report

2016

1 Team

Head of the team

Olivier Pivert, Professor, Enssat

Administrative assistant

Joëlle Thépault, Enssat, 20%
Vincent Chevrette, Enssat, 10% (up to September 2016)
Angélique Le Pennec, Enssat, 10% (since September 2016)

Université Rennes 1 personnel

Laurent D'Orazio, Professor, IUT Lannion (since September 1, 2016)
François Goasdoué, Professor, Enssat
Hélène Jaudoin, Associate Professor, Enssat
Ludovic Liétard, Associate Professor, HDR, IUT Lannion
Pierre Nerzic, Associate Professor, IUT Lannion
Daniel Rocacher, Professor, Enssat
Grégory Smits, Associate Professor, IUT Lannion
Virginie Thion, Associate Professor, Enssat

PhD students

Le Trung Dung, Vietnam government grant MOET 911, since September 2015; in Shaman since September 2016,
Ngoc Toan Duong, CIFRE with Semsoft, since October 2016
Sara El Hassad, Région Bretagne grant and Lannion Trégor Communauté grant, since October 2014
William Correa Beltran, Région Bretagne grant and Conseil Général 22 grant, since October 2012
Aurélien Moreau, DGA contract, since November 2014
Olfa Slama, DGA contract, since November 2014

2 Overall Objectives

In database research, the last two decades have witnessed a growing interest in preference queries on the one hand, and uncertain databases on the other hand.

Motivations for introducing preferences inside database queries are manifold. First, it has appeared to be desirable to offer more expressive query languages that can be more faithful to what a user intends to say. Second, the introduction of preferences in queries provides a basis for rank-ordering the retrieved items, which is especially valuable in case of large sets of items satisfying a query. Third, on the contrary, a classical query may also have an empty set of answers, while a relaxed (and thus less restrictive) version of the query might be matched by items in the database.

Approaches to database preference queries may be classified into two categories according to their qualitative or quantitative nature. In the qualitative approach, preferences are defined through binary preference relations. Among the representatives of this family of approaches, let us mention an approach based on CP-nets, and those relying on a dominance relation, e.g. Pareto order, in particular Skyline queries. In the quantitative approach, preferences are expressed quantitatively by a monotone scoring function (the overall score is positively correlated with partial scores). Since the scoring function associates each tuple with a numerical score, tuple t_1 is preferred to tuple t_2 if the score of t_1 is higher than the score of t_2. Well-known representatives of this family of approaches are top-k queries, and *fuzzy-set-based approaches*. The team Shaman particularly studies the latter, and the line followed is to focus on:

1. various types of flexible conditions, including non-trivial ones,
2. the semantics of such conditions from a user standpoint,
3. the design of query languages providing flexible capabilities in a relational setting.

Basically, a fuzzy query involves linguistic terms corresponding to gradual predicates, i.e., predicates which are more or less satisfied by a given (attribute) value. In addition, these various terms may have different degrees of importance, which means that they may be connected by operators beyond conjunction and disjunction. For instance, in the context of a search for used vehicles, a user might say that he/she wants a *compact* car *preferably French*, with a *medium* mileage, *around* 6 k$, whose color is *as close as possible* to light grey or blue. The terms appearing in this example must be specified, which requires a certain theoretical framework. For instance, one may think that "*preferably* French" corresponds to a complete satisfaction for French cars, a lower one for Italian and Spanish ones, a still smaller satisfaction for German cars and a total rejection for others. Similarly, "*medium* mileage" can be used to state that cars with less than 40000 km are totally acceptable while the satisfaction decreases as the mileage goes up to 75000 km which is an upper bound. Moreover, it is likely that some of the conditions are more important than others (e.g., the price with respect to the color). In such a context, answers are ordered according to their overall compliance with the query, which makes a major difference with respect to usual queries.

In the previous example, conditions are fairly simple, but it turns out that more complex ones can also be handled. A particular attention is paid to conditions calling on aggregate functions together with gradual predicates. For instance, one may look for departments where *most* employees are *close* to retirement, or where the average salary of *young* employees is *around* $2500. Such statements have their counterpart in regular query language, such as SQL, and the specification of their semantics, when gradual conditions come into play, is studied in the project.

Along this line, the ultimate goal of the project is to introduce gradual predicates inside database query languages, thus providing flexible querying capabilities. Algebraic languages as well as more user-oriented languages are under consideration in both the original and extended relational settings.

As to the second topic mentioned at the beginning of this introduction, i.e., uncertain databases, it already has a rather long history. Indeed, since the late 70s, many authors have

made diverse proposals to model and handle databases involving uncertain or incomplete data. In particular, the last two decades have witnessed a profusion of research works on this topic. The notion of an uncertain database covers two aspects: i) attribute uncertainty: when some attribute values are ill-known; ii) existential uncertainty: when the existence of some tuples is itself uncertain. Even though most works about uncertain databases consider probability theory as the underlying uncertainty model, some approaches rather rely on possibility theory. The issue is not to demonstrate that the possibility-theory-based framework is "better" than the probabilistic one at modeling uncertain databases, but that it constitutes an interesting alternative inasmuch as it captures a different kind of uncertainty (of a subjective, nonfrequential, nature). A typical example is that of a person who witnesses a car accident and who does not remember for sure the model of the car involved. In such a case, it seems reasonable to model the uncertain value by means of a possibility distribution, e.g., {1/Mazda, 1/Toyota, 0.7/Honda} rather than with a probability distribution which would be artificially normalized. In contrast with probability theory, one expects the following advantages when using possibility theory:

- the qualitative nature of the model makes easier the elicitation of the degrees attached to the various candidate values;
- in probability theory, the fact that the sum of the degrees from a distribution must equal 1 makes it difficult to deal with incompletely known distributions;
- there does not exist any probabilistic logic which is complete and works locally as possibilistic logic does: this can be problematic in the case where the degrees attached to certain pieces of data must be automatically deduced from those attached to some other pieces of data (e.g., when data coming from different sources are merged into a single database).

A recent research topic in Shaman concerns flexible data integration systems. One considers a distributed database environment where several data sources are available. An extreme case is that of a totally decentralized P2P system. An intermediary situation corresponds to the case where several global schemas are available and where the sources can be accessed through views defined on one of these schemas (LAV approach). The problem consists in handling a user query (possibly involving preferences conveyed by fuzzy terms) so as to forward it (or part of it) to the relevant data sources, after rewriting it using the views. The overall objective is thus to define flexible query rewriting techniques which take into account both the approximate nature of the mappings and the graded nature of the initial query. A large scale environment is aimed, and the performance aspect is therefore crucial in such a context.

3 Scientific Foundations

The project investigates the issues of flexible queries against regular databases as well as regular queries addressed to databases involving imprecise data. These two aspects make use of two close theoretic settings: fuzzy sets for the support of flexibility and possibility theory for the representation and treatment of imprecise information.

3.1 Fuzzy sets

Fuzzy sets were introduced by L.A. Zadeh in 1965 [Zad65] in order to model sets or classes whose boundaries are not sharp. This is particularly the case for many adjectives of the natural language which can be hardly defined in terms of usual sets (e.g., high, young, small, etc.), but are a matter of degree. A fuzzy (sub)set F of a universe X is defined thanks to a membership function denoted by μ_F which maps every element x of X into a degree $\mu_F(x)$ in the unit interval $[0, 1]$. When the degree equals 0, x does not belong at all to F, if it is 1, x is a full member of F and the closer $\mu_F(x)$ to 1 (resp. 0), the more (resp. less) x belongs to F. Clearly, a regular set is a special case of a fuzzy set where the values taken by the membership function are restricted to the pair $\{0, 1\}$. Beyond the intrinsic values of the degrees, the membership function offers a convenient way for ordering the elements of X and it defines a symbolic-numeric interface. The α level-cut of a fuzzy set F is defined as the (regular) set of elements whose degree of membership is greater than or equal to α and this concept bridges fuzzy sets and ordinary sets.

Similarly to a set A which is often seen as a predicate (namely, the one appearing in the intensional definition of A), a fuzzy set F is associated with a gradual (or fuzzy) predicate. For instance, if the membership function of the fuzzy set *young* is given by: $\mu_{young}(x) = 0$ for any $x \geq 30$, $\mu_{young}(x) = 1$ for any $x < 21$, $\mu_{young}(21) = 0.9$, $\mu_{young}(22) = 0.8$, ... , $\mu_{young}(29) = 0.1$, it is possible to use the predicate *young* to assess the extent to which Tom, who is 26 years old, is young ($\mu_{young}(26) = 0.4$).

The operations valid on sets (and their logical counterparts) have been extended to fuzzy sets. Their definition assumes the validity of the commensurability principle between the concerned fuzzy sets. It has been shown that it is impossible to maintain all of the properties of the Boolean algebra when fuzzy sets come into play. Fuzzy set theory starts with a strongly coupled definition of union and intersection which rely on triangular norms ($\top$) and co-norms ($\bot$) tied by de Morgan's laws. Then:

$$\mu_{A\cap B}(x) = \top(\mu_A(x), \mu_B(x)) \qquad \mu_{A\cup B}(x) = \bot(\mu_A(x), \mu_B(x))$$

The complement of a fuzzy set F, denoted by $\bar{F}$, is a fuzzy set such that: $\mu_{\bar{F}}(x) = neg(\mu_F(x))$, where *neg* is a strong negation operator and the complement to 1 is generally used. The conjunction and disjunction operators are the logical counterpart of intersection and union while the negation is the counterpart of the complement.

In practice, minimum and maximum are the most commonly used norm and co-norm because they have numerous properties among which:

- the satisfaction of all the properties of the usual intersection and union (including idempotency and double distributivity), except excluded-middle and non-contradiction laws,
- they still work with an ordinal scale, which is less demanding than numerical values over the unit interval,
- the simplicity of the underlying calculus.

[Zad65] L. Zadeh, "Fuzzy sets", *Information and Control 8*, 1965, p. 338–353.

Once these three operators given, others can be extended to fuzzy sets, such as the difference:

$$\mu_{E-F}(x) = \top(\mu_E(x), \mu_{\bar{F}}(x))$$

and the Cartesian product:

$$\mu_{E\times F}(x, y) = \top(\mu_E(x), \mu_F(y)).$$

The inclusion can be applied to fuzzy sets in a straightforward way: $E \subseteq F \Leftrightarrow \forall x, \mu_E(x) \leq \mu_F(x)$, but a gradual view of the inclusion can also be introduced. The idea is to consider that E may be more or less included in F. Different approaches can be considered, among which one is based on the notion of a fuzzy implication (the usual logical counterpart of the inclusion). The starting point is the following definition valid for sets:

$$E \subseteq F \Leftrightarrow \forall x, x \in E \Rightarrow x \in F$$

which becomes :

$$deg(E \subseteq F) = \top_x(\mu_E(x) \Rightarrow_f \mu_F(x))$$

where $\Rightarrow_f$ is a fuzzy implication whose arguments and result take their value in the unit interval. Different families of such implications have been identified (notably R-implications and S-implications) and the most common ones are:

- Kleene-Dienes implication : $a \Rightarrow_{K-D} b = \max(1 - a, b)$,
- Rescher-Gaines implication: $a \Rightarrow_{R-G} b = 1$ if $a \leq b$ and 0 otherwise,
- Gödel implication : $a \Rightarrow_{Go} b = 1$ is $a \leq b$ and b otherwise,
- Łukasiewicz implication : $a \Rightarrow_{Lu} b = \min(1, 1 - a + b)$.

Of course, fuzzy sets can also be combined in many other ways, for instance using mean operators, which do not make sense for classical sets.

3.2 Possibility theory

Possibility theory is a theory of uncertainty which aims at assessing the realization of events. The main difference with the probabilistic framework lies in the fact that it is mainly ordinal and it is not related with frequency of experiments. As in the probabilistic case, a measure (of possibility) is associated with an event. It obeys the following axioms [Zad78]:

- $\Pi(X) = 1$,
- $\Pi(\oslash) = 0$,
- $\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$,

[Zad78] L. Zadeh, "Fuzzy sets as a basis for a theory of possibility", *Fuzzy Sets and Systems 1*, 1978, p. 3–28.

where X denotes the set of all events and A, B are two subsets of X. If $\Pi(A)$ equals 1, A is completely possible (but not certain), when it is 0, A is completely impossible and the closer to 1 $\Pi(A)$, the more possible A. From the last axiom, it appears that the possibility of $\bar{A}$, the opposite event of A, cannot be calculated from the possibility of A. The relationship between these two values (for Boolean events) is:

$$max(\Pi(A),\, \Pi(\bar{A})) = 1$$

which stems from the first and third axioms (where B is replaced by $\bar{A}$).

In other words, if A is completely possible, nothing can be deduced for $\Pi(\bar{A})$. This state of fact has led to introduce a complementary measure (N), called necessity, to assess the certainty of A. $N(A)$ is based on the fact that A is all the more certain as $\bar{A}$ is impossible [DP80]:

$$N(A) = 1 - \Pi(\bar{A})$$

and the closer to 1 $N(A)$, the more certain A. From the third axiom on possibility, one derives:

$$N(A \cap B) = \min(N(A),\, N(B)))$$

and, in general:

- $\Pi(A \cap B) \leq \min(\Pi(A),\, \Pi(B))$,
- $N(A \cup B) \geq \max(N(A),\, N(B))$.

In the possibilistic setting, a complete characterization of an event requires the computation of two measures: its possibility and its certainty. It is interesting to notice that the following property holds:

$$\Pi(A) < 1 \Rightarrow N(A) = 0.$$

It indicates that if an event is not completely possible, it is excluded that it is somewhat certain, which makes it possible to define a total order over events: first, the events which are somewhat possible but not at all certain (from ($\Pi = N = 0$ to $\Pi = 1$ and $N = 0$), then those which are completely possible and somewhat certain (from $\Pi = 1$ and $N = 0$ to $\Pi = N = 1$). This favorable situation (existence of a total order) is valid for usual events, but if fuzzy ones are taken into account, this is no longer true (because A $\cup$ $\bar{A} = X$ is not true in general when A is a fuzzy set) and the only valid property is: $\forall$ A, $\Pi(A) \geq$ N(A).

The notion of a possibility distribution [Zad78], denoted by π, plays a role similar to that of a probability distribution. It is a function from the referential X into the unit interval and:

$$\forall A \subseteq X, \Pi(A) = \sup_{x \in A} \pi(x)$$

In order to comply with the second axiom above, a possibility distribution must be such that there exists (at least) an element x_0 of X for which $\pi(x_0) = 1$. Indeed, a possibility

[DP80] D. Dubois, H. Prade, *Fuzzy set and systems: theory and applications*, Academic Press, 1980.

[Zad78] L. Zadeh, "Fuzzy sets as a basis for a theory of possibility", *Fuzzy Sets and Systems 1*, 1978, p. 3–28.

distribution can be seen as a normalized fuzzy set F which represents the knowledge about a given variable. The following formula:

$$\pi(x = a) = \mu_F(a)$$

which is often used, tells that the possibility that the actual value of the considered variable x is a, equals the degree of membership of a to the fuzzy set F. For example, Paul's age may be only imprecisely known as "close to 20", where a given fuzzy set is associated with this fuzzy linguistic expression.

3.3 Fuzzy sets, possibility theory and databases

The project is situated at the crossroads of databases and fuzzy sets. Its main objective is to broaden the capabilities offered by DBMSs according to two orthogonal lines in order to separate two distinct problems:

- flexible queries against regular databases so as to provide users with a qualitative result made of ordered elements,
- Boolean queries addressed to databases containing imprecise attribute values.

Once these two aspects solved separately, the joint issue of flexible queries against databases containing imprecise attribute values will also be considered. This can be envisaged because of the compatibility between the semantics of grades (preferences) in both fuzzy sets and possibility distributions.

It turns out that fuzzy sets offer a very convenient way for modeling gradual concepts and then flexible queries. It has been proven [BP92] that many *ad hoc* approaches (e.g., based on distances) were special cases of what is expressible using fuzzy set theory. This framework makes it possible to express sophisticated queries where the semantic choices of the user can take place (e.g., the meaning of the terms or the compensatory interaction desired between the various fuzzy conditions of a query). The works conducted in Shaman aim at extending algebraic as well as user-oriented query languages in both the relational and the object-oriented (extended relational in practice) settings. The relational algebra has already been revised in order to introduce flexible queries and a particular focus has been put on the division operation. Current works are oriented towards:

- bipolar fuzzy queries (including two parts: one viewed as a constraint, the other as a wish),
- the use of a predefined fuzzy vocabulary (which raises the question of its adequacy wrt to the actual content of the database),
- fuzzy extensions of Skyline queries (based on Pareto order),
- implementation and query optimization issues.

[BP92] P. Bosc, O. Pivert, "Some approaches for relational databases flexible querying", *Journal of Intelligent Information Systems 1*, 1992, p. 323–354.

As to possibility distributions, they are used to represent imprecise (imperfect) data. By doing so, a straightforward connection can be established between a possibilistic database and regular ones. Indeed, a possibilistic database is nothing but a weighted set of regular databases (called worlds), obtained by choosing one candidate in every distribution appearing in any tuple of every possibilistic relation. According to this view, a query addressed to a possibilistic database has a natural semantics. However, it is not realistic to process it against all the worlds due to their huge number. Then, the question tied to the querying of a possibilistic database bears mainly on the efficiency, which imposes to obviate the combinatory explosion of the worlds. The objective of the project is to identify different families of queries which comply with this requirement in the context of the relational setting, even if the initial model must obviously be extended (in particular to support imprecise data).

3.4 Ontology-based data management

Data management is a longstanding research topic in *Knowledge Representation* (KR), a prominent discipline of *Artificial Intelligence* (AI), and — of course — in *Databases* (DB).

Till the end of the 20^{th} century, there have been few interactions between these two research fields concerning data management, essentially because they were addressing it from different perspectives. KR was investigating data management according to human cognitive schemes for the sake of intelligibility, e.g. using *Conceptual Graphs* [CM08] or *Description Logics* [BCM+03], while DB was focusing on data management according to simple mathematical structures for the sake of efficiency, e.g. using the *relational model* [AHV95] or the *eXtensible Markup Language* [AMR+12].

In the beginning of the 21^{st} century, these ideological stances have changed with the new era of *ontology-based data management* [Len11]. Roughly speaking, ontology-based data management brings data management one step closer to end-users, especially to those that are not computer scientists or engineers. It basically revisits the traditional architecture of database management systems by decoupling the models with which data is exposed to end-users from the models with which data is stored. Notably, ontology-based data management advocates the use of conceptual models from KR as human intelligible front-ends called *ontologies* [Gru09], relegating DB models to back-end storage.

The *World Wide Web Consortium* (W3C) has greatly contributed to ontology-based data management by providing *standards* for handling data through ontologies, the two *Semantic Web* data models. The first standard, the *Resource Description Framework* (RDF) [W3Ca], was

[CM08] M. Chein, M.-L. Mugnier, *Graph-based Knowledge Representation: Computational Foundations of Conceptual Graphs*, Springer Publishing Company, Incorporated, 2008.

[BCM+03] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider (editors), *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press, 2003.

[AHV95] S. Abiteboul, R. Hull, V. Vianu, *Foundations of Databases*, Addison-Wesley, 1995.

[AMR+12] S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset, P. Senellart, *Web Data Management*, Cambridge University Press, 2012.

[Len11] M. Lenzerini, "Ontology-based data management", 2011.

[Gru09] T. Gruber, "Ontology", *in : Encyclopedia of Database Systems*, Springer US, 2009, p. 1963–1965.

[W3Ca] W3C, "Resource Description Framework", *research report*.

introduced in 1998. It is a graph data model coming with a very simple ontology language, *RDF Schema*, strongly related to description logics. The second standard, the *Web Ontology Language* (OWL) [W3Cb], was introduced in 2004. It is actually a family of well-established description logics with varying expressivity/complexity tradeoffs.

The advent of RDF and OWL has rapidly focused the attention of academia and industry on *practical* ontology-based data management. The research community has undertaken this challenge at the highest level, leading to pioneering and compelling contributions in top venues on Artificial Intelligence (e.g. AAAI, ECAI, IJCAI, and KR), on Databases e.g. ICDT/EDBT, ICDE, SIGMOD/PODS, and VLDB), and on the Web (e.g. ESWC, ISWC, and WWW). Also, open-source and commercial software providers are releasing an ever-growing number of tools allowing effective RDF and OWL data management (e.g. Jena, ORACLE 10/11g, OWLIM, Protégé, RDF-3X, and Sesame).

Last but not least, large societies have promptly adhered to RDF and OWL data management (e.g. library and information science, life science, and medicine), sustaining and begetting further efforts towards always more convenient, efficient, and scalable ontology-based data management techniques.

4 Application Domains

Flexible queries have many potential application domains. Indeed, soft querying turns out to be relevant in a great variety of contexts, such as web search engines, yellow pages, classified advertisements, image or multimedia retrieval. One may guess that the richer the semantics of stored information (for instance images or video), the more difficult it is for the user to characterize his search criterion in a crisp way, i.e., using Boolean conditions. In this kind of situation, flexible queries which involve imprecise descriptions (or goals) and vague terms, may provide a convenient means for expressing information needs.

As for uncertain data management, many potential domains could take advantage of advanced systems capable of storing and querying databases where some pieces of information are imprecise/uncertain: military information systems, automated recognition of objects in images, data warehouses where information coming from more or less reliable sources must be fused and stored, etc.

In the near future, we intend to focus on two application domains:

- Open data management. One of the challenges in web data management today is to define adequate tools allowing users to extract the data that are the most likely to fulfill all or part of their information needs, then to understand and automatically correlate these data in order to elaborate relevant answers or analyses. Open data may be of various levels of quality: they may be imprecise, incomplete, inconsistent and/or their reliability/freshness may be somewhat questionable. An appropriate data model and suitable querying tools must then be defined for dealing with the imperfection that may pervade data in this context. On the other hand, it is of prime importance to provide

[W3Cb] W3C, "Web Ontology Language", *research report*.

end-users with simple and flexible means to better understand and analyze open data. The standards of W3C offer popular languages for representing both open and structured data. Another objective is to propose analytical tools suited to these languages through the construction of RDF data warehouses, whereas fuzzy-set-based data summarization approaches should constitute an important step towards making open data more intelligible to non-expert users.

- Environmental information systems. This work will be performed in collaboration with the Biological Station based in Roscoff (Finistère). The general objective is to define an information system architecture (along with an associated "toolbox") suited to the context of marine biodiversity monitoring and environmental protection. We intend to study three main aspects:
  - definition of a data warehouse model suited to this context, capable of dealing with missing values, imprecise information (a situation which often occurs due to the way data is collected and described, through sampling campaigns and human-performed labeling, in particular), uncertain data (uncertainty is unavoidable when data are obtained by means of predictive models, for instance).
  - identification of new needs in terms of query expression: new OLAP operators suitable for the model, making it possible to handle dimensions described by fuzzy concept trees, to manage fuzzy cardinalities, possibility distributions and so on.
  - knowledge discovery: we are notably interested in exploiting a concept that comes from artificial intelligence but has not been applied in the domain of data management yet: that of an analogical proportion, which underlies propositions of the type "A is to B as C is to D". We believe that discovering such "regularities" in a dataset could prove very useful for many purposes connected to environmental monitoring issues, in particular when it comes to predict the evolution of an ecosystem or the population of a species, etc.

5 Software

Only the most recent prototypes developed by the team are described hereafter. Some more can be found here: `http://www-shaman.irisa.fr/shaman-software/`.

- PostgreSQLF is a flexible querying prototype that aims at evaluating fuzzy queries addressed to regular databases. It is an extension of PostgreSQL which implements the fuzzy query language SQLf defined in the team. This prototype is coupled with a graphical interface names ReqFlex [SPG13] that makes it easy for an end user to specify his/her fuzzy queries.
- COKE (COnnected KEywords): Keyword queries have emerged as the most convenient way to query data sources especially for unexperienced users. Introduced initially for

[SPG13] G. Smits, O. Pivert, T. Girault, "ReqFlex: Fuzzy Queries for Everyone", *PVLDB 6*, 12, 2013, p. 1206–1209.

document retrieval on the web, such queries are defined as an enumeration of keywords corresponding to a rough description of what users are looking for. The interpretation process of keyword queries has then been adapted to handle structured data like relational databases or XML documents. Instead of considering queries as an unstructured enumeration of keywords, the approach underlying the COKE system lets users structure their keyword queries using simple but meaningful grammatical connectors. Using the data structure intensively, a COKE query is translated into SQL to retrieve exact answers. An autocompletion strategy is also proposed to help users take advantage of connectors in their keyword queries [SPJP13]. An experimentation shows that the COKE system efficiently retrieves more relevant and precise answers than classical queries made of keywords enumerations and offers a good coverage of possible query patterns.

- IKEYS [30] is an interactive and cooperative querying systems dedicated to corporate data, that allows users define unambiguous queries in an intuitive way. Users first express their information needs through coarse keyword queries (e.g. "track Jim Morrison 1971") that may then be refined with explicit projection and selection statements involving comparison operators and aggregation functions (e.g., "titles of tracks composed by Jim Morrison before 1971").

- FUDGE/SUGAR: FUDGE [PST15] is a query language allowing to query graph databases — fuzzy or not — in a flexible way. It makes it possible to express preferences queries where preference criteria may concern i) the content of the vertices of the graph and ii) the structure of the graph (which may include weighted vertices and edges when the graph is fuzzy). SUGAR [25] is a prototype, based on Neo4j, implementing the FUDGE language. More information can be found here: `https://www-shaman.irisa.fr/fudge-prototype/`.

- TAMARI (Quality Alerts Management in Graph Databases using Rabbithole) is a prototype, based on the Neo4j graph databases management system, that makes it possible to introduce some functionalities for quality management of graph databases. Based on quality annotation (tags) attached to subgraphs of the data, a quality vocabulary, and user quality profiles, TAMARI implements an extension of the Neo4j Cypher language in order to introduce quality-awareness in queries. See `https://www-shaman.irisa.fr/tamari/`.

6 New Results

6.1 Possibilistic database modeling and querying

Participants: Olivier Pivert, Ludovic Liétard.

[SPJP13] G. Smits, O. Pivert, H. Jaudoin, F. Paulus, "An Autocompletion Mechanism for Enriched Keyword Queries to RDF Data Sources", *in : Proc. of the 10th International Conference on Flexible Query Answering Systems (FQAS'13)*, 2013.

[PST15] O. Pivert, G. Smits, V. Thion, "Expression and Efficient Processing of Fuzzy Queries in a Graph Database Context", *in : Proc. of the 24th IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'15)*, Istanbul, Turkey, 2015.

On the one hand possibility theory and possibilistic logic offer a powerful representation setting in artificial intelligence for handling uncertainty in a qualitative manner. On the other hand conditional tables (c-tables for short) and their probabilistic extension provide a well-known setting for representing respectively incomplete and uncertain information in relational databases. Although these two settings rely on the idea of possible worlds, they have been developed and used independently. In [23], we investigate the links between possibility theory, possibilistic logic and c-tables, before introducing possibilistic c-tables and discussing their relation with a recent certainty-based approach to uncertain databases and their differences with probabilistic c-tables.

6.2 Flexible database querying

6.2.1 Preference queries

Participants: Olivier Pivert, Hélène Jaudoin, Grégory Smits, Virginie Thion, Ludovic Liétard, Daniel Rocacher.

The works presented hereafter deal with different aspects of preference queries (fuzzy and others) in a database context.

- *Skyline refinements.* Skyline queries are a popular and powerful paradigm for extracting interesting objects from a d-dimensional dataset. They rely on Pareto dominance principle to identify the skyline objects, i.e., the set of incomparable objects which are not dominated by any other object from the dataset. In [6], an approach is proposed, that aims at reducing the impact of exceptional points when computing skyline queries. The phenomenon that one wants to avoid is that noisy or suspect elements "hide" some more interesting answers just because they dominate them in the sense of Pareto order. The approach proposed is based on the fuzzy notion of typicality and makes it possible to distinguish between genuinely interesting points and potential anomalies in the skyline obtained. Parallel processing strategies suitable for this type of queries are proposed in [7].

- *Graph databases.* Graph databases have aroused a large interest in the last years thanks to their large scope of potential applications (e.g. social networks, biomedical networks, data stemming from the web). In a similar way as what has already been proposed in relational databases, defining a language allowing a flexible querying of graph databases may greatly improve usability of data. In a previous work, we focused on the notion of fuzzy graph database and described a fuzzy query language that makes it possible to handle such database, which may be fuzzy or not, in a flexible way. This language, called FUDGE, can be used to express preference queries on fuzzy graph databases. The preferences concern i) the content of the vertices of the graph and ii) the structure of the graph. The FUDGE language is implemented in a system, called SUGAR, that is described in [25]. In [28, 29], we deal with *fuzzy quantified queries* in a graph database context. We study a particular type of structural quantified query and show how it can be expressed in the language FUDGE that we previously proposed. A processing strategy based on a

compilation mechanism that derives regular (nonfuzzy) queries for accessing the relevant data is also described.

- *Fuzzy SPARQL.* The Resource Description Framework (RDF) is the graph-based standard data model for representing semantic web information, and SPARQL is the standard query language for querying RDF data. Because of the huge volume of linked open data published on the web, these standards have aroused a large interest in the last years. In [26, 24, 27], we propose a fuzzy extension of the SPARQL language that improves its expressiveness and usability. This extension allows i) to query a *fuzzy RDF data model*, and ii) to express *fuzzy preferences* on data and on the *structure* of the data graph, which has not been proposed in any previous fuzzy extensions of SPARQL.

6.2.2 Cooperative answering, data summarization

Participants: Grégory Smits, Olivier Pivert, Aurélien Moreau,.

The practical need for endowing information systems with the ability to exhibit cooperative behavior (thus making them more "intelligent") has been recognized at least since the early 90s. The main intent of cooperative systems is to provide correct, non-misleading and useful answers, rather than literal answers to user queries. Different aspects of this problem are tackled in the works presented hereafter.

- *Fuzzy query repair.* Telling the user that there is no result for his/her query is not informative and corresponds to the kind of situation cooperative systems try to avoid. Cooperative systems should rather explain the reason(s) of the failure, materialized by Minimal Failing Subqueries (MFS), and build alternative succeeding queries, called maXimal Succeeding Subqueries (XSS), that are as close as possible to the original query. In [8], we consider the context of fuzzy querying and we propose an efficient unified approach to the computation of gradual MFSs and XSSs that relies on a fuzzy-cardinality-based summary of a part of the database.

- *Answer characterization.* In [21, 20], we propose an approach helping users to better understand the results of their queries. These results are structured with a clustering algorithm and described using a personal fuzzy vocabulary. The goal is to find what the elements of a cluster have in common that also differentiates them from the elements of the other clusters, leveraging attributes that do not explicitly appear in the query.

- *Cluster-based summaries.* In [31], a novel approach is introduced to let users extract knowledge from a raw dataset in an intuitive way and using their own vocabulary. The inner structure of a raw dataset is first identified using a clustering algorithm, structure on which specificity-driven measures are defined to extract the most informative knowledge. To let domain experts interact with the cluster-based structure and its embedded knowledge, a graphical visualisation is proposed as well as dedicated query operators.

6.3 Distributed data management

Participants: Laurent D'Orazio, Olivier Pivert, Grégory Smits.

- *Join and recursive query processing in MapReduce.* Big Data management is a big challenge in many applications (Internet, social networks, healthcare, etc.). Paradigms for Massively Parallel Processing (MPP) have thus been proposed. One of the most famous is probably MapReduce. Unfortunately, MapReduce suffers from important limitations, especially for operations on more than one data source or based on an iterative process. [22] is about data management optimization in massively parallel environments and more particularly on optimizing joins in MapReduce. It introduces new filters, intersection filter and difference filter, enabling to reduce the amount of intermediate data, the load and the number of process for joins and recursive queries. Experiments with benchmarks demonstrate the advantages of our solutions.

- *Interactive Keyword Search.* In [30], we present IKEYS, an interactive and cooperative system aimed to query corporate linked data. With IKEYS, users first express their information needs through coarse keyword queries (e.g., "track J. Morrison 1971") that may then be refined with explicit projection and selection statements involving comparison operators and agggregates (e.g., "title of track composed by J. Morrison before 1971"). The demonstration scenario described in [30] aims to show that IKEYS makes it possible to express complex queries in a very easy way, and illustrates the fact that this approach is both more expressive than regular keyword-based techniques and much more efficient than NL-based approaches.

6.4 Ontology-based data management

Participants: Sara El Hassad, François Goasdoué, Hélène Jaudoin.

- *Efficient query answering techniques for Semantic Web data.*
  In the presence of an ontology, query answers must reflect not only data explicitly present in the database, but also implicit data, which holds due to the ontology, even though it is not present in the database. A large and useful set of ontology languages enjoys *First Order Logic (FOL) reducibility of query answering*: answering a query can be reduced to evaluating a certain first-order logic (FOL) formula (obtained from the query and ontology) against only the explicit facts. In [18], we present a *novel query optimization framework for ontology-based data access settings enjoying FOL reducibility.* Our framework is based on searching within a set of alternative equivalent FOL queries, i.e. FOL reformulations, one with minimal evaluation cost when evaluated through a relational database system. We apply this framework to the DL-lite$\mathcal{R}$ Description Logic underpinning the W3C's OWL2 QL ontology language, and we demonstrate through experiments its performance benefits when two leading SQL systems, one open-source and one commercial, are used for evaluating the FOL query reformulations. This approach has been implemented and demonstrated in [17]. See also [3].

- *Query-oriented summarization of RDF graphs.*
  The Resource Description Framework (RDF) is the W3C's graph data model for Semantic Web applications. n [19], we study the problem of RDF graph summarization:

given an input RDF graph G, find an RDF graph S_G which summarizes G as accurately as possible, while being possibly orders of magnitude smaller than the original graph. Summaries are aimed as a help for RDF graph exploration, as well as query formulation and optimization. We devise four kinds of RDF graph summaries obtained as quotient graphs, with equivalence relations reflecting the similarity between nodes wrt their types or connections. We also study whether they enjoy the formal properties of representativeness (S_G should represent as much information about G as possible) and accuracy (S_G should avoid, to the possible extent, reflecting information that is not in G). Finally, we report the experiments we made on several synthetic and real-life RDF graphs.

- *Querying inconsistent description logic knowledge bases.*
  Several inconsistency-tolerant semantics have been introduced for querying inconsistent description logic knowledge bases. In [10], we study the problem of explaining why a tuple is a (non-)answer to a query under such semantics. We define explanations for positive and negative answers under the brave, AR and IAR semantics. We then study the computational properties of explanations in the lightweight description logic DL-lite$\mathcal{R}$. For each type of explanation, we analyze the data complexity of recognizing (preferred) explanations and deciding if a given assertion is relevant or necessary. We establish tight connections between intractable explanation problems and variants of propositional satisfiability (SAT), enabling us to generate explanations by exploiting solvers for Boolean satisfaction and optimization problems. Finally, we empirically study the efficiency of our explanation framework using the well-established LUBM benchmark.

  In [12, 11], we consider the problem of query-driven repairing of inconsistent DL-Lite knowledge bases: query answers are computed under inconsistency-tolerant semantics, and the user provides feedback about which answers are erroneous or missing. The aim is to find a set of ABox modifications (deletions and additions), called a repair plan, that addresses as many of the defects as possible. After formalizing this problem and introducing different notions of optimality, we investigate the computational complexity of reasoning about optimal repair plans and propose interactive algorithms for computing such plans. For deletion-only repair plans, we also present a prototype implementation of the core components of the algorithm. See also [2].

- *Semantic search within social data.*
  Social content such as blogs, tweets, news etc is a rich source of interconnected information. In [16, 15], we identify a set of requirements for the meaningful exploitation of such rich content, and present a new data model, called S4, which is the first to satisfy them. S4 captures *social* relationships between users, and between users and content, but also the *structure* present in rich social content, as well as its *semantics*. We provide the first top-k keyword search algorithm taking into account the social, structured, and semantic dimensions and formally establish its termination and correctness. Experiments on real social networks demonstrate the efficiency and qualitative advantage of our algorithm through the joint exploitation of the social, structured, and semantic dimensions of S4. See also [1].

- *Data management tools for journalists.*

As the world's affairs get increasingly more digital, timely production and consumption of news require to efficiently and quickly exploit heterogeneous data sources. Discussions with journalists revealed that content management tools currently at their disposal fall very short of expectations. In [14, 13], we demonstrate Tatooine, a lightweight data integration prototype, which allows to quickly set up integration queries across (very) heterogeneous data sources, capitalizing on the many data links (joins) available in this application domain. Our demonstration is based on scenarios we study in collaboration with Le Monde, France's major newspaper.

- *Reasoning using ontologies.*
  Finding commonalities between descriptions of data or knowledge is a fundamental task in Machine Learning. The formal notion characterizing precisely such commonalities is known as least general generalization of descriptions and was introduced by G. Plotkin in the early 70's, in First Order Logic. Identifying least general generalizations has a large scope of database applications ranging from query optimization (e.g., to share commonalities between queries in view selection or multi-query optimization) to recommendation in social networks (e.g., to establish connections between users based on their commonalities between profiles or searches). [32] re-visits the notion of least general generalization in the entire Resource Description Framework (RDF) and popular conjunctive fragment of SPARQL, a.k.a. Basic Graph Pattern (BGP) queries. Our contributions include the definition and the computation of least general generalizations in these two settings, which amounts to finding the largest set of commonalities between incomplete databases and conjunctive queries, under deductive constraints. We also provide an experimental assessment of our technical contributions.

6.5 Data quality

Participants: Virginie Thion.

- *Quality assessment in collaborative score libraries.* In [9], we examine quality issues raised by the development of XML-based Digital Score Libraries. Based on the authors' practical experience, the paper exposes the quality shortcomings inherent to the complexity of music encoding, and the lack of support from state-of-the-art formats. We also identify the various facets of the "quality" concept with respect to usages and motivations. We finally propose a general methodology to introduce quality management as a first-level concern in the management of score collections, and an initial taxonomy of quality problems based on real use cases.

- *Survey on linked open data quality management.* Under the impulse of new technologies enabling to publish and exploit data as well as regulatory constraints forcing some companies and institutions to make their data public, the publishing of linked data has become a quickly increasing phenomenon. This huge data resource offers great possibilities, however one may notice a great variety of quality levels among the published data, which makes their use difficult and even risky. Assessing the quality of such data has thus

become a crucial challenge. In [5], we provide a state-of-the-art of the methodological and technical approaches to linked open data quality management, that covers both the dimensions and metrics, the management frameworks, the platforms and related tools, as well as use cases of quality-centered publishing and usage of linked open data. Relying on this state-of-the-art, we exhibit open problems and research perspectives in this domain.

7 Other Grants and Activities

7.1 National actions

François Goasdoué is involved in the following projects:

- Datalyse (Investissements d'Avenir, *Big Data / Cloud computing*, 2013–2016). This project deals with Big Data management in a cloud architecture. The consortium is made of industrial partners (Eolas – Business & Decision and Les Mousqueraires), academic partners (Inria, LIFL of Univ. Lille, LIG of Univ. Grenoble, LIRMM of Univ. Montpellier), as well as the city of Grenoble as an open data provider.
- ANR JCJC Pagoda (2013–2017). PAGODA (Practical algorithms for ontology-based data access) is a basic research project whose objective is to improve the efficiency and robustness of ontology-based data access by developing scalable algorithms for query answering in the presence of ontologies as well as pragmatic approaches to handling inconsistent data. Partners are from LIG of Univ. Grenoble, LIRMM of Univ. Montpellier, and LRI of Univ. Paris-Sud.
- ANR ContentCheck, whose other partners are INRIA Saclay, LIMSI (Orsay), LIRIS (Lyon) and the team in charge of the blog "Les Décodeurs" associated with the newspaper Le Monde (http://www.lemonde.fr/les-decodeurs/). This project has been accepted in August 2015 and is to start in the last quarter of 2015.

François Goasdoué, Hélène Jaudoin, Olivier Pivert, Grégory Smits, and Virginie Thion are involved in the DGA project ODIN (Open Data INtelligence) which started in November 2014. The other partners involved are Semsoft and INRIA Saclay. The ODIN project aims to propose a data management and business intelligence solution for big data, i.e., large-scale heterogeneous and imperfect data distributed over several sources. For doing so, we intend to conceive a data processing and multidimensional analysis chain suitable for RDF data, taking into account the data quality aspect.

Grégory Smits and Olivier Pivert are involved in the project 360 Predict (Projet PME du pôle Images et Réseaux), which aims at developing a web tool for predictive scoring. The other partners are two start-ups: PredicSis (Lannion) and Semsoft (Rennes).

Virginie Thion coordinates the project GioQoso (défi CNRS mastodons 2016) about quality management of open musical scores (see `https://gioqoso.irisa.fr/` for more details). Apart from IRISA/Shaman, the other participants are the teams CNAM/CEDRIC (Paris), CNRS/IREMUS (Paris) and CESR (Tours). Olivier Pivert, from Shaman, is also involved in this project.

7.2 International actions

- Grégory Smits gave a Master's course about Fuzzy Preferences Queries at the Hanoi University of Science and Technology (HUST) in January 2015.

8 Dissemination

8.1 Teaching

Project members give lectures in different faculties of engineering, in the third cycle University curriculum: "Bases de données avancées" and "Web data management" in the speciality "Interaction Intelligente avec l'Information" of the Master's degree in computer science at University of Rennes 1, and at Enssat (third year level cursus).

8.2 Scientific activities

8.2.1 Highlights of the year

- William Correa defended his Ph.D. thesis [4] on July 18, 2016.
- A new permanent member, Laurent D'Orazio, joined the team in September 2016.

8.2.2 Program committees

Laurent D'Orazio served as a member of the following program committees:

- International Workshop on Multi-Engine Data AnaLytics (MEDAL@EDBT 2016);
- International Conference on Business Process Management (BPM 2016) (demos);
- Workshop on Testing and Quality Assurance and Services for Big Data and Application Systems @SEKE 2016;
- International Symposium on Information and Communication Technology (SoICT 2016);
- Colloque sur l'Optimisation et les Systèmes d'Information (COSI 2016);
- Journées Francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2016).

François Goasdoué served as a member of the following program committees:

- Journées Bases de Données Avancées (BDA), Poitiers, France, November 15-18, 2016;
- 22$^{\text{nd}}$ European Conference on Artificial Intelligence (ECAI), The Hague, Holland, August 29-September 2, 2016;
- 13$^{\text{rd}}$ European Semantic Web Conference (ESWC), Heraklion, Crete, Greece, May 29-June 2, 2016;

- 32nd International Conference on Data Engineering (ICDE), Helsinky, Finland, May 16-20, 2016;
- 25th International Joint Conference on Artificial Intelligence (IJCAI), New York, USA, July 9-15, 2016;
- International Conference on Scientific and Statistical Database Management (SSDBM), Budapest, Hungary, July 18-20, 2016.

H. Jaudoin served as a member of the following program committee:

- Journées Bases de Données Avancées (BDA), Poitiers, France, November 15-18, 2016.

L. Liétard served as a member of the following program committees:

- 31st ACM Symposium on Applied Computing (SAC 2016), Pisa, Italy, April 4-8, 2016;
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2016), La Rochelle, France, November 3-4, 2016.

O. Pivert served as a member of the following program committees:

- 31st ACM Symposium on Applied Computing (SAC 2016), Pisa, Italy, April 4-8, 2016;
- IEEE International Conference on Fuzzy Systems (Fuzz-IEEE 2016), Vancouver, Canada, July 25-29, 2016.
- 16th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2016), Eindhoven, The Netherlands, June 20-24, 2016.
- International Conference on Fuzzy Management Methods (ICFMsquare'16), Fribourg, Switzerland, September 29-30, 2016.
- 17th International Conference on Web Information Systems Engineering (WISE 2016), Shanghai, China, November 7-10, 2016;
- 27th International Conference on Database and Expert Systems Applications (DEXA 2016), Porto, Portugal, September 5-8, 2016;
- 10th International Conference on Scalable Uncertainty Management (SUM 2016), Nice, France, September 21-23, 2016;
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2016), La Rochelle, France, November 3-4, 2016.
- Atelier Big data et Intelligence Artificielle (BigIA 2016), Lyon, December 2, 2016.

D. Rocacher served as a member of the following program committee:

- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2016), La Rochelle, France, November 3-4, 2016.

G. Smits served as a member of the following program committees:

- 27^{th} International Conference on Database and Expert Systems Applications (DEXA 2016), Porto, Portugal, September 5-8, 2016;

- 16^{th} International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2016), Eindhoven, The Netherlands, June 20-24, 2016.

V. Thion served as a member of the following program committee:

- Workshop QLOD 2016 (Quality of Linked Open data) associated with the conference EGC 2016, Reims, France, January 19, 2016.

8.2.3 Editorial boards

Olivier Pivert is a member of the following editorial boards:

- Journal of Intelligent Information Systems,

- Fuzzy Sets and Systems,

- International Journal of Fuzziness, Uncertainty and Knowledge-Based Systems,

8.2.4 Steering committees

O. Pivert is as a member of the steering committee of the French-speaking conference "Rencontres Francophones sur la Logique Floue et ses Applications" (LFA).

8.2.5 International advisory boards

O. Pivert is as a member of the international advisory board of the International Conference on Flexible Query-Answering Systems (FQAS).

8.2.6 Invited talks

- Grégory Smits gave an invited talk about "An Agile Business Intelligence Approach Based on Soft Computing" at the seminary DAPA (Données et Apprentissage Artificiel) organized by LIP6 (Laboratoire d'Informatique de Paris 6) on June 9, 2016.

9 Bibliography

Major publications by the team in recent years

[1] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Explaining Inconsistency-Tolerant Query Answering over Description Logic Knowledge Bases", *in : Proc. of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*, Phoenix, Arizona, USA, 2016.

[2] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Query-driven Repairing of Inconsistent DL-Lite Knowledge Bases", *in : Proc. of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*, New York, NY, USA, 2016.

[3] P. Bosc, O. Pivert, "On a fuzzy bipolar relational algebra", *Inf. Sci. 219*, 2013, p. 1–16.

[4] D. Bursztyn, F. Goasdoué, I. Manolescu, "Teaching an RDBMS about Ontological Constraints", *in : Proc. of the 42nd International Conference on Very Large Data Bases (PVLDB'16)*, New Delhi, India, 2016.

[5] O. Pivert, P. Bosc, *Fuzzy Preference Queries to Relational Databases*, Imperial College Press, London, UK, 2012.

[6] O. Pivert, H. Prade, "A Certainty-Based Model for Uncertain Databases", *IEEE Trans. Fuzzy Systems 23*, 4, 2015, p. 1181–1196.

[7] O. Pivert, G. Smits, V. Thion, "Expression and Efficient Processing of Fuzzy Queries in a Graph Database Context", *in : Proc. of the 24th IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'15)*, Istanbul, Turkey, 2015.

[8] G. Smits, O. Pivert, T. Girault, "ReqFlex: Fuzzy Queries for Everyone", *PVLDB 6*, 12, 2013, p. 1206–1209.

Doctoral dissertations and "Habilitation" theses

[1] R. Bonaque, *Recherche Top-k pour le contenu riche du Web social*, PhD Thesis, University of Paris Saclay – École doctorale STIC, September 30, 2016, supervised by B. Cautis, F. Goasdoué and I. Manolescu.

[2] C. Bourgaux, *Algorithmique robuste pour l'accès aux données en présence d'ontologies*, PhD Thesis, University of Paris Saclay – École doctorale STIC, September 29, 2016, supervised by M. Bienvenu and F. Goasdoué.

[3] D. Bursztyn, *Modèles et algorithmes pour Big Data sémantique*, PhD Thesis, University of Paris Saclay – École doctorale STIC, December 15, 2016, supervised by F. Goasdoué and I. Manolescu.

[4] W. Correa Beltran, *Discovery and Exploitation of Analogical Proportions in Relational Databases*, PhD Thesis, University of Rennes 1 – École doctorale Matisse, July 18, 2016, supervised by O. Pivert and H. Jaudoin.

Articles in referred journals and book chapters

[5] D. Barrau, N. Barthélémy, Z. Kedad, B. Laboisse, S. Nugier, V. Thion, "Gestion de la qualité des données ouvertes liées – État des lieux et perspectives", *RNTI Journal Special Issue on Open Data*, 2016.

[6] H. Jaudoin, P. Nerzic, O. Pivert, D. Rocacher, "On Making Skyline Queries Resistant to Outliers", *in : Advances in Knowledge Discovery and Management, vol. 6*, F. Guillet, B. Pinaud, and G. Venturini (editors), *Studies in Computational Intelligence*, Springer, 2016, p. 19–38.

[7] P. Nerzic, H. Jaudoin, O. Pivert, "Parallel Processing Strategies for Skyline Queries Tolerant to Outliers", *International Journal of Intelligent Systems*, 2016.

[8] G. Smits, O. Pivert, "Une approche coopérative d'aide à la réparation de requêtes floues", *Ingéniérie des Systèmes d'Information 21*, 3, 2016, p. 11–30.

Publications in Conferences and Workshops

[9] V. Besson, M. Gurrieri, P. Rigaux, A. Tacaille, V. Thion, "A Methodology for Quality Assessment in Collaborative Score Libraries", *in : Proc. of the 17th International Society for Music Information Retrieval (ISMIR'16)*, New York City, NY, USA, 2016.

[10] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Explaining Inconsistency-Tolerant Query Answering over Description Logic Knowledge Bases", *in : Proc. of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*, Phoenix, Arizona, USA, 2016.

[11] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Query-driven Repairing of Inconsistent DL-Lite Knowledge Bases", *in : Proc. of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*, New York, NY, USA, 2016.

[12] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Query-driven Repairing of Inconsistent DL-Lite Knowledge Bases (Extended Abstract)", *in : Proc. of the 28th International Workshop on Description Logics (DL'16)*, Cape Town, South Africa, 2016.

[13] R. Bonaque, T.-D. Cao, B. Cautis, F. Goasdoué, J. Letelier, I. Manolescu, O. Mendoza, S. Ribeiro, X. Tannier, M. Thomas, "Interrogation d'instance mixte : une architecture d'intégration légère pour le journalisme de données", *in : Actes des 32es Journées bases de Données Avancées (BDA'16), session démonstration*, Poitiers, France, 2016.

[14] R. Bonaque, B. Cautis, F. Goasdoué, J. Letelier, I. Manolescu, O. Mendoza, S. Ribeiro, X. Tannier, M. Thomazo, "Mixed-Instance Querying: a Lightweight Integration Architecture for Data Journalism", *in : Proc. of the 42nd International Conference on Very Large Data Bases (VLDB'16), demo paper*, New Delhi, India, 2016.

[15] R. Bonaque, B. Cautis, F. Goasdoué, I. Manolescu, "Social, Structured and Semantic Search", *in : Actes des 32es Journées bases de Données Avancées (BDA'16)*, Poitiers, France, 2016.

[16] R. Bonaque, B. Cautis, F. Goasdoué, I. Manolescu, "Structured, Social and Semantic Search", *in : Proc. of the 19th International Conference on Extending Database Technology (EDBT'16)*, p. 41–52, Bordeaux, France, 2016.

[17] D. Bursztyn, F. Goasdoué, I. Manolescu, "Optimizing FOL Reducible Query Answering: Understanding Performance Challenges", *in : Proc. of the 15th International Semantic Web Conference (ISWC'16), demo paper*, Kobe, Japan, 2016.

[18] D. Bursztyn, F. Goasdoué, I. Manolescu, "Teaching an RDBMS about Ontological Constraints", *in : Proc. of the 42nd International Conference on Very Large Data Bases (PVLDB'16)*, New Delhi, India, 2016.

[19] S. Cebiric, F. Goasdoué, I. Manolescu, "Query-Oriented Summarization of RDF Graphs", *in : Actes des 32esJournées bases de Données Avancées (BDA'16)*, Poitiers, France, 2016.

[20] A. Moreau, O. Pivert, G. Smits, "Caractérisation floue de clusters de réponses", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'16)*, La Rochelle, France, 2016.

[21] A. Moreau, O. Pivert, G. Smits, "A Fuzzy Approach to the Characterization of Database Query Answers", *in : Proc. of the 16th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'16), CCIS vol. 611*, p. 329–340, Eindhoven, The Netherlands, 2016.

[22] T. Phan, T. Tran, L. D'Orazio, "Filtres pour jointures et requêtes récursives en MapReduce", *in : Actes des 32esJournées bases de Données Avancées (BDA'16)*, Poitiers, France, 2016.

[23] O. Pivert, H. Prade, "Possibilistic Conditional Tables", *in : Proc. of the 9th International Symposium on Foundations of Information and Knowledge Systems (FoIKS'16), LNCS vol. 9616*, p. 42–61, Linz, Austria, 2016.

[24] O. Pivert, O. Slama, G. Smits, V. Thion, "A Fuzzy Extension of SPARQL for Querying Gradual RDF Data", *in : Proc. of the 10th IEEE International Conference on Research Challenges in Information Science (RCIS'16), poster session*, Grenoble, France, 2016.

[25] O. Pivert, O. Slama, G. Smits, V. Thion, "SUGAR: A Graph Database Fuzzy Querying System", *in : Proc. of the 10th IEEE International Conference on Research Challenges in Information Science (RCIS'16), demo session*, Grenoble, France, 2016.

[26] O. Pivert, O. Slama, V. Thion, "An Extension of SPARQL with Fuzzy Navigational Capabilities for Querying Fuzzy RDF Data", *in : Proc. of the 25th IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'16)*, Vancouver, Canada, 2016.

[27] O. Pivert, O. Slama, V. Thion, "FURQL : une extension floue du langage SPARQL", *in : Actes des 32esJournées bases de Données Avancées (BDA'16)*, Poitiers, France, 2016.

[28] O. Pivert, O. Slama, V. Thion, "Fuzzy Quantified Structural Queries to Fuzzy Graph Databases", *in : Proc. of the 10th International Conference on Scalable Uncertainty Management (SUM'16), LNAI vol. 9858*, p. 260–273, Nice, France, 2016.

[29] O. Pivert, O. Slama, V. Thion, "Requêtes quantifiées floues structurelles sur des bases de données graphe", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'16)*, La Rochelle, France, 2016.

[30] G. Smits, K. Dramé, O. Pivert, "IKEYS: Interactive Keyword Search Dedicated to Corporate Data", *in : Proc. of the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW'16), demo paper*, Bologna, Italy, 2016.

[31] G. Smits, O. Pivert, R.R. Yager, "A Soft Computing Approach to Agile Business Intelligence", *in : Proc. of the 25th IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'16)*, Vancouver, Canada, 2016.

Internal Reports

[32] S. El Hassad, F. Goasdoué, H. Jaudoin, "Learning Commonalities in RDF and SPARQL", *research report*, 2016, https://hal.inria.fr/hal-01386237.