
# Activity Report 2018

## Team SHAMAN

## A Symbolic and Human-Centric View of Data Management

D7 – Data and Knowledge Management

# 1 Team composition

**Researchers and faculty**

Ahmed Abid, post-doctoral researcher from Sep. 2017 to Jul. 2018
Laurent D'Orazio, Professor, IUT Lannion
François Goasdoué, Professor, Enssat
Hélène Jaudoin, Associate Professor, Enssat
Ludovic Liétard, Associate Professor, HDR, IUT Lannion
Pierre Nerzic, Associate Professor, IUT Lannion
Olivier Pivert, Professor, Enssat, head of the team
Daniel Rocacher, Professor, Enssat
Grégory Smits, Associate Professor, IUT Lannion
Virginie Thion, Associate Professor, Enssat

**PhD students** Maxime Buron, IPL iCoda grant, since Oct. 2017
Trung Dung Le, Vietnam government grant MOET 911, since Sep. 2015; in Shaman since Sep. 2016,
Ngoc Toan Duong, CIFRE with Semsoft, from Oct. 2016 to July 2018
Ludivine Duroyon, ANR ContentCheck grant, since Oct. 2017
Sara El Hassad, Région Bretagne grant and Lannion Trégor Communauté grant, Oct. 2014 – Feb. 2018
Cheikh-Brahim El Vaigh, IPL iCoda, since Oct. 2017
Aurélien Moreau, DGA contract, since Nov. 2014
Thi To Quyen, Vietnam government grant MOET 911, since Oct. 2017
Van Hoang Tran, PEC grant, since Dec. 2017
Olfa Slama, DGA contract, Nov. 2014 – Nov. 2017, then ATER till Aug. 18

**Administrative assistant**

Joëlle Thépault, Enssat, 20%
Angélique Le Pennec, Enssat, 20%

# 2 Overall objectives

## 2.1 Overview

The overall goal pursued by Shaman is to improve the data management methods currently used in commercial systems, which suffer from a severe lack of flexibility in several respects. In particular, with the techniques currently available, it is difficult for a user to *i)* understand the data he/she has access to, and to *ii)* specify his/her information needs in an intuitive though sufficiently expressive way. Moreover, these systems/approaches have limited capabilities when it comes to handling imperfect data, in particular in a context where data come from different sources. Shaman addresses these shortcomings and strives to devise new tools with the objective of helping end users and/or database conceptors:

- *model* and *integrate* the data — possibly *heterogeneous* and/or *imperfect* — that are relevant in a given applicative context;
- *understand* the data (structure and semantics) that are accessible to them;
- *query* and *analyze* these data, taking into account their *preferences*, by means of a mechanism as *cooperative* as possible.

We favor *symbolic* approaches for the sake of intelligibility/ease of use (again, the objective is to define *human-centric* data management methods). Fuzzy set theory (and the closely related possibility theory) constitutes a natural and intuitive symbolic/numerical interface, between the symbolic aspect of a linguistic variable and the numerical nature of the corresponding characteristic function valued in the unit interval. Fuzzy set theory can be used to model preference queries, data summaries, and cooperative answering strategies, as well as to define a new data model and querying framework based on *clusters* instead of tables. On the other hand, possibility theory can serve as a basis to the modeling of uncertain databases where uncertainty is assumed to be of a *qualitative*, nonfrequential, nature.

Ontology-based data management is another central topic in Shaman inasmuch as ontologies *i)* are a powerful tool to make data more *intelligible* to users, and to *mediate* between data sources whose schemas differ, *ii)* make it possible to enhance data management systems with *reasoning capabilities*, thus to handle data in a more "intelligent" way.

A strong point of Shaman lies in its positioning at the junction between the Databases and Artificial Intelligence domains. Up to now, these two research communities have stayed much apart from each other, whereas we believe that data management should highly benefit from a cross-fertilization between DB technologies and AI approaches. Historically, the members of the team were always sensitized to this challenge, making use for instance of theoretical tools coming from fuzzy logic for making database querying more flexible. This trend also corresponds to an evolution of the data management landscape itself: the rise of the internet made it necessary to manage open and linked data, using methods that involve reasoning capabilities (i.e., what is called the Semantic Web).

## 2.2 Scientific foundations

### 2.2.1 Fuzzy logic applied to databases

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 $\mu_F$ 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.

Since Lotfi Zadeh introduced fuzzy set theory in 1965, many applications of fuzzy logic to various domains of computer science have been achieved. As far as databases are concerned, the potential interest of fuzzy sets in this area has been identified as early as 1977, by V. Tahani [Tah77] — then a Ph.D. student supervised by L.A. Zadeh — who proposed a simple fuzzy query language extending SEQUEL. This first attempt was then followed by many researchers who strove to exploit fuzzy logic for giving database languages more expressiveness and flexibility. Then, in 1978, Zadeh coined possibility theory [Zad78], a model for dealing with uncertain information in a qualitative way, which also opened new perspectives in the area of uncertain databases. The pioneering work by Prade and Testemale [PT84] has had a rich posterity and the issue of modeling/querying uncertain databases in the framework of possibility theory is still an active topic of research nowadays. Beside these two main research lines, several other ways of exploiting fuzzy logic have been proposed along the years for dealing with various other aspects of data management, for instance *fuzzy data summaries*. More recently, fuzzy logic has also been applied — notably by the Shaman team — to model and query non-relational databases such as RDF databases or graph databases.

### 2.2.2 Ontology-based data management

Till the end of the 20$^{\text{th}}$ 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

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

[Tah77] V. Tahani, "A Conceptual Framework for Fuzzy Query Processing — A Step Toward Very Intelligent Database Systems", *Information Processing and Management 13*, 5, 1977, p. 289–303.

[Zad78] L. Zadeh, "Fuzzy Sets as a Basis for a Theory of Possibility", *Fuzzy Sets and Systems 1*, 1978, p. 3–28.

[PT84] H. Prade, C. Testemale, "Generalizing database relational algebra for the treatment of incompleteuncertain information and vague queries", *Information Sciences 34*, 1984, p. 115–143.

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 21st 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 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.

[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*.

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

### 2.2.3 Big Data management

Managing large volumes of data (with respect to the available resources) has been an important issue for decades. As an illustration, the first Very Large Data Bases (VLDB) conference was organized in 1975. Main contributions in the domain include parallel and distributed systems [DG92] with different approaches, in particular shared-nothing architectures [Sto86].

The deployment of large data centers consisting of thousand of commodity hardware-based node have lead to massively parallel processing systems. In particular, large scale distributed file system such as Google File System [GGL03], parallel processing paradigm/environment like MapReduce [DG08] have been the foundation of a new ecosystem with data management contributions in major conferences and journals on databases, such as VLDB, VLDBJ, SIGMOD, TODS, ICDE, IEEE DEB, ICDE and EDBT. Different (often open-source) systems have been provided such as Pig [ORS+08], Hive [TSJ+10] or more recently Spark [ZCD+12] and Flink [CKE+15], making it easier to use data centers resources for managing big data.

## 2.3 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

[DG92] D. J. DeWitt, J. Gray, "Parallel Database Systems: The Future of High Performance Database Systems", *Communications of the {ACM} 35*, 6, 1992, p. 85–98.

[Sto86] M. Stonebraker, "The Case for Shared Nothing", *IEEE Database Engineering Bulletin 9*, 1, 1986, p. 4–9.

[GGL03] S. Ghemawat, H. Gobioff, S.-T. Leung, "The Google file system", *in : Proceedings of the Symposium on Operating Systems Principles (SOSP)*, p. 29–43, Bolton Landing, NY, USA, 2003.

[DG08] J. Dean, S. Ghemawat, "MapReduce: simplified data processing on large clusters", *Communications of the ACM 51*, 1, 2008, p. 107–113.

[ORS+08] C. Olston, B. Reed, U. Srivastava, R. Kumar, A. Tomkins, "Pig latin: a not-so-foreign language for data processing", *in : Proceedings of the SIGMOD International Conference on Management of Data*, p. 1099–1110, Vancouver, BC, Canada, 2008.

[TSJ+10] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Anthony, H. Liu, R. Murthy, "Hive - a petabyte scale data warehouse using Hadoop", *in : Proceedings of the International Conference on Data Engineering ({ICDE})*, p. 996–1005, Long Beach, California, {USA}, 2010.

[ZCD+12] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker, I. Stoica, "Resilient Distributed Datasets: {A} Fault-Tolerant Abstraction for In-Memory Cluster Computing", *in : Proceedings of the {USENIX} Symposium on Networked Systems Design and Implementation (NSDI)*, p. 15–28, San Jose, CA, USA, 2012.

[CKE+15] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, K. Tzoumas, "Apache Flink{\texttrademark}: Stream and Batch Processing in a Single Engine", *{IEEE} Data Engineering Bulletin 38*, 4, 2015, p. 28–38.

descriptions (or goals) and vague terms, may provide a convenient means for expressing information needs.

As for uncertain/inconsistent data management, many potential domains could take advantage of advanced systems capable of storing and querying databases where some pieces of information are imperfect: 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 the following 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 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.

- Data journalism. Fact-checking is the task of assessing the factual accuracy of claims, typically prior to publication. Modern fact-checking is faced with a triple revolution in terms of scale, complexity, and visibility: many more claims are made and disseminated through Web and social media, they represent a complex reality and their investigation requires using multiple heterogeneous data source; finally, fact-checking outputs themselves are interesting for the public wishing to cross-check the process. The ANR project ContentCheck, in which Shaman participates, brings together academic labs with expertise in data management, natural language processing, automated reasoning and data mining, and a fact-checking team of journalists from a major French Web media. The aims are to establish fact-checking as a data management problem, endow it with sound foundations from the literature and/or new models as needed, design and deploy novel algorithms for automating fact-checking, and validate them by close interaction with the journalists.

- Cybersecurity. Security monitoring is one subdomain of cybersecurity. It aims to guarantee the safety of systems, continuously monitoring unusual events analyzing logs. The notion of a system in this context is very variable. It can actually be an information system in any institution or any device, like a laptop, a smartphone, a smartwatch, a vehicle (car, plane, etc.), a television, etc. Hence, the data to be managed with a high Velocity, are Voluminous with a high Variety. Security monitoring can thus be seen as a concrete use case of Big Data. Shaman is involved

in several projects related to security monitoring, in particular SERBER funded by the Pôle d'Excellence Cyber. One of the main goals is to provide a Big Data platform applied to security monitoring. This makes it mandatory to address several issues like efficient big fuzzy joins, data management with new hardware (FPGA) or optimization on encrypted data.

- Maritime transportation of goods. Shaman participates in the project CREDOC (2018 – 2021), founded by the EU and the region Britanny, whose objective is to conceive a solution for automating the controls performed by financiary institutions related to the maritime transportation of goods (an important partner in the project is the banking company HSBC). These controls aim to check i) the coherence between the data contained in the documents describing the transaction and those related to the effective path and transportation mode of the goods; ii) the conformity of the transport wrt. the rules of international trade (embargoed countries, piracy, etc.). For doing so, it is necessary to i) aggregate the data provided by different sources: maritime transportation companies, sites devoted to ship tracking, sites specialized in risk detection and fraud management, maritime weather forecast information, customs, etc.); ii) correlate all these data according to precise business rules in order to detect suspicious activities. The approach advocated by Shaman involves two steps; First, one needs to model complex fuzzy concepts based on the combination of different dimensions (e.g., a batch of containers may be considered *suspicious* if its rotation frequency is *high*, the loading intervals are *long*, and if they come from a company *under surveillance*). Then one needs to conceive knowledge discovery tools working on a unified representation of the data in the form of linguistic summaries.

# 3 Scientific achievements

## 3.1 Flexible database querying

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

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 [9], 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.

- *Fuzzy Query By Example.* In [17], we describe *Fuzzy Query By Example*, an approach helping users retrieve data without any prior knowledge of the database schema or any formal querying language. The user is solicited to evaluate, in a binary way, pre-selected items of the database. We provide a characterization-based strategy that identifies the properties shared by the examples (resp. counter-examples) positively (resp. negatively) evaluated by the user. These properties are expressed using linguistic terms from a fuzzy vocabulary to ensure that the user has a good understanding of the inferred query.

- *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 [22], we propose a fuzzy extension of the SPARQL language, named FUDGE, 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. [10] deals with *fuzzy quantified queries* in FUDGE. A processing strategy based on a compilation mechanism that derives regular (nonfuzzy) queries for accessing the relevant data is described. Some experiments are performed that show the tractability of this approach.

- *Fuzzy preference queries to NoSQL graph database.* Graph databases raise new challenges in terms of flexible querying since two aspects can be involved in the preferences that a user may express: i) the content of the nodes/edges and ii) the structure of the graph itself. In [6], we present a language, named FUDGE, that extends the well-known language Cypher so as to make it possible to express fuzzy preferences queries over graph databases.

### 3.2 Cooperative answering, data summarization

**Participants**: Pierre Nerzic, Olivier Pivert, Grégory Smits.

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.

- *Interactive data exploration on top of linguistic summaries.* The added value of a dataset lies in the knowledge a domain expert can extract from it. Considering the continuously increasing volume and velocity of these datasets, efficient tools have to be defined to generate meaningful, condensed and human-interpretable representations of big datasets. In [12], soft computing techniques are used to define an interface between the numerical and categorical space of data definition and the linguistic space of human reasoning. Based on the expert's own vocabulary about the data, a personal summary composed of linguistic terms is efficiently generated and graphically displayed as a term cloud offering a synthetic view of

the data properties. Using dedicated indexation strategies linking data and their subjective linguistic rewritings, exploration functionalities are then provided on top of the summary to let the user browse the data. Experimentations confirm that the space change operates in linear time wrt. the size of the dataset making the approach tractable on large scale data. The approach is also briefly described in [27].

- *Dissimilarity measures at the fuzzy partition level.* On the one hand, a user vocabulary is often used by soft-computing-based approaches to generate a linguistic and subjective description of numerical and categorical data. On the other hand, knowledge extraction strategies (as e.g. association rules discovery or clustering) may be applied to help the user understand the inner structure of the data. To apply knowledge extraction techniques on subjective and linguistic rewritings of the data, one first has to address the question of defining a dedicated distance metric. Many knowledge extraction techniques indeed rely on the use of a distance metric, whose properties have a strong impact on the relevance of the extracted knowledge. In [26, 25], we propose a measure that computes the dissimilarity between two items rewritten according to a user vocabulary.

- *Efficient generation of linguistic summaries.* Summarizing data with linguistic statements is a crucial and topical issue that has been largely addressed by the soft computing community. The goal of such summarization strategies is to generate statements that linguistically describe the properties observed in a dataset. In [23, 24], we address the issue of efficiently extracting these summaries and rendering them to the final user, in the case where the data to be summarized are stored in a relational database: it proposes a novel strategy that leverages the statistics about the data distribution maintained by the database system. We show that reliable summaries can be very efficiently estimated based on these statistics only and without any costly data access. Additionally, it proposes a visualization of the set of extracted summaries that offers a fruitful interactive exploration tool to the user. Experiments performed on two real data bases show the relevance and efficiency of the proposed approach: with a negligible loss of accuracy, we provide the first linguistic summarization approach whose processing time does not depend on the size of the dataset. The generation of estimated linguistic summaries takes less than one second even for dataset containing millions of tuples.

### 3.3 Ontology-based data management

**Participants**: François Goasdoué, Hélène Jaudoin.

- *Querying inconsistent description logic knowledge bases.* Several inconsistency-tolerant semantics have been introduced for querying inconsistent description logic knowledge bases. In [5], our first contribution is a practical approach for computing the query answers under three well-known such semantics, namely the AR, IAR and brave semantics, in the lightweight description logic DL-lite$_{\mathcal{R}}$. We show that query answering under the intractable AR semantics can be performed efficiently by using IAR and brave semantics as tractable approximations and encoding the

AR entailment problem as a propositional satisfiability (SAT) problem. Our second contribution is explaining why a tuple is a (non-)answer to a query under these 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 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 SAT, enabling us to generate explanations by exploiting solvers for Boolean satisfaction and optimization problems. Finally, we empirically study the efficiency of our query answering and explanation framework using a benchmark we built upon the well-established LUBM benchmark.

- *Summarization of RDF graphs.* The explosion in the amount of the available RDF data has lead to the need to explore, query and understand such data sources. Due to the complex structure of RDF graphs and their heterogeneity, the exploration and understanding tasks are significantly harder than in relational databases, where the schema can serve as a first step toward understanding the structure. Summarization has been applied to RDF data to facilitate these tasks. Its purpose is to extract concise and meaningful information from RDF knowledge bases, representing their content as faithfully as possible. There is no single concept of RDF summary, and not a single but many approaches to build such summaries; each is better suited for some uses, and each presents specific challenges with respect to its construction. The survey [7] is the first to provide a comprehensive survey of summarization method for semantic RDF graphs. We propose a taxonomy of existing works in this area, including also some closely related works developed prior to the adoption of RDF in the data management community; we present the concepts at the core of each approach and outline their main technical aspects and implementation. We hope the survey will help readers understand this scientifically rich area, and identify the most pertinent summarization method for a variety of usage scenarios.

- *Exploration of RDF graphs.* The Web of Data is growing fast, as exemplified by the evolution of the Linked Open Data (LOD) cloud over the last ten years. One of the consequences of this growth is that it is becoming increasingly difficult for application developers and end-users to find the datasets that would be relevant to them. Semantic Web search engines, open data catalogs, datasets and frameworks such as LODStats and LOD Laundromat, are all useful but only give partial, even if complementary, views on what datasets are available on the Web. In [19], we introduce LODAtlas, a portal that enables users to find datasets of interest. Users can make different types of queries about both the datasets' metadata and contents, aggregated from multiple sources. They can then quickly evaluate the matching datasets' relevance, thanks to LODAtlas' summary visualizations of their general metadata, connections and contents.

- *RDF integration of data sources.* In [14], we propose an integration architecture to access data from remote heterogeneous data sources as an RDF graph; we study in this setting the problem of computing the certain query answers. Existing approaches to query answering in the presence of knowledge (expressed here

as an RDFS ontology and entailment rules) involve either the materialization of inferences in the data or the reformulation of the query. Both approaches have well-known drawbacks. We introduce a new approach to query answering, based on a reduction to view-based query answering. This approach avoids both materialization in the data and query reformulation. We define restrictions of our general architecture under which our method is correct, and formally prove its correctness.

### 3.4 Big data management

**Participants**: Laurent D'Orazio.

- *Improving Hamming distance-based fuzzy join in MapReduce using Bloom Filters.* Join operation is one of the key ones in databases, allowing to cross data from several tables. Two tuples are crossed when they share the same value on some attribute(s). A fuzzy or similarity join combines all pairs of tuples for which the distance is lower than or equal to a pre-specified threshold epsilon from one or several relations. Fuzzy join has been studied by many researchers because its practical application. However, join is the most costly and may even not be possible to compute on large databases. In [28], we thus propose the optimization for MapReduce algorithms to process fuzzy joins of binary strings using Hamming Distance. In particular we propose to use an extension of Bloom Filters to eliminate the redundant data, reduce the unnecessary comparisons, and avoid the duplicate output. We compare and evaluate analytically the algorithms with a cost model.

- *Semantic caching framework, an application to FPGA-based application for IoT security monitoring.* Security monitoring is one subdomain of cybersecurity which aims to guarantee the safety of systems, continuously monitoring unusual events. The development of Internet Of Things leads to huge amounts of information, being heterogeneous and requiring to be efficiently managed. Cloud Computing provides software and hardware resources for large scale data management. However, performances for sequences of on-line queries on long term historical data may be not compatible with the emergency security monitoring. In [8] we address this problem by proposing a semantic caching framework and its application to acceleration hardware with FPGA for fast- and accurate-enough logs processing for various data stores and execution engines.

- *A Method to build a Geolocalized Food Price Time Series Knowledge Base analyzable by Everyone.* Time-series analysis is a very challenging concept in Data Science for companies and industries. Harvesting prices of agricultural production (e.g. vegetable, fruit, milk...) as time series is key to operating reliable dish cost prediction at scale to ensure for example that the market price is valid. In [18], we describe initial stakeholder needs, the service and engineering contexts in which the challenge of time-serie harvesting and management arose, and theoretical and architectural choices we made to implement a solution of historical food

prices. For this, we use scrappers through the TOR network. We also propose a knowledge map approach to make the data accessible to any type of users.

- *NSGA-G: a genetic algorithm for cloud computing* Cloud computing provides computing resources with elasticity following a pay-as-you-go model. This raises Multi-Objective Optimization Problems (MOOP), in particular to find Query Execution Plans (QEPs) with respect to users' preferences being for example response time, money, quality, etc. In such a context, MOOP may generate Pareto-optimal fronts with high complexity. Pareto-dominated based Multiple Objective Evolutionary Algorithms (MOEA) are often used as an alternative solution, like Non-dominated Sorting Genetic Algorithms (NSGAs) that provide better computational complexity. In [16], we present NSGA-G, a NSGA based on Grid Partitioning for improving complexity and quality of current NSGAs. Experiments on DTLZ test problems using Generational Distance (GD), Inverted Generational Distance (IGD) and Maximum Pareto Front Error prove the relevance of our solution.

- *Adaptive Time, Monetary Cost Aware Query Optimization on Cloud DataBase* Most of the existing database query optimization techniques are designed targeting the traditional database systems, and the objective of optimization is one-dimensional. These techniques usually aim to reduce either the query response time or the I/O cost of the query. Evidently, these optimization algorithms are not suitable for cloud database systems because these systems are provided to users as on-demand services where users are charged for the actual usage of the services. In this case, users will take both query response time and monetary cost to be paid to the cloud service providers into consideration for selecting a database system product. Thus, query optimization for cloud database systems needs to target reducing the monetary cost in addition to query response time. This means that query optimization has multiple objectives which are more challenging than that for traditional database systems. Similar problems exist when incorporating query re-optimization into the query execution process to obtain more accurate cost estimates when considering multiple objectives. In [29], we present a query optimization method that achieves two goals: 1) identifying a query execution plan that satisfies the multi-objectives provided by the user; and 2) reducing the costs of running the query execution plan by performing adaptive query re-optimization during the query execution. The experimental results show that the proposed method performs can save either the time cost or the monetary cost based on the types of queries.

### 3.5 Data quality and uncertain data management

**Participants**: Olivier Pivert, Grégory Smits, Virginie Thion.

- *Handling uncertainty in relational databases with possibility theory.* Mainstream approaches to uncertainty modeling in databases are probabilistic. Still some researchers persist in proposing representations based on possibility theory. They

are motivated by the ability of this latter setting for modeling epistemic uncertainty and by its qualitative nature. Interestingly enough, several possibilistic models have been proposed over time, and have been motivated by different application needs ranging from database querying, to database design and to data cleaning. The surveys [20, 21, 11] distinguish between four different frameworks ordered according to an increasing representation power: databases with i) layered tuples; ii) certainty-qualified attribute values; iii) attribute values restricted by general possibility distributions; iv) possibilistic c-tables. In each case, we discuss the role of the possibility-necessity duality, the limitations and the benefit of the representation settings, and their suitability with respect to different tasks.

- *Data quality in Digital Libraries of Scores.* Much published data suffers from quality problems [ZRM+16]. It is now well-recognised that these endemic problems may lead to severe consequences, and that managing the quality of data conditions the success of most existing information systems. In our research, we deal with the quality management of open data stored in digital libraries of scores. The first and fundamental step of the data quality management process consists in eliciting data quality requirements. Because data quality is defined as being the fitness for use of data (meaning that the notion of data quality depends on the context), it is a conceptually complex notion, whose implementation for a given use case is not trivial. Context-dependant guidelines are needed in order to help users to define data quality in their context. This is the problem that we tackle in [30], by proposing a set of quality rules specific to DSLs, which can serve as a basis in order to elaborate users' quality requirements. In [13] and [15], we consider the problem of implementing them in a real digital library of scores, which is managed by computer scientists and musicologists.

# 4 Software development

## 4.1 PostgreSQLF

**Participants**: Olivier Pivert, Grégory Smits.

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 that makes it easy for an end user to specify his/her fuzzy queries.

## 4.2 Ikeys

**Participants**: Olivier Pivert, Grégory Smits.

[ZRM+16] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, "Quality assessment for Linked Data: A Survey", *Semantic Web 7*, 1, 2016, p. 63–93, http://dx.doi.org/10.3233/SW-150175.

IKEYS 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").

### 4.3 Fudge/Sugar

**Participants**: Olivier Pivert, Virginie Thion.

FUDGE 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 is a prototype, based on Neo4j, implementing the FUDGE language. More information can be found here: `https://www-shaman.irisa.fr/fudge-prototype/`.

### 4.4 OntoSQL

**Participants**: François Goasdoué.

ONTOSQL is a Java-based tool that provides two main functionalities: (i) loading RDF graphs (consisting of RDF assertions and possibly an RDF Schema) into a relational database; the data is integer-encoded and indexed; (ii) querying the loaded RDF graphs through conjunctive SPARQL queries, a.k.a. basic graph pattern queries. ONTOSQL not only evaluates queries, it answers them, that is: its answers accounts for both the data explicitly present in the database, as well as the implicit data begotten by the ontology knowledge. To this aim, ONTOSQL supports both materialization (aka saturation), and reformulation-based query answering.

### 4.5 RDFQuotient

**Participants**: François Goasdoué.

RDFQUOTIENT is a Java-based tool that allows summarizing RDF graphs for first-sight visualization. It gives users as much information as possible about the graph structure, without requiring them to provide an input or tune parameters. In particular, the summarization technique of RDFQUOTIENT builds on graph quotients. It uses a novel graph node equivalence relation based on the transitive cooccurrence of edges, particularly suited to the high and meaningful compression of heterogeneous graphs.

# 5 Contracts and collaborations

## 5.1 National Initiatives

### 5.1.1 ODIN

**Participants**: François Goasdoué, Hélène Jaudoin, Olivier Pivert, Grégory Smits, Virginie Thion.

The DGA project ODIN (Open Data INtelligence) (Nov. 2014 – Oct. 2018) 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. The other partners involved are Semsoft and INRIA Saclay.

## 5.2 ContentCheck

**Participants**: François Goasdoué.

The ANR project ContentCheck brings together experts in data management, natural language processing, automated reasoning and data mining from Inria, LIMSI/CNRS and U. Paris Saclay, Univ. Rennes 1, Univ. Lyon 1, and the fact-checking team "Les Décodeurs" from Le Monde, the leading French national newspaper. The aim of the project is to design and deploy novel algorithms for automating fact-checking, and validate them by close interaction with the journalists.

## 5.3 iCoda

**Participants**: François Goasdoué.

The INRIA Project Lab iCoda — Knowledge-mediated Content and Data Analytics (2017–2020) — gathers INRIA Montpellier (Graphik), INRIA Saclay (Cedar & Ilda), INRIA/IRISA Rennes (LinkMedia & Shaman), as well as AFP, Ouest France and Le Monde. The goal of this project is the design of algorithms that allow analysts to efficiently infer useful information and knowledge by collaboratively inspecting heterogeneous information sources, from structured data to unstructured content, taking data journalism as an emblematic use-case.

## 5.4 GioQoso

**Participants**: Virginie Thion.

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).

### 5.5 Collaborations

Ronald R. Yager, from the Machine Intelligence Institute (Iona College, New Rochelle, NY, USA) is a pioneer in the domain of fuzzy logic, and one of the most prolific researchers in this area. Grégory Smits visited him in Jan. 2018, and we started a collaboration with Prof. Yager, whose goal is to define a soft computing approach to Big Data summarization.

## 6 Dissemination

### 6.1 Promoting scientific activities

#### 6.1.1 Scientific Events Selection

**Member of Conference Program Committees**

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

- 20$^{th}$ International Conference on Big Data Analytics and Knowledge Discovery (DaWaK), 2018
- 1$^{st}$ International Workshop on Data Engineering meets Intelligent Food and COoking Recipe(DECOR@ICDE), 2018
- European Conference on Information Systems (ECIS), 2018
- 2$^{nd}$ International Conference on Computer Science and Application Engineering (CSAE), 2018
- 3$^{rd}$ International Conference on Intelligent Information Technologies (ICIIT), 2018
- Journées Francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA), 2018
- 15$^{th}$ Colloque sur l'Optimisation et les Systèmes d'Information (COSI), 2018
- International Workshop on Information Search, Integration and Personalization (ISIP) 2018

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

- 37$^{th}$ ACM Symposium on Principles Of Database Systems (PODS), 2018
- 23$^{rd}$ European Conference on Artificial Intelligence (ECAI) & 27$^{th}$ International Joint Conference on Artificial Intelligence (IJCAI), 2018
- 44$^{th}$ International Conference on Very Large Data Bases (PVLDB), 2018

- 34th Journées Bases de Données Avancées (BDA), 2018.

Ludovic Liétard served as a member of the following program committees:

- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2018), Arras, France, November 8-9, 2018.

Olivier Pivert served as a member of the following program committees:

- 33rd ACM Symposium on Applied Computing (SAC 2018), Pau, France, April 9-13, 2018;
- 19th International Conference on Web Information Systems Engineering (WISE 2018), Dubai, United Arab Emirates, November 12-15, 2018;
- 12th International Conference on Scalable Uncertainty Management (SUM 2018), Milan, Italy, October 3-5, 2018;
- 24th International Symposium on Methodologies for Intelligent Systems (ISMIS 2018), Limassol, Cyprus, October 29-31, 2018;
- 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2018), Cádiz, Spain, June 11-15, 2018;
- 36e Conférence INFORSID, Nantes, France, May 28-31, 2018;
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2018), Arras, France, November 8-9, 2018.

Daniel Rocacher served as a member of the following program committee:

- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2018), Arras, France, November 8-9, 2018.

Grégory Smits served as a member of the following program committees:

- 24th International Symposium on Methodologies for Intelligent Systems (ISMIS 2018), Limassol, Cyprus, October 29-31, 2018;
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2018), Arras, France, November 8-9, 2018.

Virginie Thion served as a member of the following program committee:

- Workshop Qualité des Données du Web (QLOD), in conjunction with the 18th Conference on Knowledge Extraction and Management (EGC), Paris, France, January 22-26, 2018.

**Reviewer**

Hélène Jaudoin served as an external reviewer for:

- the 33rd ACM Symposium on Applied Computing (SAC 2018), Pau, France, April 9-13, 2018;
- the 34th Journées Bases de Données Avancées (BDA), 2018.

### 6.1.2 Journal

**Member of 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,
- Ingénierie des Systèmes d'Information.

**Reviewer - Reviewing Activities**

- François Goasdoué reviewed for Information Systems; Semantic Web Journal; Revue d'Intelligence Artificielle.
- Hélène Jaudoin reviewed for Fuzzy Sets and Systems.
- Olivier Pivert reviewed for the Journal of Official Statistics; IEEE Transactions on Fuzzy Systems; the International Journal of Fuzziness, Uncertainty and Knowledge-Based Systems; Technique et Science Informatiques.
- Laurent d'Orazio reviewed for IEEE Transactions on Parallel and Distributed Systems (TPDS), International Journal of Information Technology and Decision Making (IJITDM), International Journal of Approximate Reasoning (IJA).

### 6.1.3 Invited Talks

Grégory Smits gave an invited talk about "Fuzzy Querying: From Theory to pratice" at the 12th International Conference on Scalable Uncertainty Management (SUM 2018), that took place in Milan (Italy), from Oct. 3 to Oct. 5, 2018.

Laurent d'Orazio gave a tutorial on "Big Data Management" at École Thématique BDA (MDD 2018), that took place in Aussois, from June 17 to June 22, 2018.

### 6.1.4 Leadership within the Scientific Community

Olivier Pivert is a member of the permanent steering committees of

- the French-speaking conference "Rencontres Francophones sur la Logique Floue et ses Applications" (LFA);

- the International Symposium on Methodologies for Intelligent Systems (ISMIS);
- the International Conference on Flexible Query-Answering Systems (FQAS).

#### 6.1.5 Scientific Expertise

Olivier Pivert is an expert for the Czech Science Foundation.

Laurent d'Orazio has been asked to review proposals by ANR and STIC AmSud.

#### 6.1.6 Research Administration

François Goasdoué is a member of the Scientific Advisory Committee of IRISA UMR 6074.

### 6.2 Teaching, supervision

#### 6.2.1 Teaching

Several members of the Shaman team give courses in the Enssat track of the Master's degree curriculum in Computer Science at University of Rennes 1: Olivier Pivert and Grégory Smits teach a course about *Advanced Databases*, Hélène Jaudoin teaches a part of the course on *Machine Learning*, and François Goasdoué and Hélène Jaudoin teach a course on *Web data Management*.

#### 6.2.2 Supervision

- Ph.D.: Sara El Hassad, Learning Commonalities in RDF and SPARQL, defended Feb. 2, 2018, François Goasdoué and Hélène Jaudoin;
- Ph.D.: Aurélien Moreau, How Fuzzy Set Theory Can Help Make Database Systems More Cooperative, defended June 29, 2018, Olivier Pivert and Grégory Smits;
- Ph.D. in progress: Maxime Buron, Efficient reasoning on heterogeneous large-scale graphs, started Oct. 2017, François Goasdoué and Ioana Manolescu (INRIA/Cedar) and Marie-Laure Mugnier (LIRMM/GraphIK);
- Ph.D. in progress: Le Trung Dung, Data Management in cloud federation, started Sep. 2016, Laurent D'Orazio and Verena Kantere (Univ. Ottawa, Canada);
- Ph.D. in progress: Ludivine Duroyon, Data management models, algorithms and tools for fact-checking, started Oct. 2017, François Goasdoué and Ioana Manolescu (INRIA/Cedar);
- PhD in progress: Cheikh Brahim El Vaigh, Incremental content to data linking leveraging ontological knowledge in data journalism, started Oct. 2017, François Goasdoué, Guillaume Gravier (IRISA/LinkMedia) and Pascale Sébillot (IRISA/LinkMedia);
- Ph.D. in progress: Thi To Quyen, Filter-based fuzzy big joins, started Oct. 2017, Laurent D'Orazio, Anne Laurent (LIRMM/Fado)) and Thuong Cang Phan (Can Tho Univ., Vietnam);

- Ph.D. in progress: Van Hoang Tran, Encrypted big log management, started Dec. 2017, Laurent D'Orazio, Tristan Allard (IRISA/Druid) and Amr El Abbadi (Univ. of California Santa Barbara, USA);

#### 6.2.3 Juries

Laurent D'Orazio

- Ph.D., reviewer, Amadou Fall Dia, Sorbonne Université (CNAM-ISEP)
- Ph.D., president, David Pierrot, Université de Lyon 2
- Ph.D., president, Kemp Gavin, Université de Lyon 1

François Goasdoué

- HDR, reviewer, Nicolas Travers, Sorbonne Université (CNAM)
- Ph.D., president, Amadou Fall Dia, Sorbonne Université (CNAM-ISEP)

Olivier Pivert

- Ph.D., member (supervisor), Aurélien Moreau, Univ. Rennes 1
- HDR, member, Grégory Smits, Univ. Rennes 1

### 6.3 Popularization

## 7 Bibliography

### Major publications by the team in recent years

[1] 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.

[2] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Computing and Explaining Query Answers over Inconsistent DL-Lite Knowledge Bases", *Journal of Artificial Intelligence Reasearch (JAIR)*, 2018, to appear.

[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] F. Goasdoué, Z. Kaoudi, I. Manolescu, J. Quiané-Ruiz, S. Zampetakis, "CliqueSquare: Flat plans for massively parallel RDF queries", *in: 31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, 2015.

[6] S. E. Hassad, F. Goasdoué, H. Jaudoin, "Learning Commonalities in SPARQL", *in : 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I*, 2017.

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

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

[9] 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.

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

## Books and Monographs

[1] O. Pivert (editor), *NoQSL Data Models — Trends and Challenges*, ISTE, London, UK, 2018.

## Doctoral dissertations and "Habilitation" theses

[2] S. El Hassad, *Learning Commonalities in RDF & SPARQL*, PhD Thesis, University of Rennes 1 – École doctorale MathSTIC, February 2, 2018, supervised by F. Goasdoué and H. Jaudoin.

[3] A. Moreau, *How Fuzzy Set Theory can Help Make Database Systems More Cooperative*, PhD Thesis, University of Rennes 1 – École doctorale MathSTIC, June 29, 2018, supervised by O. Pivert and G. Smits.

[4] G. Smits, *Personnalisation et enrichissement des méthodes d'accès aux données*, Habilitation à diriger des recherches en informatique, University of Rennes 1, March 14, 2018.

## Articles in referred journals and book chapters

[5] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Computing and Explaining Query Answers over Inconsistent DL-Lite Knowledge Bases", *Journal of Artificial Intelligence Reasearch (JAIR)*, 2018, to appear.

[6] A. Castelltort, A. Laurent, O. Pivert, O. Slama, V. Thion, "Fuzzy Preference Queries to NoSQL Graph Databases", *in : NoQSL Data Models — Trends and Challenges*, O. Pivert (editor), ISTE, London, UK, 2018, p. 167–201.

[7] S. Cebiric, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, M. Zneika, "Summarizing Semantic Graphs: A Survey", *VLDB J. 27*, 2018.

[8] L. d'Orazio, J. Lallet, "Semantic caching framework, an application to FPGA-based application for IoT security monitoring", *Open Journal of Internet of Things (OJIOT)*, 2018, `https://hal.archives-ouvertes.fr/hal-01857359`.

[9] P. Nerzic, H. Jaudoin, O. Pivert, "Parallel Processing Strategies for Skyline Queries Tolerant to Outliers", *International Journal of Intelligent Systems 33*, 10, 2018, p. 1992–2018.

[10] O. Pivert, O. Slama, V. Thion, "Fuzzy Quantified Queries to Fuzzy Graph Databases", *Fuzzy Sets and Systems*, 2018, to appear.

[11] O. Pivert, G. Smits, "Fuzzy Extensions of Databases", *in : A Fuzzy Dictionary of Fuzzy Modelling. Common Concepts and Perspectives*, C. Marsala and M.-J. Lesot (editors), Springer, 2018, to appear.

[12] G. Smits, O. Pivert, R. Yager, P. Nerzic, "A Soft Computing Approach to Big Data Summarization", *Fuzzy Sets and Systems 348*, 2018, p. 4–20.

## Publications in Conferences and Workshops

[13] V. Besson, D. Fiala, P. Rigaux, V. Thion, "Gioqoso, an Online Quality Evaluation Tool for MEI Scores", *in : Proc. of the Music Encoding Conference 2018 (MEC 2018)*, University of Maryland, Washington DC, USA., May 2018, `https://hal.archives-ouvertes.fr/hal-01708859`.

[14] M. Buron, F. Goasdoué, I. Manolescu, M.-L. Mugnier, "Rewriting-Based Query Answering for Semantic Data Integration Systems", *in : 34ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA 2018)*, Bucarest, Romania, October 2018, `https://hal.archives-ouvertes.fr/hal-01927282`.

[15] F. Foscarin, D. Fiala, F. Jacquemard, P. Rigaux, V. Thion, "Gioqoso, an online Quality Assessment Tool for Music Notation", *in : 4th International Conference on Technologies for Music Notation and Representation (TENOR'18)*, 2018. Poster.

[16] T.-D. Le, V. Kantere, L. d'Orazio, "An efficient multi-objective genetic algorithm for cloud computing: NSGA-G", *in : International Workshop on Benchmarking, Performance Tuning and Optimization for Big Data Applications (BPOD)*, Seattle, United States, December 2018, `https://hal.archives-ouvertes.fr/hal-01962235`.

[17] A. Moreau, O. Pivert, G. Smits, "Fuzzy Query by Example", *in : Proc. of the 33rd ACM Symposium on Applied Computing (SAC'18)*, p. 688–695, Pau, France, 2018.

[18] J. Papin, F. Andrès, L. d'Orazio, "A Method to build a Geolocalized Food Price Time Series Knowledge Base analyzable by Everyone", *in : Latin America Data Science Workshop (LADaS@VLDB)*, Rio de Janeiro, Brazil, 2018, `https://hal.archives-ouvertes.fr/hal-01857388`.

[19] E. Pietriga, H. Gozukan, C. Appert, M. Destandau, Š. Čebirić, F. Goasdoué, I. Manolescu, "Browsing Linked Data Catalogs with LODAtlas", *in : International Semantic Web Conference (ISWC)*, 2018.

[20] O. Pivert, H. Prade, "Handling Uncertainty in Relational Databases with Possibility Theory – A Survey of Different Modelings", *in : Proc. of the 12th International Conference on Scalable Uncertainty Management (SUM'18)*, Milan, Italy, 2018.

[21] O. Pivert, H. Prade, "Modèles possibilistes de bases de données incertaines – Une vue d'ensemble", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'18)*, Arras, France, 2018.

[22] O. Pivert, O. Slama, V. Thion, "FURQL: une extension floue de SPARQL", *in : Actes du 36e Congrès INFORSID*, p. 169–178, Nantes, France, 2018.

[23] G. Smits, P. Nerzic, O. Pivert, M.-J. Lesot, "Efficient Generation of Reliable Estimated Linguistic Summaries", *in : Proc. of the 27th IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'18)*, Rio de Janeiro, Brazil, 2018. ** Best Paper Award **.

[24] G. Smits, P. Nerzic, O. Pivert, M.-J. Lesot, "Génération efficace de résumés linguistiques estimés", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'18)*, Arras, France, 2018.

[25] G. Smits, O. Pivert, T. Ngoc Duong, "Mesurer la dissimilarité au niveau d'une partition floue", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'18)*, Arras, France, 2018.

[26] G. Smits, O. Pivert, T. Ngoc Duong, "On Dissimilarity Measures at the Fuzzy Partition Level", *in : Proc. of the 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'18)*, p. 301–312, Cádiz, Spain, 2018.

[27] G. Smits, O. Pivert, "Exploration de résumés personnalisés de données", *in : Actes de la 18e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'18) – Atelier Visualisation d'informations, Interactions, et Fouille de Données*, Paris, France, 2018.

[28] T. Tran, T. Phan, A. Laurent, L. d'Orazio, "Improving Hamming distance-based fuzzy join in MapReduce using Bloom Filters", *in : FUZZ-IEEE 2018: International Conference on Fuzzy Systems*, Rio de Janeiro, Brazil, 2018, `https://hal.archives-ouvertes.fr/hal-01857386`.

[29] C. Wang, Z. Arani, L. Gruenwald, L. d'Orazio, "Adaptive Time, Monetary Cost Aware Query Optimization on Cloud Database Systems", *in : International workshop on Scalable Cloud Data Management (SCDM@BigData)*, Seattle, United States, December 2018, `https://hal.archives-ouvertes.fr/hal-01962218`.

## Internal Reports

[30] D. Fiala, P. Rigaux, A. Tacaille, V. Thion, m. o. The Gioqoso project, "Data Quality Rules for Digital Score Libraries", *Research report*, IRISA, Université de Rennes, 2018, `https://hal.inria.fr/hal-01734821`.