UMR IRISA

Activity Report 2019

Team SHAMAN

A Symbolic and Human-Centric View of Data Management

D7 – Data and Knowledge Management

1 Team composition

Researchers and faculty

François Goasdoué, Professor, ENSSAT, head of the team
Hélène Jaudoin, Associate Professor, ENSSAT
Ludovic Liétard, Associate Professor, HDR, IUT Lannion
Pierre Nerzic, Associate Professor, IUT Lannion
Laurent d'Orazio, Professor, IUT Lannion
Olivier Pivert, Professor, ENSSAT
Daniel Rocacher, Professor, ENSSAT
Amit Shukla, Postdoc from March 2019 to September 2020
Grégory Smits, Associate Professor, HDR, IUT Lannion
Virginie Thion, Associate Professor, HDR, ENSSAT

Non-permanent members

Maxime Buron, PhD student, IPL iCoda grant, since Oct. 2017
Trung Dung Le, PhD student then Engineer; in Shaman since Sep. 2016,
Ludivine Duroyon, PhD student, ANR ContentCheck grant, since Oct. 2017
Cheikh-Brahim El Vaigh, PhD student, IPL iCoda, since Oct. 2017
Mohamed Handaoui, PhD student, B-Com, since october 2019
Thi To Quyen, PhD student, Vietnam gov. grant MOET 911, since Oct. 2017
Van Hoang Tran, PhD student, PEC grant, since Dec. 2017

Administrative assistants

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 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 nodes have led to massively parallel processing systems. In particular, large scale distributed file systems such as Google File System [GGL03], parallel processing paradigm/environment like MapReduce [DG08] have been the foundations 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 center resources for managing big data.

2.2.2 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

[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.

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

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.

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.

2.2.3 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 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*

[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.

[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] 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.

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

[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*.

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/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.

We currently 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 ContentCheck (2015-2020) and IPL iCoda (2017-2020) projects, 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 at guaranteeing the safety of systems, continuously monitoring unusual events by analyzing logs. The notion of a system in this context is very variable. It can actually be an information system in any organization 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 Big data management

Participants: Laurent d'Orazio.

- *PINED-RQ++.* Privacy is a major concern in cloud computing since clouds are considered as untrusted environments. In [19], we address the problem of privacy-preserving range query processing on clouds. Several solutions have been proposed in this line of work, however, they become inefficient or impractical for many monitoring applications, including real-time monitoring and predicting the spatial spread of seasonal epidemics (e.g., H1N1 influenza). In this case, a system often confronts a high rate of incoming data. Prior schemes may thus suffer from potential performance issues, e.g., overload or bottleneck. We introduce an extension of PINED-RQ to address these limitations. We also demonstrate experimentally that our solution outperforms PINED-RQ.

- *NSGA-G.* Cloud-based systems enable to manage ever-increasing medical data. The Digital Imaging and Communication in Medicine (DICOM) standard has

been widely accepted to store and transfer the medical data, which uses single (row/column) or hybrid data storage technique (row-column). In particular, hybrid systems leverage the advantages of both techniques and allow to take into account various kinds of queries from full records retrieval (online transaction processing) to analytics (online analytical processing) queries. Additionally, the pay-as-you-go model and elasticity of cloud computing raise an important issue regarding to Multiple Objective Optimization (MOO) to find a data configuration according to users preferences such as storage space, process-ing response time, monetary cost, quality, etc. In such a context, the considerable space of solutions in MOO leads to generation of Pareto-optimal front with high complexity. Pareto-dominated based Multiple Objective Evolutionary Algorithms are often used as an alternative solution, e.g., Non-dominated Sorting GeneticAlgorithms (NSGA) which provide less computational complexity. [15] presents NSGA-G, an NSGA based on Grid Partitioning to improve the complexity and quality of current NSGAs and to obtain efficient storage and querying of DICOM hybrid data.Experimental results on DTLZ test problems and DICOM hybrid data prove the relevance of the proposed algorithm

- *DREAM.* Data sharing is important in the medical domain. Sharing data allows large-scale analysis with many data sources to provide more accurate results (especially in the case of rare diseases with small local datasets). Cloud federations consist in a major progress in sharing medical data stored within different cloud platforms, such as Amazon, Microsoft, Google Cloud, etc. It also enables to access distributed data of mobile patients. The pay-as-you-go model in cloud federations raises an important issue in terms of Multi-Objective Query Processing (MOQP) to find a Query ExecutionPlan according to users preferences, such as response time, money, quality, etc. However, optimizing a query in a cloud federation is complex with increasing heterogeneity and additional variance, especially due to a wide range of communications and pricing models. Indeed, in such a context, it is difficult to provide accurate estimation to make relevant decision. To address this problem, we present Dynamic Regression Algorithm (DREAM) [14], which can provide accurate estimation in a cloud federation with limited historical data. DREAM focuses on reducing the size of historical data while maintaining the estimation accuracy. The proposed algorithm is integrated in Intelligent Resource Scheduler, a solution for heterogeneous databases, to solve MOQP in cloud federations and validate with preliminary experiments on a decision support benchmark (TPC-H benchmark).

- *ML-based re-optimization.* Many of the existing cloud database query optimization algorithms target reducing the monetary cost paid to cloud service providers in addition to query response time. These query optimization algorithms rely on an accurate cost estimation so that the optimal query execution plan (QEP) is selected. The cloud environment is dynamic, meaning the hardware configuration, data usage, and workload allocations are continuously changing. These dynamic changes make an accurate query cost estimation difficult to obtain. Concurrently, the query execution plan must be adjusted automatically to address these changes. In order to optimize the QEP with a more accurate cost estimation, the query needs to be optimized multiple times during execution. On top of this, the most

updated estimation should be used for each optimization. However, issues arise when deciding to pause the execution for minimum overhead. In [20], we present our vision of a method that uses machine learning techniques to predict the best timings for optimization during execution.

- *Complex Value Relations in Hive.* In [17], we raise the question "how data architects model their data for processing in Apache Hive?". This well-known SQL-on-Hadoop engine supports complex value relations, where attribute types need not be atomic. In fact, this feature seems to be one of the prominent selling points, e.g., in Hive reference books. In an empirical study, we analyze Hive schemas in open source repositories. We examine to which extent practitioners make use of complex value relations and accordingly, whether they write queries over complex types. Understanding which features are actively used will help make the right decisions in setting up benchmarks for SQL-on-Hadoop engines, as well as in choosing which query operators to optimize for.

3.2 Flexible, cooperative and quality-aware data management

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

- *Answer characterization.* In [6], 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.

- *Fuzzy Querying of graph databases.* We have recently proposed 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. [7] 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.

- *Linguistic summarization of data* In [18, 16], we describe an approach that makes it possible to efficiently summarize large datasets stored in relational databases, by leveraging some statistics about data distributions that are maintained by any RDBMS. Summaries take the form of fuzzy quantified statements that involve some linguistic terms from a user-defined vocabulary.

- *Processing fuzzy relational queries using fuzzy views.* In [10], we propose two original approaches to the processing of fuzzy queries in a relational database context. The general idea is to use views, either materialized or not. In the first case, materialized views are used to store the satisfaction degrees related to user-defined fuzzy predicates, instead of calculating them at runtime by means of user functions embedded in the query (which induces an important overhead). In the second case, abstract views are used to efficiently access the tuples that belong to the α-cut of the query result, by means of a derived Boolean selection condition.

3.3 Ontology-based data management

Participants: François Goasdoué, Hélène Jaudoin.

- *Querying RDF graphs.* Query answering in RDF knowledge bases has traditionally been performed either through graph saturation, i.e., adding all implicit triples to the graph, or through query reformulation, i.e., modifying the query to look for the explicit triples entailing precisely what the original query asks for. The most expressive fragment of RDF for which Reformulation-based query answering exists is the so-called database fragment, in which implicit triples are restricted to those entailed using an RDF Schema (RDFS) ontology. Within this fragment, query answering was so far limited to the interrogation of data triples (non-RDFS ones); however, a powerful feature specific to RDF is the ability to query data and schema triples together. In [8], we address the general query answering problem by reducing it, through a pre-query reformulation step, to that solved by a state of the art query reformulation technique. We also report on experiments demonstrating the low cost of our reformulation algorithm.

- *Querying inconsistent description logic knowledge bases.* Several inconsistency-tolerant semantics have been introduced for querying inconsistent description logic knowledge bases. In [2], 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.

- *Survey on RDF graph summarization.* 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 [4] 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.

- *Summarization of RDF graphs.* Realizing the full potential of Linked Open Data sharing and reuse is currently limited by the difficulty users have when trying to understand the data modelled within an RDF graph, in order to determine whether or not it may be useful for their need. In [13], we demonstrate our RDFQuotient tool, which builds compact summaries of heterogeneous RDF graphs for the purpose of first-sight visualizations. An RDFQuotient summary provides an overview of the complete structure of an RDF graph, while being typically many orders of magnitude smaller, thus can be easily grasped by new users. Our summarization algorithms are time linear in the size of the input graph and incremental: they incrementally update a summary upon addition of new data. For the demo, we plan to show the visualizations of our summaries obtained from well-known synthetic and real data sets. Further, attendees will be able to add data to the summarized RDF graphs and visually witness the incurred changes.

- *RDF data management for data journalism.* A frequent journalistic fact-checking scenario is concerned with the analysis of statements made by individuals, whether in public or in private contexts, and the propagation of information and hearsay ("who said/knew what when"). In [11, 9], inspired by our collaboration with fact-checking journalists from Le Monde – France's leading newspaper –, we describe and demonstrate a Linked Data (RDF) model, endowed with formal foundations and semantics, for describing facts, statements, and beliefs. Our model combines temporal and belief dimensions to trace propagation of knowledge between agents along time, and can answer a large variety of interesting questions through RDF query evaluation. A preliminary feasibility study of our model incarnated in a corpus of tweets demonstrates its practical interest.

- *Knowledge-based entity linking.* Entity linking is a core task in textual document processing, which consists in identifying the entities of a knowledge base (KB) that are mentioned in a text. Approaches in the literature consider either independent linking of individual mentions or collective linking of all mentions. Regardless of this distinction, most approaches rely on the Wikipedia encyclopedic KB in order to improve the linking quality, by exploiting its entity descriptions (web pages) or its entity interconnections (hyperlink graph of web pages). In [12], we devise a novel collective linking technique which departs from most approaches in the literature by relying on a structured RDF KB. This allows exploiting the semantics of the interrelationships that candidate entities may have at disambiguation time rather than relying on raw structural approximation based on Wikipedia's hyperlink graph. The few approaches that also use an RDF KB simply rely on the existence of a relation between the candidate entities to which mentions may be linked. Instead, we weight such relations based on the RDF KB structure

and propose an efficient decoding strategy for collective linking. Experiments on standard benchmarks show significant improvement over the state of the art.

4 Software development

4.1 FuzzViz

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

FuzzViz is a research prototype that acts as a business intelligence tool offering functionalities of knowledge discovery from massive raw data. It provides very efficiently an interactive summary of the data [18, 16] and help users identify interesting data properties (correlations, atypical properties, etc.)

4.2 Ikeys

Participants: Olivier Pivert, Grégory Smits.

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 NSGA-G

Participants: Laurent d'Orazio.

NSGA-G is a prototype of a genetic algorithm of the NSGA family aiming to provide a trade-off between performances and diversity, in particular with respect to NSGA II and NSGA III.

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 PINED-RQ++

Participants: Laurent d'Orazio.

PINED-RQ++ is a prototype that addresses the scalability and privacy of range queries in the cloud. It is an extension of PINED-RQ, which implements a template index and a parallel version of a collector.

4.6 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.

4.7 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.

4.8 Sugar

Participants: Olivier Pivert, Virginie Thion.

Sugar is a prototype, based on the Neo4j graph database management system, which allows querying 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).

4.9 Tamari

Participants: Virginie Thion.

Tamari is software add-on, based on the Neo4j graph database management system, which allows introducing data quality-awareness when querying a graph database.

Based on quality annotations that denote quality problems appearing in data (the annotations typically result from collaborative practices in the context of open data usage like e.g. users' feedbacks), and on a user's profile defining usage-dependant quality requirements, the TAMARI prototype computes a quality level of each retrieved answer.

5 Contracts and collaborations

5.1 International Initiatives

5.1.1 AGRI-WATCH

Participants: Laurent d'Orazio.

The PHC SIAM project AGRI-WATCH (2018-2019) brings together experts in data management from Univ. Rennes 1, Univ. Cergy Pontoise, U. Paris Saclay and Kasetsart University. The aim of the project is to design and deploy a novel architecture for smart farming.

5.1.2 MOCCAD

Participants: Laurent d'Orazio.

The NSF project MOCCAD (2013-2019) brings together experts in data management from University of Oklahoma and Univ. Rennes 1. The aim of the project is to design and deploy a novel architecture for mobile cloud computing.

5.2 National Initiatives

5.2.1 ContentCheck

Participants: François Goasdoué.

The ANR project ContentCheck (2015-2020) brings together experts in data management, natural language processing, automated reasoning and data mining from Inria, Univ. Lyon 1, Univ. Paris Saclay, Univ. Rennes 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.

5.2.2 CQFD

Participants: François Goasdoué, Hélène Jaudoin.

The ANR project CQFD (2019-2023) brings together experts in automated reasoning, data management and knowledge representation from Inria, Telecom ParisTech, Univ. Bordeaux, Univ. Grenoble, Univ. Montpellier and Univ. Rennes 1. The aim of the

project is to devise data management algorithms for distributed knowledge-based data management systems.

5.3 CreDoc

Participants: Hélène jaudoin, Olivier Pivert, Grégory Smits.

The project CreDoc, funded by the Region Bretagne and the FEDER, aims to conceive a solution automating the tracking and the controls related to the maritime transportation of goods, in order to help financial institutions fight fraud and financial crime. Apart from IRISA/Shaman, the other participants are CLS (Brest), HSBC and the startup Semsoft based in Rennes.

5.3.1 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).

5.3.2 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 SERBER

Participants: Laurent d'Orazio.

The project SERBER, funded by the Region Bretagne and LTC aims at developing a big data management platform for security monitoring.

5.5 Think Cities

Participants: Laurent d'Orazio.

The project Think Cities, funded by the Region Bretagne aims at developing a digital tool for smart city to evaluate urban projects. Apart from IRISA/Shaman, the other participants are SETUR (Rennes) and SenX (Brest).

5.6 Collaborations

Grégory Smits still collaborates with Ronald R. Yager, head of the Machine Intelligence Institute from Iona College New-York. He has obtained a grant from the Region Brittany (Boost'Mobility) in 2019 to work with him on theoretical aspect of knowledge management. 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 Feb. 2019.

In 2019, Olivier Pivert and Grégory Smits have reinforced their collaborations with Marie-Jeanne Lesot from Sorbonne Université Paris. They have worked on the creation of novel approaches to knowledge management that are at the intersection of machine learning, database and soft computing. Three papers have been submitted at PAKDD20, FuzzIEEE20 and IPMU20 as results of this fruitful collaboration.

6 Dissemination

6.1 Promoting scientific activities

6.1.1 Scientific Events Selection

Member of Conference Program Committees

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

- AAAI Conference on Artificial Intelligence (AAAI)
- Atelier Web des Données (AWD)
- International Conference on Extending Database Technology (EDBT)
- Conference on Extraction et Gestion de Connaissances (EGC)
- International Joint Conference on Artificial Intelligence (IJCAI)

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

- International Conference on Big Data Analytics and Knowledge Discovery (DAWAK@DEXA)
- International Workshop on Data Engineering meets Intelligent Food and COoking Recipe (DECOR@ICDE)
- International Workshop on Benchmarking, Performance Tuning and Optimization for Big Data Applications (BPOD@BigData)

- International Workshop on BI and BIG Data APPlications (BBIGAPP@ADBIS)
- International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)
- International Conference on Intelligent User Interfaces (IUI)
- International Conference on Artificial Intelligence Testing (AI Testing)
- Journées Bases de Données Avancées (BDA)
- Colloque sur l'Optimisation et les Systèmes d'Information (COSI)

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

- International Conference on Flexible Query Answering Systems (FQAS)
- ACM Symposium on Applied Computing (ACM SAC)
- IEEE International Conference on Fuzzy Systems (Fuzz-IEEE)
- International Conference on Scalable Uncertainty Management (SUM)
- European Society for Fuzzy Logic and Technology Conference (EUSFLAT)
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA)

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

- AAAI Conference on Artificial Intelligence (AAAI-student)
- European Conference on Artificial Intelligence (ECAI)
- International Conference on Flexible Query Answering Systems (FQAS)
- IEEE International Conference on Fuzzy Systems (Fuzz-IEEE)
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA)

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

- Conference on Extraction et Gestion de Connaissances (EGC)
- Congrès INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID)

6.1.2 Journals

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 Communications of the ACM (CACM); Information Systems (IS).
- Laurent d'Orazio reviewed for Computing, International Journal of Data Warehousing and Mining (IJDWM).
- Olivier Pivert reviewed for the International Journal of Approximate Reasoning (IJAR), Information Fusion, Fuzzy Sets and Systems (FSS), and the Journal of Intelligent Information Systems (JIIS).
- Grégory Smits reviewed for Transactions on Fuzzy Systems (TFS), Fuzzy Sets and Systems (FSS) and International Journal of Intelligent & Fuzzy Systems (JIFS).
- Virginie Thion reviewed for Fuzzy Sets and Systems (FSS) and the International Journal of Intelligent & Fuzzy Systems (JIFS).

6.1.3 Invited Talks

- François Goasdoué gave an invited talk on "Learning commonalities between RDF graphs and between SPARQL queries" at the Reasoning on Data (RoD) workshop of the Madics CNRS Research Group (Big data and data science), June 27, 2019.
- François Goasdoué gave an invited talk on "Query answering in ontological databases" at the Automata, Logic, Games & Algebra (ALGA) workshop of the Informatique Mathématique CNRS Research Group (Logic and complexity), October 11, 2019.
- Laurent d'Orazio gave an invited talk on "Big Data technologies, an application to smart farming" at Kasetsart University, June 3rd 2019.
- Laurent d'Orazio gave an invited talk on "Big Data technologies and security monitoring" at "Tour de France de la Cyber Sécurité", Lannion, June 27th 2019.
- Laurent d'Orazio gave an invited talk on "Big Logs Management" at "Journées Francophones sur les Entrepôts de Données (EDA)", Montpellier, October 2nd 2019.

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.

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

- PhD 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);
- PhD in progress: Le Trung Dung, Data Management in cloud federation, defended in July 2019, Laurent D'Orazio and Verena Kantere (Univ. Ottawa, Canada);
- PhD 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);

- PhD in progress: Mohamed Handaoui, Big data applications scheduling on ephemeral and heterogeneous Cloud resources, the case of learning algorithms, started Oct. 2019, Jalil Boukhobza (LabSTIC), Olivier Barais (IRISA/DiverSE), and Laurent D'Orazio;
- PhD 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);
- PhD 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

François Goasdoué

- PhD, president, Ibrahim Dellal, ENSMA

Olivier Pivert

- HDR, member, Virginie Thion, Univ. Rennes 1

Laurent d'Orazio

- PhD, referee, George Ajam, University of New South Wales
- PhD, referee, Mariem Brahem, Université Paris Saclay
- PhD, member, Van Bao Nguyen, Université Grenoble Alpes

Grégory Smits

- PhD, member, Martin Lenart, Thales Polska, Sorbonne Université, AGH Université de Cracovie

6.3 Popularization

The newsletter n°105 of the French Association for Artificial Intelligence bulletin presents the SHAMAN research contributions related to Artificial Intelligence.

The Sciences Ouest magazine, in its issue n°375 of October 2019, presents the SHAMAN research contributions related to fact checking.

7 Bibliography

Major publications by the team in recent years

[1] S. Aridhi, L. d'Orazio, M. Maddouri, E. M. Nguifo, "Density-based data partitioning strategy to approximate large-scale subgraph mining", *Inf. Syst. 48*, 2015, p. 213–223, https://doi.org/10.1016/j.is.2013.08.005.

[2] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Explaining Inconsistency-Tolerant Query Answering over Description Logic Knowledge Bases", *in : AAAI Conference on Artificial Intelligence*, Phoenix, United States, 2016, https://hal.inria.fr/hal-01277086.

[3] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Query-Driven Repairing of Inconsistent DL-Lite Knowledge Bases", *in : IJCAI: International Joint Conference on Artificial Intelligence*, New York, United States, 2016, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01367864.

[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] 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.

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

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

[8] 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.

[9] G. Smits, P. Nerzic, O. Pivert, M. Lesot, "Efficient Generation of Reliable Estimated Linguistic Summaries", *in : Proc. of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2018)*, 2018.

[10] T. Tran, T. Phan, A. Laurent, L. d'Orazio, "Improving Hamming distance-based fuzzy join in MapReduce using Bloom Filters", *in : IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, p. 1–7, Rio de Janeiro, Brazil, 2018.

Doctoral dissertations and "Habilitation" theses

[1] V. Thion, *Quality Management of Information Systems - A Human-Centric Point of View*, Habilitation à Diriger des Recherches (HDR), Université Rennes 1, October 2019, https://hal.inria.fr/tel-02407434.

Articles in referred journals and book chapters

[2] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Computing and Explaining Query Answers over Inconsistent DL-Lite Knowledge Bases", *Journal of Artificial Intelligence Research 64*, March 2019, p. 563–644, https://hal.inria.fr/hal-02066288.

[3] A. Castelltort, A. Laurent, O. Pivert, O. Slama, V. Thion, "Fuzzy Preference Queries to NoSQL Graph Databases", *in : NoSQL Data Models – Trends and Challenges*, O. Pivert (editor), *1*, Chapter 6, 2018, p. 167–201, https://hal.inria.fr/hal-01962965.

[4] S. Cebiric, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, M. Zneika, "Summarizing Semantic Graphs: A Survey", *The VLDB Journal 28*, 3, June 2019, https://hal.inria.fr/hal-01925496.

[5] F. Goasdoué, "SHAMAN : Symbolic and Human-centric view of dAta MANagement", *Bulletin de l'Association Française pour l'Intelligence Artificielle*, July 2019, https://hal.inria.fr/hal-02345067.

[6] A. Moreau, O. Pivert, G. Smits, "Linguistically Characterizing Clusters of Database Query Answers", *Fuzzy Sets and Systems 366*, 2019, p. 18–33.

[7] O. Pivert, O. Slama, V. Thion, "Fuzzy Quantified Queries to Fuzzy Graph Databases", *Fuzzy Sets and Systems 366*, 2019, p. 3–17.

Publications in Conferences and Workshops

[8] M. Buron, F. Goasdoué, I. Manolescu, M.-L. Mugnier, "Reformulation-based query answering for RDF graphs with RDFS ontologies", *in : ESWC 2019 - European Semantic Web Conference*, Portoroz, Slovenia, March 2019, https://hal.archives-ouvertes.fr/hal-02051413.

[9] T.-D. Cao, L. Duroyon, F. Goasdoué, I. Manolescu, X. Tannier, "BeLink: Querying Networks of Facts, Statements and Beliefs", *in : ACM CIKM : 28th International Conference on Information and Knowledge Management*, Beijing, China, November 2019, https://hal.inria.fr/hal-02269134.

[10] E. Doumard, O. Pivert, G. Smits, V. Thion, "Processing Fuzzy Relational Queries Using Fuzzy Views", *in : Proc. of the 28th IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'19)*, New Orleans, LA, USA, 2019.

[11] L. Duroyon, F. Goasdoué, I. Manolescu, "A Linked Data Model for Facts, Statements and Beliefs", *in : International Workshop on Misinformation, Computational Fact-Checking and Credible Web, WWW '19 Companion - Proceedings of the 2019 World Wide Web Conference*, San Francisco, United States, May 2019, https://hal.inria.fr/hal-02057980.

[12] C. B. El Vaigh, F. Goasdoué, G. Gravier, P. Sébillot, "Using Knowledge Base Semantics in Context-Aware Entity Linking", *in : DocEng 2019 - 19th ACM Symposium on Document Engineering*, ACM, p. 1–10, Berlin, Germany, September 2019, https://hal.inria.fr/hal-02171981.

[13] F. Goasdoué, P. Guzewicz, I. Manolescu, "Incremental structural summarization of RDF graphs", *in : EDBT 2019 - 22nd International Conference on Extending Database Technology*, Lisbon, Portugal, March 2019, https://hal.inria.fr/hal-01978784.

[14] T. Le, V. Kantere, L. d'Orazio, "Dynamic Estimation for Medical Data Management in a Cloud Federation", *in : Proceedings of the Workshops of the EDBT/ICDT*, Lisbon, Portugal, 2019.

[15] T. Le, V. Kantere, L. d'Orazio, "Optimizing DICOM Data Management with NSGA-G", *in : Proceedings of the International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP@EDBT/ICDT)*, Lisbon, Portugal, 2019.

[16] M. Lesot, G. Smits, P. Nerzic, O. Pivert, "Génération efficace d'estimations fiables de résumés linguistiques", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'19)*, 2019.

[17] M. Pilven, S. Scherzinger, L. D'orazio, "On Complex Value Relations in Hive", *in : International Workshop on Modeling and Management of Big Data*, Salvador, Bahia, Brazil, November 2019, https://hal.archives-ouvertes.fr/hal-02290732.

[18] G. Smits, P. Nerzic, O. Pivert, M.-J. Lesot, "FRELS: Fast and Reliable Estimated Linguistic Summaries", *in : Proc. of the 28th IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'19)*, New Orleans, LA, USA, 2019.

[19] H. V. Tran, T. Allard, L. d'Orazio, A. El Abbadi, "Range Query Processing for Monitoring Applications over Untrustworthy Clouds", *in : International Conference on Extending Database Technology (EDBT)*, p. 666–669, Lisbon, Portugal, 2019.

[20] C. Wang, Z. Arani, L. Gruenwald, L. d'Orazio, "A Vision of a Decisional Model for Re-optimizing Query Execution Plans Based on Machine Learning Techniques", *in : Proceedings of the International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP@EDBT/ICDT)*, Lisbon, Portugal, 2019.