UMR IRISA

Activity Report 2023

Team SHAMAN

Symbolic and Human-centric view of dAta MANagement

D7 – Data & Knowledge Management

Université de Rennes | ENS rennes | INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES | cnrs | Inria | Inserm | CentraleSupélec | IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom | Université Bretagne Sud ubs

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
Virginie Thion, Associate Professor, HDR, ENSSAT

Research engineers, technical staff

Pierre Alain, Research engineer (PhD), ENSSAT (20%), up to Aug. 23

PhD students

Adel Aly, PhD student, ENSSAT, since Oct. 23
Wafaa El Husseini, PhD student, ENSSAT, from Sep. 20 to Dec. 23
Yamen Haddad, PhD student, Inria Saclay, from Jan. 21 to Jan. 23
Vincent Lannurien, PhD student, ENSTA Bretagne, since Oct. 21
Vérone Yepmo, PhD student, ENSSAT, from Sep. 20 to Dec. 23

Administrative assistant

Angélique Le Pennec, team assistant, ENSSAT (20%)
Joëlle Thépault, team assistant, 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.

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

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

[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*.

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.

- 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 that was funded by the Pôle d'Excellence Cyber. One of the main goals was to provide a Big Data platform applied to security monitoring. The team is still investigated this direction with an informal collaboration with Thales. We are considering 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 Sea Defender (2020 – 2024), founded by the DGA (Direction Générale de l'Armement), 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.

- Digital score libraries. *Sheet music scores* have been the traditional way to preserve and disseminate western classical music works for centuries. Nowadays, their content can be encoded in digital formats that yield a very detailed representation of music content expressed in the language of *music notation*. These digitized music scores constitute, therefore, an invaluable asset for digital library services. Shaman studies the data management of digitized music score data, including the design of intuitive and effective querying process and the data quality management of such data. This axis involves collaboration with the Dastum association, a Breton cultural organization based in Rennes (Brittany, France), whose mission is "to collect, protect and promote the cultural heritage of Brittany.

3 Scientific achievements

3.1 Big data management

Participants: Laurent d'Orazio, Vincent Lannurien, Remi Uhartegaray.

- HeROfake: Heterogeneous Resources Orchestration in a Serverless Cloud - An Application to Deepfake Detection [4]. Serverless is a trending service model for cloud computing. It shifts a lot of the complexity from customers to service providers. However, current serverless platforms mostly consider the provider's infrastructure as homogeneous, as well as the users' requests. This limits possibilities for the provider to leverage heterogeneity in their infrastructure to improve function response time and reduce energy consumption. We propose a heterogeneity-aware serverless orchestrator for private clouds that consists of two components: the autoscaler allocates heterogeneous hardware resources (CPUs, GPUs, FPGAs) for function replicas, while the scheduler maps function executions to these replicas. Our objective is to guarantee function response time, while enabling the provider to reduce resource usage and energy consumption. This work considers a case study for a deepfake detection application relying on CNN inference. We devised a simulation environment that implements our model and a baseline Knative orchestrator, and evaluated both policies with regard to consolidation of tasks, energy consumption and SLA penalties. Experimental results show that our platform yields substantial gains for all those metrics, with an average of 35% less energy consumed for function executions while consolidating tasks on less than 40% of the infrastructure's nodes, and more than 60% less SLA violations.

- Scalable Computation of Fuzzy Joins Over Large Collections of JSON Data [9]. Fuzzy joins are widely used in a variety of data analysis applications such as data integration, data mining, and master data management. In the context of Big Data, computing fuzzy joins is challenging due to the high computational cost required and the communication cost. While on one hand big fuzzy joins on relational data and on the other hand joins on tree-structured data have been considered in the literature, to the best of our knowledge, combining the two is still an open problem. In this context, we study methods for leveraging distributed

environments in order to compute fuzzy joins over large collections of JSON documents. Our algorithms take into account both the text-similarity of the joining data, as well as its structural similarity.

3.2 Flexible, cooperative and quality-aware data management

Participants: Ludovic Liétard, Pierre Nerzic, Olivier Pivert, Grégory Smits, Virginie Thion, Véronne Yepmo.

- *Data Partitioning and Anomaly Detection Based on an Isolation Forest.* In [10], we introduce an approach that constitutes a step towards the generation of constrastive explanations for data anomalies. It is based on a variant of the Isolation Forest algorithm, with the main objective of preserving the structure of regular data in order to make its reconstitution easier. Experiments on synthetic datasets show that this method deteriorates the structure of regular data significantly less than the classical method. It can thus serve as a basis to a unified approach aimed to both detect and explain the anomalies. Such a unified approach has indeed been defined and is described in detail in [1].
- *Fuzzy Partition Generation Based on Data Density.* Fuzzy partitions associated with linguistic variables are particularly useful to provide users with a description of the data. However, designing a fuzzy vocabulary that makes it possible to linguistically describe the data distribution and its inner structure is a tedious task. In [6, 5], we introduce a novel strategy to infer possible fuzzy partitions from the data distribution with the objective to have available modalities to describe both dense and sparse regions. A data inner structure as well as the anomalies are then identified using this vocabulary whose terms are also used to provide users with contrastive explanations about the found anomalies.
- *Diversifying Top-k Answers in a Query By Example Setting.* For a given database T and a user query Q, the top-k answers are the k tuples from T that best match Q. The integration of a diversity constraint aims at avoiding returning redundant tuples, that are too similar one to another. In [8, 7], we address the diversification problem in the Query By Example setting, taking into account the situation where very different representative examples are provided by the user. The approach defined in [8, 7] proposes a new definition for diversity that depends on the query, in order to guarantee that the result set illustrates the diversity of the representative examples provided by the user. The approach includes a numerical measure to assess diversity in that sense, and an algorithm that identifies such a diversified top-k set, maximizing both the query satisfaction and the diversity measure.

3.3 Ontology-based data management

Participants: Wafaa El Husseini, Cheikh-Brahim El Vaigh, François Goasdoué, Hélène Jaudoin.

- *OptiRef: Query Optimization for Knowledge Bases.* Ontology-mediated query answering (OMQA) consists in asking database queries on a knowledge base (KB); a KB is a set of facts, the KB's database, described by domain knowledge, the KB's ontology. FOL-rewritability is one of the main OMQA technique: it reformulates a query w.r.t. the KB's ontology so that the evaluation of the reformulated query on the KB's database computes the correct answers. However, because this technique embeds the domain knowledge relevant to the query into the reformulated query, a reformulated query may be complex and its optimization is the crux of efficiency. In [3], we showcase OptiRef that implements a novel, general optimization framework for efficient query answering on datalog+/-, description logic, existential rules, OWL and RDF/S KBs. OptiRef optimizes reformulated queries by rapidly computing, based on a KB's database summary, simpler (contained) queries with the same answers. We demonstrate OptiRef's effectiveness on well-established benchmarks: performance is significantly improved in general, up to several orders of magnitude in the best cases.

4 Software development

4.1 FuzViz

Participants: Pierre Nerzic, Grégory Smits.

FuzViz includes three fuzzy vocabulary elicitation methods based on the distribution of the data estimated from statistics, and a scalable linguistic summarization strategy. The goal of this prototype is to show how complementary our scientific contributions are and that they provide pragmatic solutions to concrete needs. In terms of functionalities, FuzViz provides fluid and intuitive exploration methods and interactive views of massive relational data. We are currently collaborating with the SATT Ouest Valorisation company and Stratinnov to obtain a software maturation funding and to reach companies interested in such functionalities.

4.2 Musypher

Participants: Virginie Thion.

Musypher is an application that makes it possible to transcript a music score, encoded in a XML dialect (MEI or MusicXML), into an attributed graph database hosted by a Neo4j database management system. Our goal is to illustrate the relevancy (expressiveness, efficiency) of managing music scores over a graph-based data model.

4.3 The SKRID platform

Participants: Pierre Alain, Vincent Barreaud, Virginie Thion.

The skrid platform is a digital score library that makes available some Traditional Breton music scores. The skrid platform is still under development but already makes

available some music scores that were collected and encoded by Shaman. The platform will integrate the approaches developed in the *Music score management* research axis of Shaman (under development). The SKRID platform is available at `https://shaman.enssat.fr/skrid/`

4.4 Smarten

Participants: Olivier Pivert, Virginie Thion.

SMARTEN is an application that allows extending a mind map by querying data stemming from graph databases. It implements a theoretical framework that uses fuzzy set theory in order to identify the graph databases concepts that could contribute to the extension of the mind map, and also to compute scores (a relevancy score and an originality acore) associated with each suggestion.

4.5 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).

4.6 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.

4.7 OptiRef

Participants: Pierre Alain, Wafaa El Husseini, Cheikh-Brahim El Vaigh, François Goasdoué, Hélène Jaudoin.

OPTIREF is a JAVA tool built on top of ontology-based data management systems in order to optimize them. It features a PHP/JSP/jQuery-based GUI in order to examine the performance it brings to off-the-shelf ontology-based data management systems.

4.8 FRESQUE

Participants: Hoang Van Tran, Laurent d'Orazio.

FRESQUE is a framework for secure range query processing, that enables a scalable consumption throughput while still maintaining strong privacy protection for outsourced data.

4.9 Time-Series Semantic Caching

Participants: Trung Dung Le, Laurent d'Orazio.

Time-Series Semantic Caching is a form-based semantic caching for Time Series Data (TSD) system. The approach reduces both query result storing based on semantic caching technique and the data transfer between clients and servers.protection for outsourced data.

4.10 MASCARA

Participants: Van Long Nguyen Huu, Laurent d'Orazio.

MASCARA is a FPGA-based semantic caching. The approach relies on hardware acceleration to improve performances (in particular response times and energy consumption) in big data processing.

4.11 HeROfake

Participants: Vincent Lannurien, Laurent d'Orazio.

HeROfake is a heterogeneity-aware serverless orchestrator for private clouds that consists of two components: the autoscaler allocates heterogeneous hardware resources (CPUs, GPUs, FPGAs) for function replicas, while the scheduler maps function executions to these replicas. Our objective is to guarantee function response time, while enabling the provider to reduce resource usage and energy consumption.

4.12 OntoSQL

Participants: Maxime Buron, Cheikh-Brahim El Vaigh, 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.

5 Contracts and collaborations

5.1 International Initiatives

5.1.1 DODAM

Participants: Wafaa El Husseini, Cheikh-Brahim El Vaigh, Francçois Goasdoué, Hélène Jaudoin.

The Stic-AmSud project DODAM (2022-2024) brings together experts from artificial intelligence and data management from Univ. Rennes and Univ. Sorbonnes Universités in France as well as from Univ. Adolfo Ibanez (Chile), Univ. Buenos Aires (Argentina), Univ. de la Republica (Uruguay) in South America. The goal of this project is to study how knowledge representation and reasoning can improve performance, interpretability and explanability of machine learning and data analytics.

5.2 National Initiatives

5.2.1 CQFD

Participants: Wafaa El Husseini, Cheikh-Brahim El Vaigh, François Goasdoué, Hélène Jaudoin.

The ANR project CQFD (2019-2024) 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.

5.2.2 SeaDefender

Participants: Grégory Smits, Olivier Pivert, Pierre Nerzic.

Sea Defender is a project funded by the DGA that involves the Semsoft company (located in Rennes) and the SHAMAN team. The goal of this project is to provide a novel anomaly detection workflow dedicated to the particular cases of under and upper pricing, which the main cause of money laundering in the world. To solve this issue, two scientific issues are addressed by the shaman team : the detection of contextual anomalies and the explanation of the found anomalies. This two tasks form the basis of research subjects studied by Véronne Yepmo (PhD) and Rahul Nath (research engineer).

6 Dissemination

6.1 Promoting scientific activities

6.1.1 Scientific Events Organisation

General Chair, Scientific Chair

François Goasdoué was the Chair of the expert committee for the PhD Award 2023 of the French Community on Data Management (BDA)

6.1.2 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)
- European Conference on Artificial Intelligence (ECAI)
- 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 Intelligent Data - From Data to Knowledge (DOING@ADBIS)
- Journées Bases de Données Avancées (BDA)
- Extraction et Gestion des Connaissances (EGC), démonstrations

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

- International Conference on Flexible Query Answering Systems (FQAS'23)
- ACM Symposium on Applied Computing (SAC'23)
- IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'23)
- European Society for Fuzzy Logic and Technology Conference (EUSFLAT'23)
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA'23)

Hélène Jaudoin served as a member of the following program committee:

- International Conference on Flexible Query Answering Systems (FQAS'23)

6.1.3 Journal

Member of the 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

Laurent d'Orazio served as a reviewer of:

- Information Sciences
- Future Generation Computer Systems (FGCS)
- Expert Systems with Applications (ESWA)
- Internet of Things (IoT)
- International Journal of Computational Intelligence Systems (IJCIS)
- Computer Science and Information Systems (ComSIS)

6.1.4 Invited Talks

Laurent d'Orazio has been invited to present his work at the National Institute of Informatics, Tokyo, Japan: Secure and efficient big data analysis: (some) challenges and perspectives.

6.1.5 Leadership within the Scientific Community

François Goasdoué is a member of the IJCAI Program Committee Board, from 2022 to 2024.

François Goasdoué is a member of the Steering Committee of "Communauté Francophone en Gestion de Données : Principes, Technologies et Applications" (BDA).

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).

6.1.6 Scientific Expertise

Laurent d'Orazio is an expert for the Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur (Hcéres).

Laurent d'Orazio is an expert for the Direction générale de la recherche et de l'innovation (DGRI).

Olivier Pivert is an expert for the Czech Science Foundation.

6.1.7 Research Administration

François Goasdoué is a member of the Scientific Steering Committee of IRISA UMR 6074, since 2013.

François Goasdoué is a member of the Laboratory council of IRISA UMR 6074, since 2022.

François Goasdoué is the head of the Shaman team of IRISA, since 2019.

François Goasdoué is the head of the Lannion branch of IRISA, since 2020.

François Goasdoué is the head of the Scientific Committee of ENSSAT, since 2022.

Laurent d'Orazio is the co-head of the International Relationship office of IRISA UMR 6074, since 2022.

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 teaches a course about *Advanced Databases*, François Goasdoué and Hélène Jaudoin teach a course on *Web Data Management*, and Laurent d'Orazio teaches a part of the course on *Data analysis and data mining*.

6.2.2 Supervision

- PhD in progress: Adel Aly, Towards Flexible Querying of Musical Score Databases, advisor: Virginie Thion.
- PhD in progress: Vincent Lannurien, Big data applications scheduling on heterogeneous Cloud resources, advisors: Laurent d'Orazio, Jalil Boukhobza and Olivier Barais.
- PhD: Efficient Ontology-Mediated Data Management, defended December 2023, advisors: François Goasdoué and Hélène Jaudoin.
- PhD: Véronne Yepmo, Contribution to Anomaly Detection and Explanation: A Unified Method based on Isolation Forest, defended in December 2023, advisors: Olivier Pivert and Grégory Smits.

6.2.3 Juries

François Goasdoué

- PhD, president, Ambre Ayats, Université de Rennes
- PhD, referee, Guillaume Pérution, Université de Montpellier
- PhD, referee, Hui Yang, Université de Saclay
- PhD, referee, Adam Sanchez, Université de Grenoble
- HDR, referee, Ouassila Narsis, Université de Dijon

Laurent d'Orazio

- PhD, president, François Mentec, Université de Rennes
- PhD, referee, Redha Benhissen, Université Lumière Lyon 2
- PhD, Houssameddine Yousfi, Ecole Nationale Supérieure de Mécanique et d'Aérotechnique
- HDR, referee, Giang Nguyen, Slovak University of Technology in Bratislava

7 Bibliography

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`.

M. Buron, F. Goasdoué, I. Manolescu, M.-L. Mugnier, "Ontology-Based RDF Integration of Heterogeneous Data", *in : EDBT/ICDT 2020 - 23rd International Conference on Extending Database Technology*, Copenhagen, Denmark, March 2020, `https://hal.inria.fr/hal-02446427`.

W. El Husseini, C. B. El Vaigh, F. Goasdoué, H. Jaudoin, "Query Optimization for Ontology-Mediated Query Answering", *in : The ACM Web Conference (WWW)*, Singapore, Singapore, May 2024, https://hal.science/hal-04470002.

F. Goasdoué, P. Guzewicz, I. Manolescu, "RDF graph summarization for first-sight structure discovery", *The VLDB Journal 29*, 5, April 2020, p. 1191–1218, https://hal.inria.fr/hal-02530206.

M. Georgoulakis Misegiannis, L. D'orazio, V. Kantere, "From Cloud to Serverless: MOO in the new Cloud epoch", *in : International Conference on Extending Database Technology (EDBT)*, Virtual, United Kingdom, March 2022, https://hal.inria.fr/hal-03925696.

V. L. Nguyen Huu, J. Lallet, E. Casseau, L. D'Orazio, "MASCARA-FPGA cooperation model: Query Trimming through accelerators", *in : SSDBM 2021 - 33rd International Conference on Scientific and Statistical Database Management*, ACM, p. 203–208, Tampa, United States, July 2021, https://hal.inria.fr/hal-03503635.

O. Pivert, E. Scholly, G. Smits, V. Thion, "Fuzzy quality-aware queries to graph databases", *Information Sciences 521*, February 2020, p. 160–173, https://hal.inria.fr/hal-02484041.

O. Pivert, O. Slama, V. Thion, "Expression and efficient evaluation of fuzzy quantified structural queries to fuzzy graph databases", *Fuzzy Sets and Systems 366*, July 2019, p. 3–17, https://hal.inria.fr/hal-02444573.

H. Van Tran, T. Allard, L. D'orazio, A. El Abbadi, "FRESQUE: A Scalable Ingestion Framework for Secure Range Query Processing on Clouds", *in : EDBT 2021 - 24th International Conference on Extending Database Technology*, Nicosia, Cyprus, March 2021, https://hal.inria.fr/hal-03198346.

V. Yepmo, G. Smits, O. Pivert, "Anomaly Explanation : A Review", *Data and Knowledge Engineering*, November 2021, https://hal.archives-ouvertes.fr/hal-03449887.

Doctoral dissertations and "Habilitation" theses

[1] V. Yepmo Tchaghe, *Contribution to Anomaly Detection and Explanation*, Theses, Université de Rennes, December 2023, https://hal.science/tel-04465556.

Articles in referred journals and book chapters

[2] P. Rigaux, V. Thion, "Exploration de partitions musicales modélisées sous forme de graphe", *Revue des Sciences et Technologies de l'Information - Série ISI : Ingénierie des Systèmes d'Information*, 2023.

Publications in Conferences and Workshops

[3] W. El Husseini, C. B. El Vaigh, F. Goasdoué, H. Jaudoin, "OptiRef: Query Optimization for Knowledge Bases", *in : WWW 2023 - The International World Wide*

Web Conference 2023, Austin, United States, April 2023, https://inria.hal.science/hal-04023665.

[4] V. Lannurien, L. d'Orazio, O. Barais, E. Bernard, O. Weppe, L. Beaulieu, A. Kacete, S. Paquelet, J. Boukhobza, "HeROfake: Heterogeneous Resources Orchestration in a Serverless Cloud – An Application to Deepfake Detection", *in : CCGrid 2023 - IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing*, IEEE, p. 154–165, Bangalore, India, May 2023, https://inria.hal.science/hal-04165179.

[5] R. Nath, G. Smits, O. Pivert, "Détection et Explication des Données Régulières et Irrégulières", *in : Rencontres francophones sur la logique floue et ses applications*, INSA Centre Val de Loire, Bourges, France, November 2023, https://hal.science/hal-04186040.

[6] R. Nath, G. Smits, O. Pivert, "Fuzzy-Vocabulary-Based Detection and Explanation of Anomalies", *in : 2023 IEEE International Conference on Fuzzy Systems*, Incheon, South Korea, August 2023, https://hal.science/hal-04122596.

[7] G. Smits, M.-J. Lesot, O. Pivert, M. Reformat, "Diversification des k meilleures réponses à des requêtes par l'exemple Diversifying top-k Answers in a Query by Example Setting", *in : Rencontres francophones sur la logique floue et ses applications*, INSA Centre Val de Loire, Bourges, France, November 2023, https://hal.science/hal-04185985.

[8] G. Smits, M.-J. Lesot, O. Pivert, M. Reformat, "Diversifying top-k Answers in a Query by Example Setting", *in : Flexible Query Answering System*, Palma de Mallorca, Spain, September 2023, https://hal.science/hal-04122580.

[9] R. Uhartegaray, L. d'Orazio, M. Damigos, E. Kalogeros, "Scalable Computation of Fuzzy Joins Over Large Collections of JSON Data", *in : 2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, Incheon, South Korea, August 2023, https://hal.science/hal-04354170.

[10] V. Yepmo, G. Smits, M.-J. Lesot, O. Pivert, "Vers un partitionnement des données à partir d'une forêt d'isolation", *in : Conférence Extraction et Gestion de Connaissances 2023, Revue des Nouvelles Technologies de l'Information, Extraction et Gestion des Connaissances, RNTI-E-39*, Association EGC, p. 163–174, Lyon, France, January 2023, https://hal.science/hal-03972677.