UMR IRISA

Activity Report 2021

Team SHAMAN

A Symbolic and Human-Centric View of Data Management

D7 – Data and Knowledge Management

CentraleSupélec · cnrs · ENS rennes · IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom · Inria · INSA RENNES · Université Bretagne Sud ubs · UNIVERSITÉ DE RENNES 1

1 Team composition

Researchers and faculty

Cheikh-Brahim El Vaigh, ATER, ENSSAT, from Sep. 20 to Aug. 22
François Goasdoué, Professor, ENSSAT, head of the team
Hélène Jaudoin, Associate Professor, ENSSAT
Ludovic Liétard, Associate Professor, HDR, IUT Lannion
Cyrielle Mallart, ATER, ENSSAT, from Dec. 21 to Aug. 22
Rahul Nath, Postdoc, ENSSAT, from Sep. 21 to Mar. 22
Pierre Nerzic, Associate Professor, IUT Lannion
Laurent d'Orazio, Professor, IUT Lannion
Olivier Pivert, Professor, ENSSAT
Grégory Smits, Associate Professor, HDR, IUT Lannion
Virginie Thion, Associate Professor, HDR, ENSSAT

Research engineers, techical staff

Pierre Alain, Research engineer (PhD), ENSSAT (20%)

PhD students

Taras Basiuk, PhD student, University of Oklahoma, since Mar. 19
Ludivine Duroyon, PhD student, ENSSAT, from Sep. 17 to Aug. 21
Wafaa El Husseini, PhD student, ENSSAT, since Sep. 20
Yamen Haddad, PhD student, Inria Saclay & École Polytechnique, since Jan. 21
Vincent Lannurien, PhD student, ENSTA Bretagne, since Oct. 21
Van Long Nguyen Huu, PhD student ENSSAT Nokia, since Jan. 20
Chenxiao Wang, PhD student University of Oklahoma from Jul. 15 to Dec 21.
Vérone Yepmo, PhD student, ENSSAT, since Sep. 20

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 (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

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.

- 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 (2015-2021) and IPL iCoda (2017-2021) 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 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 *digital scores* constitute, therefore, an invaluable asset for digital library services such as search, analysis, clustering, recommendations, and synchronization with audio files. Digital scores, like any other published data, may suffer from quality problems. For instance, they can contain incomplete or inaccurate elements. As a "dirty" dataset may be an irrelevant input for some use cases, users need to be able to estimate the quality level of the data they are about to use. In furtherance of the GioQoso (défi CNRS mastodons 2016-2019), Shaman still studies, through a collaboration with the CNAM Paris, the problem of managing the data quality in digital score libraries.

3 Scientific achievements

3.1 Big data management

Participants: Taras Basiuk, Laurent d'Orazio, Vincent Lannurien, an Long Nguyen Huu, Hoang Van Tran, Quyen Tran Thi To, Le Trung Dung, Chenxiao Wang.

- MASCARA: The use of Field Programmable Gate Arrays (FPGA) has become attractive in recent years to accelerate database analysis. Meanwhile, Semantic Caching (SC) is a technique for optimizing the evaluation of database queries by exploiting the knowledge and resources contained in the queries themselves. Organizing SC on FPGA is relevant in terms of response time and quality of results to increase system performance. To make SC scalable on FPGAs, we have proposed a ModulAr Semantic CAching fRAmework (MASCARA) in which relevant stages or modules could be convertible as accelerators on FPGAs. Therefore, we present a complementary query processing platform based on the cooperation model between MASCARA and FPGA [13]. This novel approach extends the advantage of the classical SC, which is mainly based on Central Processing Unit (CPU),

by offloading computationally intensive phases to FPGA. Moreover, MASCARA-FPGA presents the workflow of query rewriting and partial query execution in a pipelined execution model where multiple accelerators can run in parallel. In our experiments, the Query Trimming can reduce the response time by up to 3.96 times with only one accelerator used.

- FRESQUE: Performing non-aggregate range queries over encrypted data stored on untrusted clouds has been considered by a large body of work over the last years. However, prior schemes mainly concentrate on improving query performance while the scalability dimension still remains challenging. Due to heavily pre-processing incoming data at a trusted component such as encrypting data and building secure indexes, existing solutions cannot provide a satisfactory ingestion throughput. We overcome this limitation by introducing a framework for secure range query processing, FRESQUE [16], that enables a scalable consumption throughput while still maintaining strong privacy protection for outsourced data. Our experiments on real-world datasets show that FRESQUE can support over 160 thousand record insertions in a second, when running on a 12-computing node cluster. It also significantly outperforms one of the most efficient schemes such as PINED-RQ++ by 43 times on ingestion throughput.

- Query processing on cloud database systems is a challenging problem due to the dynamic cloud environment. The configuration and utilization of the distributed hardware used to process queries change continuously. A query optimizer aims to generate query execution plans (QEPs) that are optimal meet user requirements. In order to achieve such QEPs under dynamic environments, performing query re-optimizations during query execution has been proposed in the literature. In cloud database systems, besides query execution time, users also consider the monetary cost to be paid to the cloud provider for executing queries. Thus, such query re-optimizations are multi-objective optimizations which take both time and monetary costs into consideration. However, traditional re-optimization requires accurate cost estimations, and obtaining these estimations adds overhead to the system, and thus causes negative impacts on query performance. To fill this gap, in this work, we introduce ReOptRL [17], a novel query processing algorithm based on deep reinforcement learning. It bootstraps a QEP generated by an existing query optimizer and dynamically changes the QEP during the query execution. It also keeps learning from incoming queries to build a more accurate optimization model. In this algorithm, the QEP of a query is adjusted based on the recent performance of the same query so that the algorithm does not rely on cost estimations. Our experiments show that the proposed algorithm performs better than existing query optimization algorithms in terms of query execution time and query execution monetary costs.

- AGRI-WATCH: In this work [12] we introduce a Data and Knowledge Integration Model and a Collaborative Platform for fact-oriented Agricultural Biodiversity Management that is inspired by the conservation and sustainable use of biodiversity within agricultural landscapes, which is essential for the future of agriculture and food security. We demonstrate and validate our proposal in a realistic case study that was carried out with stakeholders from educational institutes includ-

ing several government agencies from five Ministries, i.e. Ministry of Agricultural and Cooperative, Ministry of Natural Resources and Environment, Ministry of Public Health, Ministry of Commerce and Ministry of Higher Education, Science, Research and Innovation. Key challenges are how to make data inter-operation across these agencies when re-engineering the existing information system and how to make trustworthy platform for data collecting, integrating and sharing, especially, how to keep these agencies engaged throughout the project. The resulting Syntax-Semantic-Organizational Interoperability model was proposed to provide a candidate best practice for engineering data and knowledge integration through a community-shared and reusable Data Reference Model. The resulting Data Governance Implementation across government agencies, by using BIO-AGRI-WATCH as a case study, has significant consequences regarding communication and engagement with stakeholders and dedicated team for increasing their trust in digital data sharing platform.

- ASSIST: There are fewer female authors than male authors in the field of scientific research. However, there is not yet a system that provides a way to analyze the data that is available, and to backup that claim. This work [11] illustrates the upgrade of a tool previously made, in order to make it more efficient and add new features. Such new features are the keywords cloud or the new statistical functionality. Sources, references and other information on the article will be displayed for each articles retrieved. Genders of the authors will be determined using a database linking first names to genders, to be able to get accurate statistics on a large number of gathered articles.

3.2 Flexible, cooperative and quality-aware data management

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

- *Data quality management in digital score libraries.* In [3], we proposed a data quality management framework for digital score libraries. Such a framework relies on a *content model* that identifies several information levels that are unfortunately blurred out in digital score encodings. This content model then serves as a foundation to organize the categories of quality issues that can occur in a music score, leading to a *quality model*. The quality model also positions each issue with respect to potential usage contexts, allowing attachment of a consistent set of indicators that together measure how a given score is *fit* to a specific usage. The framework was implemented in the online digital score library called NEUMA (`http://neuma.huma-num.fr`).

- *Anomaly detection and explanation.* The SHAMAN team is involed in a DGA project about anomaly detection to maritime transportation control. To this purpose a first study of the existing approaches to anomaly detection and explanation has been done. This work has led to the categorization of the different strategies that may be envisaged to perform this task [18].

- *Disjunctive concepts modelling and inference.* The notion of conjunctive concept is an old problem in Artificial Intelligence that has been studied mostly during the 90's. In [15], it has been shown that the modelling of complex concepts, that are not necessarily conjunctive, may be interesting to flexible querying. Implemented on top a relational DB management system, the proposed approach thus allows for the inference, from a few provided examples, of complex search conditions that can then be integrated into flexible queries.

3.3 Ontology-based data management

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

- *Querying RDF graphs.* Answering queries on RDF knowledge bases is a crucial data management task, usually performed through either graph saturation or query reformulation. In [6], we optimize our recent state-of-the-art query reformulation technique for RDF graphs with RDFS ontologies, and we report on preliminary encouraging experiments showing performance improvement by up to two orders of magnitudes!

- *Querying description logic knowledge bases.* Ontology-mediated query answering (OMQA) is a recent data management trend in the Artificial Intelligence, Database and Semantic Web areas, which aims at answering database queries on knowledge bases. Because it is an intricate combination of automated reasoning and database query evaluation, it raises major performance challenges. In [7], we showcase a decade of OMQA optimization to understand "Where do we stand now and how did we get there?" and we highlight a promising new OMQA optimization that brings further significant performance improvement to discuss "What's next?".

- *Summarizing description logic knowledge bases.* The quotient operation from graph theory offers an elegant graph summarization framework that has been widely investigated in the literature, notably for the exploration and efficient management of large graphs; it consists in fusing equivalent vertices according to an equivalence relation. In [10], we study whether a similar operation may be used to summarize description logic (DL) databases, i.e., ABoxes. Towards this goal, we define and examine the quotient operation on an ABox: we establish that a quotient ABox is more specific than the ABox it summarizes, and characterize to which extent it is more specific. This preliminary investigation validates the interest of a quotient-based ABox summarization framework, and paves the way for further studies on it in the DL setting, e.g., to devise equivalence relations suited to the optimization of typical DL data management and reasoning tasks on large ABoxes or to the visualization of large ABoxes, and on its utilization in related settings, e.g., Semantic Web.

3.4 Machine learning

Participants: Cheikh-Brahim El Vaigh, Cyrielle Mallart.

- In [9], we present in this paper our participation to the task of fake news conspiracy theories detection from tweets. We rely on a variant of BERT-based classification approach to devise a first classification method for the three different tasks. Moreover, we propose a multitask learning approach to perform the three different tasks at once. Finally, we developed a prompt-based approach to generate classifications thanks to a TinyBERT pre-trained model. Our experimental results show the multitask model to be the best on the three tasks.

- The rise of digitization of cultural documents offers large-scale contents, opening the road for development of AI systems in order to preserve, search, and deliver cultural heritage. To organize such cultural content also means to classify them, a task that is very familiar to modern computer science. Contextual information is often the key to structure such real world data, and we propose to use it in form of a knowledge graph. Such a knowledge graph, combined with content analysis, enhances the notion of proximity between artworks so it improves the performances in classification tasks. In [8], we propose a novel use of a knowledge graph, that is constructed on annotated data and pseudo-labeled data. With label propagation, we boost artwork classification by training a model using a graph convolutional network, relying on the relationships between entities of the knowledge graph. Following a transductive learning framework, our experiments show that relying on a knowledge graph modeling the relations between labeled data and unlabeled data allows to achieve state-of-the-art results on multiple classification tasks on a dataset of paintings, and on a dataset of Buddha statues. Additionally, we show state-of-the-art results for the difficult case of dealing with unbalanced data, with the limitation of disregarding classes with extremely low degrees in the knowledge graph.

- Buddha statues are a part of human culture, especially of the Asia area, and they have been alongside human civilisation for more than 2,000 years. As history goes by, due to wars, natural disasters, and other reasons, the records that show the built years of Buddha statues went missing, which makes it an immense work for historians to estimate the built years. In [14], we pursue the idea of building a neural network model that automatically estimates the built years of Buddha statues based only on their face images. Our model uses a loss function that consists of three terms: an MSE loss that provides the basis for built year estimation; a KL divergence-based loss that handles the samples with both an exact built year and a possible range of built years (e.g., dynasty or centuries) estimated by historians; finally a regularisation that utilises both labelled and unlabelled samples based on manifold assumption. By combining those three terms in the training process, we show that our method is able to estimate built years for given images with 37.5 years of a mean absolute error on the test set.

4 Software development

4.1 FuzViz

Participants: Pierre Nerzic, Grégory Smits.

FuzViz aims at turning two scientific contributions into an operational research prototype. It includes a fuzzy vocabulary elicitation method 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 view of massive relational datas. 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 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.

4.4 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).

4.5 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.

4.6 OptiRef

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

OptiRef is a PHP/JSP/jQuery-based GUI that offers a set of visualizations in order to examine the performance challenges and advances in ontology-based data management optimization.

4.7 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.

4.8 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.

4.9 ComIn

Participants: François Goasdoué, Hélène Jaudoin.

Com'In offers a PHP/JSP-based GUI to graphically illustrate previously published works that allow to compute commonalities between RDF descriptions and between SPARQL queries. ComÍn makes it possible to upload two RDF descriptions, respectively two SPARQL queries, in N-Triples format, resp. in an extended N-Triples format, to

visualize them as graphs, to compute their commonalities and to display the result as a graph.

4.10 OntoSQL

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. This year, we improved reformulation-based query answering by optimizing the produced query reformulation using a summary of the queried data. This work is described in [6].

5 Contracts and collaborations

5.1 International Initiatives

[to be removed] Below is an example of a project description that you can use for almost all input with minor adaptation, except maybe for the 'Collaboration' section in which you should list non-contractual collaborations, e.g., with foreign institutions. Formal academic collaborations with contractual funding (Inria associate teams, PICS, ANR PRCI, STIC AmSud, etc.) should be described within the 'International initiatives' section. Formal industry collaboration, including CIFRE, should go into the 'Bilateral industry initiatives' section.

5.2 National Initiatives

5.2.1 ContentCheck

Participants: Ludivine Duroyon, François Goasdoué.

The ANR project ContentCheck (2015-2021) 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: 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.

5.2.3 iCoda

Participants: Cheikh-Brahim El Vaigh, François Goasdoué.

The INRIA Project Lab iCoda — Knowledge-mediated Content and Data Analytics (2017–2021) — 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.

5.2.4 Smarten

Participants: Olivier Pivert, Virginie Thion.

A mind map is a graphical representation of thoughts introduced in the 60's by the psychology scholar Tony Buzan. Mind maps have turned out to be an effective tool for expliciting and organizing individual and collective knowledge, in learning and early-stage ideation processes. The design of a mind map is primarily based on tacit human skills. This task is difficult and expensive. In the Smarten (Défi scientifique Univ. Rennes 1), we consider the problem of aiding the conception of a mind map by using data stemming from any graph database.

5.2.5 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).

5.3 Hardware acceleration, an application to big data analytics in security monitoring

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

The project Think Cities, funded by a CIFRE grant aims at developing optimization techniques, namely semantic caches, on top of new hardware and more precisely FPGAs, with an application in Cyber Security and security monitoring. Apart from IRISA/Shaman, the other participant is Nokia/Alctal Lucent Bell Labs (Lannion).

5.4 Think Cities

Participants: Trung Dung Le, 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).

6 Dissemination

6.1 Promoting scientific activities

6.1.1 Scientific Events Organisation

General Chair, Scientific Chair

Member of the Organizing Committees

François Goasdoué served as a member of the following organizing committee:

- Bases de Donné Avancées (BDA)

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

- Bases de Donné Avancées (BDA)

Laurent d'Orazio served as a member of the following organizing committee:

- Bases de Donné Avancées (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)
- Journées Bases de Données Avancées (BDA)
- Conference on Extraction et Gestion de Connaissances (EGC)
- International Joint Conference on Artificial Intelligence (IJCAI), Senior PC

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

- Conference on Extraction et Gestion de Connaissances (EGC)

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)
- International Workshop on Benchmarking, Performance Tuning and Optimization for Big Data Applications (BPOD@BigData)
- International Workshop on Intelligent Data - From Data to Knowledge (DOING@ADBIS)
- Journés Francophones sur les Entrepôts de Donnés et l'Analyse en ligne (EDA)
- Conference on Extraction et Gestion de Connaissances (EGC)

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

- Journées Bases de Données Avancées (BDA) - Session démonstration
- International Conference on Flexible Query Answering Systems (FQAS)

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

- International Conference on Flexible Query Answering Systems (FQAS'21),
- ACM Symposium on Applied Computing (SAC'21),
- IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'21),
- World Congress of the International Fuzzy Systems Association / Conference of the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT'21),
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA'21).

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

- Rencontres sur la logique floue et ses applications (LFA),
- International Conference on Flexible Query Answering Systems (FQAS),
- Information Processing and Management of Uncertainty (IPMU),
- IEEE conference on Fuzzy Systems (Fuzz IEEE).

Reviewer

Virginie Thion served as a reviewer for the following conferences:

- European Conference on Information Systems (ECIS)
- IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)
- Conference on Extraction et Gestion de Connaissances (EGC)

Grégory Smits served as a reviewer for the following conferences :

- European Conference on Artificial Intelligence (ECAI),
- AAAI Conference on Artificial Intelligence (AAAI student session),
- International journal on Transactions on Fuzzy Systems (TFS),
- International journal on Fuzzy Sets and Systems (FSS),
- International Journal of Intelligent & Fuzzy Systems (JIFS).

6.1.3 Journal

Member of the Editorial Boards

François Goasdoué served as a Guest Editor of the following journal:

- Special issue on Data Management - Principles, Technologies and Applications (BDA'20) - of Transactions on Large-Scale Data and Knowledge-Centered Systems (TLDKS journal) [1]

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,
- Revue Ouverte d'Ingénierie des Systèmes d'Information.

Reviewer - Reviewing Activities

François Goasdoué served as a reviewer for the following journals:

- Artificial Intelligence Journal (AIJ)
- IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE)
- Information Systems (IS)
- Journal of Intelligent Information Systems (JIIS)
- VLDB Journal - The International Journal on Very Large Data Bases (VLDBJ)

Laurent d'Orazio served as a reviewer for the following journals:

- Distributed and Parallel Databases (DAPD)
- Internet of Things (IoT)
- Transactions on Emerging Telecommunications Technologies (ETT)

6.1.4 Invited Talks

Grégory Smits has given an invited talk at LFA'2021 (Paris).

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

Olivier Pivert is an expert for the Czech Science Foundation.

6.1.7 Research Administration

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

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

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: Taras Basiuk, Resources allocation in cloud computing, started in Mar. 19, Laurent d'Orazio and Le Gruenwald
- PhD: Ludivine Duroyon, Data management models, algorithms and tools for fact-checking, started Oct. 2017, François Goasdoué and Ioana Manolescu (INRIA/Cedar)
- PhD in progress: Wafaa El Husseini, Efficient ontology-based data management, started in Oct. 20, François Goasdoué and Hélène Jaudoin
- PhD: Cheikh-Brahim El Vaigh, Content and data linking leveraging ontological knowledge in data journalism, defended in Jan. 21, François Goasdoué, Guillaume Gravier and Pascale Sébillot
- PhD in progress: Yamen Haddad, Adaptative query planning and execution on heterogeneous data sources, started in Jan. 21, Angelos-Christos Anadiotis, François Goasdoué and Ioana Manolescu
- PhD: Van Long Nguyen Huu, Hardware acceleration, an application to big data analytics in security monitoring, started in Jan. 21, Laurent d'Orazio, Emmanuel Casseau and Julien Lallet
- PhD: Vincent Lannurien, Big data applications scheduling on heterogeneous Cloud resources, started in Oct. 21, Laurent d'Orazio, Jalil Boukhobza and Olivier Barais
- PhD: Chenxiao Wang, Multi-objective and adaptative optimization in clouds, defended in Dec. 21, Laurent d'Orazio and Le Gruenwald
- PhD in progress: Véronne Yepmo, Anomaly detection and explanation, started in Nov. 20, Grégory Smits and Olivier Piver

6.2.3 Juries

François Goasdoué

- PhD, referee, Sabiha Tahra, Université de Paris
- PhD, member, Pawel Guzewicz, Institut Polytechnique de Paris

Laurent d'Orazio

- PhD, referee and chair, Juba Agoun, Université de Lyon
- PhD, referee, Zahi Al Chami, Université de Pau et des Pays de l'Adour
- PhD, referee, Mehrdad Farokhnejad, Université de de Grenoble Alpes

Grégory Smits

- PhD, referee, Marcin Lenart, Sorbonne Universités Paris

6.3 Popularization

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.

S. El Hassad, F. Goasdoué, H. Jaudoin, "Learning Commonalities in SPARQL", *in : International Semantic Web Conference (ISWC)*, Vienna, Austria, October 2017, https://hal.inria.fr/hal-01572691.

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.

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.

G. Smits, O. Pivert, R. Yager, P. Nerzic, "A soft computing approach to big data summarization", *Fuzzy Sets and Systems 348*, October 2018, p. 4–20, https://hal.inria.fr/hal-01962961.

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.

Books and Monographs

[1] B. Amann, F. Goasdoué, *Transactions on Large-Scale Data- and Knowledge-Centered Systems XLIX*, *Lecture Notes in Computer Science*, *12920*, Springer, 2021, https://hal.inria.fr/hal-03347656.

Doctoral dissertations and "Habilitation" theses

[2] C. B. El Vaigh, *Content and data linking leveraging ontological knowledge in data journalism*, Theses, Université Rennes 1, January 2021, https://hal.inria.fr/tel-03131484.

Articles in referred journals and book chapters

[3] F. Foscarin, P. Rigaux, V. Thion, "Data Quality Assessment in Digital Score Libraries. The GioQoso Project", *International Journal on Digital Libraries 22*, 2, 2021, p. 159–173, https://hal.archives-ouvertes.fr/hal-03163156.

[4] G. Smits, O. Pivert, "Fuzzy Extensions of Databases", *in : Fuzzy Approaches for Soft Computing and Approximate Reasoning: Theories and Applications*, *Studies in Fuzziness and Soft Computing*, *394*, Springer International Publishing, October 2021, p. 191–200, https://hal.archives-ouvertes.fr/hal-03539556.

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

Publications in Conferences and Workshops

[6] M. Buron, C. B. El Vaigh, F. Goasdoué, "Towards Faster Reformulation-based Query Answering on RDF Graphs with RDFS Ontologies", *in : International Semantic Web Conference (ISWC)*, Online, United States, October 2021, https://hal.inria.fr/hal-03347679.

[7] W. El Husseini, C. B. El Vaigh, F. Goasdoué, H. Jaudoin, "Ontology-Mediated Query Answering: Performance Challenges and Advances", *in : International Semantic Web Conference (ISWC)*, Online, United States, October 2021, https://hal.inria.fr/hal-03347688.

[8] C. B. El Vaigh, N. Garcia, B. Renoust, C. Chu, Y. Nakashima, H. Nagahara, "GCNBoost: Artwork Classification by Label Propagation through a Knowledge Graph", *in : ICMR 2021 - ACM International Conference on Multimedia Retrieval*, Taipei, Taiwan, August 2021, https://hal.inria.fr/hal-03228787.

[9] C. B. El Vaigh, T. Girault, C. Mallart, D. H. Nguyen, "Detecting Fake News Conspiracies with Multitask and Prompt-Based Learning", *in : MediaEval 2021 - MediaEval Multimedia Evaluation benchmark. Workshop*, p. 1–3, Online, Netherlands, December 2021, https://hal.inria.fr/hal-03482254.

[10] C. B. El Vaigh, F. Goasdoué, "A Well-founded Graph-based Summarization Framework for Description Logics", *in : International worshop on Description Logics*, Bratislava, Slovakia, September 2021, https://hal.inria.fr/hal-03347664.

[11] J. Fouillé, T. L. H. Nguyen, B. Alix, B. Becker, M. Rochard, H. De Ribaupierre, L. d'Orazio, "ASSIST: Article eXtraction and statIstical AnalysiS", *in : International Workshop on Data science for equality, inclusion and well-being challenges (DS4EIW@BigData)*, Virtuelle, France, December 2021, https://hal.archives-ouvertes.fr/hal-03522313.

[12] A. Kawtrakul, H. Chanlekha, K. WAIYAMAI, T. Kangkachit, L. d'Orazio, D. Kotzinos, D. Laurent, N. Spyratos, "Towards Data-and-Innovation Driven Sustainable and Productive Agriculture: BIO-AGRI-WATCH as a Use Case Study", *in : Workshop on Smart Farming, Precision Agriculture, and Supply Chain (SmartFarm@BigData)*, Virtuelle, France, December 2021, https://hal.archives-ouvertes.fr/hal-03522308.

[13] 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.

[14] Y. Qian, C. B. El Vaigh, Y. Nakashima, B. Renoust, H. Nagahara, Y. Fujioka, C. Brahim, E. Vaigh, "Built Year Prediction from Buddha Face with Heterogeneous Labels", *in : SUMAC'21: 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents*, Chengdu, China, October 2021, https://hal.inria.fr/hal-03520715.

[15] G. Smits, M.-J. Lesot, O. Pivert, R. R. Yager, "Flexible Querying using Disjunctive Concepts", *in : International Conference on Flexible Query Answering Systems*, Bratislava, Slovakia, September 2021, https://hal.archives-ouvertes.fr/hal-03276700.

[16] 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.

[17] C. Wang, L. Gruenwald, L. D'orazio, E. Leal, "Cloud Query Processing with Reinforcement Learning-based Multi-Objective Re-optimization", *in : International Conference on Model & Data Engineering (MEDI)*, Tallinn, Estonia, June 2021, https://hal.archives-ouvertes.fr/hal-03522314.

[18] V. Yepmo, G. Smits, O. Pivert, "A Classification of Anomaly Explanation Methods", *in : Advances in Interpretable Machine Learning and Artificial Intelligence (AIMLAI)*, (Online), France, September 2021, https://hal.archives-ouvertes.fr/hal-03337036.