UMR IRISA

Activity Report 2025

Team SHAMAN

Symbolic and Human-centric view of dAta MANagement

D7 – Data & Knowledge Management

Université de Rennes · cnrs · CentraleSupélec · ens rennes école normale supérieure · IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom · Inria · INSA Institut National des Sciences Appliquées Rennes · Inserm · Université Bretagne Sud ubs

1 Team composition

Researchers and faculty

Amine Boulhamel, Temporary teaching and research assistant, ENSSAT, since Nov. 24
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

PhD students

Adel Aly, PhD student, ENSSAT, since Oct. 23
Phan-An Truong Tran, ENSSAT, since Sep. 25
Claire Vanoni, ENSSAT, since Oct. 25

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 20th 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.

- 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 the music content expressed in the language of *music notation*. These digitized music scores constitute, therefore, an invaluable asset for digital library services. In this context, 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 collaborations with the Dastum association, a cultural organization based in Rennes (Brittany, France), whose mission is to collect, protect and promote the cultural heritage of Brittany, and with teachers of the traditional music department of the Music Conservatory of Lannion Trégor.

3 Scientific achievements

3.1 Big data management

Participants: Laurent d'Orazio, An-Truong Tran Phan.

- The demand for environmentally friendly cloud computing is on the rise, leading cloud service providers to focus on reducing carbon emissions by using renewable energy sources and energy-efficient computing models. This study [8] assesses the performance and energy consumption of serverless and serverful architectures, specifically looking at join operations using Apache Spark for big data processing in a private cloud combined with Kubernetes. By using the TPC-DS benchmark, we examine the impact of cold-start and warm-start phases in the serverless environment, as well as the auto-scaling capabilities of Spark in serverless environments within the private cloud. The results show that the efficient and flexible resource management in serverless environments in private clouds leads to more optimal processing times and energy consumption compared to serverful architectures, especially in warm-start scenarios. These findings offer valuable insights for organizations seeking to streamline their big data infrastructure while also making a positive environmental impact within the IT industry.

- Merging robotic technologies, sensor networks, and Geographic Information Systems (GIS) offers significant potential across various domains, including agriculture and urban planning. However, a critical challenge lies in the lack of interoperability between data generated by these technologies and existing GIS tools. The EU-funded GIS4IoRT project [9] addresses this gap by developing a plug-and-play and cloud-based middleware. This middleware facilitates seamless integration and visualization of multi-dimensional and multi-modal data within GIS environments. Key GIS4IoRT components include: a middleware architecture, a scalable cloud-based infrastructure, real-time robot querying capabilities, data quality assurance, spatio-temporal query support within the cloud, integration with GIS tools, and adherence to relevant standards. The middleware supports diverse data types, including LiDaR, imagery, and sensor data. This study (1) presents an initial data integration architecture specifically designed for the sustainable architecture domain, (2) outlines the challenges encountered in designing such an architecture, and (3) explores novel data processing paradigms enabled by the architecture.

- The increasing prevalence of JSON documents as a standard format for data storage and exchange in diverse applications has led to the need for efficient methods to compare and analyze hierarchical data structures. Traditional comparison methods often struggle with scalability and fail to account for structural variations in complex data formats. These limitations become particularly problematic in applications such as duplicate detection, anomaly analysis, and data integration, where accurate and scalable comparison of large JSON datasets is essential.To address these challenges, this work presents a scalable framework for comparing JSON documents using the Aho-Hopcroft-Ullman (AHU) algorithm and the MapReduce paradigm. By generating canonical labels for hierarchical structures, the AHU algorithm captures structural similarities between JSON trees. The framework employs structural similarity measures, such as Longest Common Substring (LCS) and Levenshtein Distance, to quantify resemblance. MapReduce enables efficient processing of large JSON datasets, with a mapping phase for parsing and labeling and a reducing phase for similarity computations. This approach offers a robust and scalable solution for hierarchical data comparison, facilitating critical applications in duplicate detection and anomaly analysis.

- We have worked on a demonstrator [7] to propose a multi-dimensional exploration of media collection metadata. Two user interfaces, a virtual reality explorer and a web-based explorer, will be used to explore three different collections of lifelog images, general video clips, and music using a variety of metadata attributes.

3.2 Flexible, cooperative and quality-aware data management

Participants: Adel Aly, Pierre Nerzic, Olivier Pivert, Virginie Thion.

- *Fuzzy Retrieval of Musical Scores Based on Melodic Patterns.* Retrieving music pieces based on their content is a growing area of research in the field of database management. In [3], we addressed the challenge of flexible musical score retrieval based on melodic patterns, offering two contributions. First, we introduced a theoretical framework that leverages fuzzy logic to enable flexibility. The approach allows a user to define fuzzy tolerance thresholds for pitch, duration, and sequencing, and returns ranked answers that "more or less" match a given pattern, offering both explainability and customization. Second, we presented Maelis, an implementation of this framework as an extension of the Cypher graph-pattern query language. Maelis is integrated inside the Skrid platform, a digital library that stores musical scores of folk music from the French region of Brittany.

- *A Flexible Framework for Transposition-Aware Querying of a Musical Score Database.* In [2], we extended the framework proposed in [3] so as to make it tolerant to transposition. The extended framework is capable of retrieving musical scores that contain a given pattern, even if transposed on the pitch axis.

- *MAELIS4SKRID: An Approximate Query Engine for an Online Graph-Based Music Score Library.* In the demo paper [4] presents the Skrid platform, an online Digital Score Library (DSL) that utilizes a graph-based storage for the scores' musical content. To our knowledge, Skrid is the only real DSL that relies on a property graph data model). it also presents Maelis, a flexible querying module implemented inside Skrid, that enables melodic pattern approximate search, ranks the results by relevance, and provides a detailed explanation of the answers. The demonstration iss composed of two scenarios that showcase the key features of Maelis embedded in Skrid. The scenarios are accompanied by an online interaction inviting the audience to engage with the system.

3.3 Ontology-based data management

Participants: Maxime Buron (Université Clermont Auvergne), Théo Ducros (Université Clermont Auvergne), François Goasdoué, Hélène Jaudoin, Farouk Toumani (Université Clermont Auvergne).

- *Optimization for knowledge-based data management.* We recently optimized ontology-mediated query answering (OMQA) that consists in asking database queries on knowledge bases (KBs); a KB is a set of facts called the KB's database, which is described by domain knowledge called the KB's ontology. A widely-investigated OMQA technique is FO-rewriting: every query asked on a KB is reformulated w.r.t. the KB's ontology, so that its answers are computed by the relational evaluation of the query reformulation on the KB's database. Crucially, because FO-rewriting compiles the domain knowledge relevant to queries into their reformulations, query reformulations may be complex and their optimization is the crux of efficiency. In [EHEVGJ24,EHEVGJ23] we devised a novel optimization framework for a large set of OMQA settings that enjoy FO-rewriting: conjunctive queries, i.e., the core select-project-join queries, asked on KBs expressed using datalog±, description logics, existential rules, OWL, or RDF/S. We optimized the query reformulations produced by state-of-the-art FO-rewriting algorithms by computing rapidly, with the help of a KB's database summary, simpler (contained) queries with the same answers that can be evaluated faster by RDBMSs. We shown on a well-established OMQA benchmark that time performance is significantly improved by our optimization framework in general, up to three orders of magnitude in our experiments. We are extending these results on optimizing query answering to optimizing consistency checking, another central data management task.

- *Pattern matching and containment in description logics.* We started a collaboration with colleagues from Université Clermont Auvergne (UCA) to study query answering and query containment for recursive structural queries within the description logic (DL) setting. Such queries can be specified using concept patterns, i.e., concept descriptions containing variables. We have defined the novel $\mathcal{EL}_{\mathcal{RV}}$ DL to model such queries, which extends the classical $\mathcal{EL}$ DL with role variables that represent unknown relationships within concept descriptions. In particular, we equip these role variables with a new refreshing semantics to capture recursion under the well-known greatest fixpoint semantics. We are now investigating the two fundamental reasoning problems of *pattern matching* and *pattern containment.* We have shown that these two problems are EXPTIME-complete. Our main technical results have been derived by establishing a correspondence between the $\mathcal{EL}_{\mathcal{RV}}$ DL and a new variant of variable automata.

4 Software development

4.1 Software development

4.2 FuzViz

Participants: Pierre Nerzic.

[EHEVGJ24] 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`.

[EHEVGJ23] W. El Husseini, C. B. El Vaigh, F. Goasdoué, H. Jaudoin, "OptiRef: Query Optimization for Knowledge Bases", *in: The ACM Web Conference (WWW)*, Austin, United States, April 2023, `https://inria.hal.science/hal-04023665`.

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.

4.3 The SKRID platform

Participants: Adel Aly, Vincent Barreaud, Olivier Pivert, Virginie Thion.

The skrid platform is a digital score library that makes available some Traditional Breton music scores. It is a collaborative effort with Dastum [1], an cultural organization dedicated to preserving and disseminating the cultural heritage of Brittany. This platform is available at `https://shaman.enssat.fr/skrid/`

4.4 Maelis

Participants: Adel Aly, Olivier Pivert, Virginie Thion.

Maelis is a flexible querying module, implemented in the Skrid platform, which allows melodic pattern approximate searching in the graoh-based music score databases of Skrid.

4.5 Musypher

Participants: Adel Aly, 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. It allows illustrating the relevancy (expressiveness, efficiency) of managing music scores over a graph-based data model. Musypher is used as the populating tool of the Skrid platform.

4.6 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

[1] `https://www.dastum.bzh/association/`

the vertices of the graph and ii) the structure of the graph (which may include weighted vertices and edges when the graph is fuzzy).

4.7 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.

4.8 OptiRef

Participants: 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.

4.9 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.

4.10 HeROSim

Participants: Vincent Lannurien, Laurent d'Orazio.

HeROsim is an open source simulation environment for the evaluation of serverless resources allocation and tasks scheduling policies. Its main goal is to relieve researchers from the implementation of discrete-event simulation boilerplate in the context of the dynamic process of cloud orchestration.

4.11 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.

4.12 QRLIT

Participants: Diogo Barbosa, Laurent d'Orazio.

QRLIT is an algorithm that uses the power of quantum computing and reinforcement learning for database index tuning.

4.13 BLOSSOM

Participants: Chanattan Sok, Laurent d'Orazio.

BLOSSOM is a Rust-based experimental platform, to give insights into Wasm-based joins.

4.14 GUESS

Participants: Chanattan Sok, Laurent d'Orazio.

GUESS (monitorinG join qUery Execution in Serverless and Serverful spark) is a monitoring system called GUESS to compare the performance and energy consumption of join query processing on Spark in serverless and serverful environments.

5 Contracts and collaborations

5.1 International Initiatives

5.1.1 Advancing OT Simulations in Cyber Ranges

Participants: Laurent d'Orazio.

Advancing OT Simulations in Cyber Ranges is a B-Monde project, funded by Conseil Régional de Bretagne. The Cyber Innovation Hub, in collaboration with the University of Rennes, Brittany, seeks Agile Cymru funding to recruit a student at Cardiff University to inte-grate and advance existing Operational Technology (OT) simulation environments into Cyber Range platforms. Both universities currently use the commercial Thales (Diateam) Cyber Range, and students will explore its capabilities to simulate near-realistic OT environments. Additionally, the project will investigate integrating these environments into Ludus Cyber Range, an open-source alternative that leverages Ansible roles and templates to create complex OT networks. This approach offers a versatile foundation for developing and customizing OT simula-tions. Depending on the project timeline, student may also enhance these simu-lations by developing user interfaces or adding complexity to bring them closer to realistic OT infrastructures, using Siemens industry processes as a model. Upon completion, the University of Rennes will deploy these projects within their infra-structure to assess their effectiveness. The University of Rennes also plans to apply for B-monde funding to hire additional students, ensuring the continuation and expansion of this pioneering work. This initiative targets the Brittany region and aligns perfectly with Agile Cymru's goals of fostering cross-border innovation, advancing cybersecurity, and developing essential digital skills.

5.1.2 IDENTITY: Big Data management in multimedia exploration in virtual reality

Participants: Laurent d'Orazio.

IDENTITY is a PHC Jules Verne (with Iceland) project.The Icelandic team has been developing a novel approach to multimedia exploration in Virtual Reality (VR), which leads to database queries that require significant query processing resources. The French team has been studying data caching methods, under the umbrella term of Semantic Caching, that hold a promise to significantly improve the query processing cost of the approach. The French team has also been studying several multimedia applications, particularly in the medical domain, which could benefit from the new approach to multimedia exploration. The goal of the exchange is thus twofold: (a) in the short term, to apply the multimedia exploration methodology to novel applications, and (b) in the long term, to study and quantify the impact of Semantic Caching on VR-based multimedia exploration.

5.1.3 GIS4IoRT - Development of a Plug-and-Play Middleware for Integrating Robot Sensor Data with GIS Tools in a Cloud Environment

Participants: Laurent d'Orazio.

The convergence of robotic technologies and sensor networks has generated vast amounts of data with potential applications in several domains like agriculture and urban planning. However, the lack of interoperability between this data and Geographic Information System (GIS) tools poses a challenge. The objective of GIS4IoRT (Geographic Information Systems for Internet of Robotic Things) project is the development of a "plug and play" and cloud-based middleware to bridge this gap, enabling seamless integration and visualization of multidimensional IoRT datasets within GIS environments.

5.2 National Initiatives

5.2.1 EXPAND

Participants: François Goasdoué, Hélène Jaudoin, Claire Vanoni.

The ANR project EXPAND (2025-2030) brings together experts in automated reasoning, data management and knowledge representation from Inria, Univ. Bordeaux, Univ. Montpellier and Univ. Rennes. The aim of the project is to devise practical algorithms for efficient, expressive and explainable knowledge-based data management.

6 Dissemination

6.1 Promoting scientific activities

6.1.1 Scientific Events Selection

Member of Conference Program Committees

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

- AAAI Conference on Artificial Intelligence (AAAI'25)
- Bases de Donnés Avancés (BDA'25)
- European Confenrence on Artificial Intelligence (ECAI'25)
- Extraction et Gestion de Connaissances (EGC'25)
- International Conference on Principles of Knowledge Representation and Reasoning (KR'25)
- International Joint Conference on Artificial Intelligence (IJCAI'25)
- The ACM Web Conference (WWW'25)

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

- International Conference on Flexible Query Answering Systems (FQAS'25)

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

- Atelier sur la Similarité de Données Séquentielles massives : définition, calcul et optimisation (SiDoS@EGC'25)
- International Conference on Scientific and Statistical Database Management (SSDBM'25)
- International Conference on Computer Communications and Networks (ICCCN'25)

- International Conference on Database and Expert Systems Applications (DEXA'25)
- International Conference on Big Data Analytics and Knowledge Discovery (DaWaK'25)
- International Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KMIS'25)
- Colloque sur l'Optimisation et les Systèmes d'Information (COSI'25)

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

- International Conference on Flexible Query Answering Systems (FQAS'25)
- IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'25)
- European Society for Fuzzy Logic and Technology Conference (EUSFLAT'25)
- World Congress of the International Fuzzy Systems Association / Conference of the North American Fuzzy Information Society (IFSA-NAFIPS'25)
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA'25)

Reviewer

6.1.2 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 Uncertainty, Fuzziness and Knowledge-based Systems.

Reviewer - Reviewing Activities
Laurent d'Orazio was a reviewer for:

- IEEE Internet Computing
- Future Generation Computer Systems (FGCS)
- Distributed and Parallel Databases (DAPD)
- Springer Nature: SN Computer Science (SNCS).

6.1.3 Invited Talks

Laurent d'Orazio, RESILIENCE Workshop at EPHE Paris 23-25 September 2025: Big Data in museum, a brief history of cloud data management and perspectives.

6.1.4 Leadership within the Scientific Community

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

François Goasdoué is a member of the Steering Committee of the CNRS GDR MaDICS since 2024.

Olivier Pivert is a member of the permanent Steering Committees of

- the French-speaking conference "Rencontres Francophones sur la Logique Floue et ses Applications" (LFA)
- the International Symposium on Methodologies for Intelligent Systems (ISMIS)
- the International Conference on Flexible Query-Answering Systems (FQAS).

6.1.5 Scientific Expertise

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

Olivier Pivert is an expert for the Czech Science Foundation.

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

6.1.6 Research Administration

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 council of ENSSAT, since 2022.

Laurent d'Orazio is the co-head of the International Relationalships of IRISA, 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: Olivier Pivert teaches a course on *AI for Database Querying*, 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, Storage and Querying of Musical Score Databases, advisor: Virginie Thion.
- PhD in progress: Claire Vanoni, Comparison of descriptions w.r.t. domain knowledge, advisors: François Goasdoué and Hélène Jaudoin.
- PhD in progress: Truong Tran Phan, Big joins in serverful and serverless computing, advisor: Laurent d'Orazio.

6.2.3 Juries

François Goasdoué

- HDR, referee, Mohamed-Amine Baazizi, Université de Paris.
- PhD, referee, Hassan Abdallah, Université de Tours.
- PhD, president, William Charles, Université de Toulouse.

Laurent d'Orazio

- PhD, referee, Nour Kired, Université de Toulouse.
- PhD, referee, Oleksandr Lytvyn, Slovak University of Technology in Bratislava.
- PhD, referee, Ján Mach, Slovak University of Technology in Bratislava.
- PhD, referee, Fagniné Alassane Coulibaly, Université Clermont Auvergne.

7 Bibliography

A. Aly, O. Pivert, V. Thion, "A Flexible Framework for Transposition-Aware Querying of a Musical Score Database (Best Paper Award RCIS 2025)", *in: Research Challenges in Information Science (RCIS'25)*, Seville, Spain, May 2025, https://hal.science/hal-05011084.

A. Aly, O. Pivert, V. Thion, “Maelis4Skrid: an Approximate Query Engine for an Online Graph-Based Musical Score Library”, *in: Companion Proceedings of the ACM Web Conference 2025, WWW 2025*, Sydney, Australia, April 2025, `https://hal.science/hal-04904525`.

A. Boulahmel, F. Djelil, G. Smits, “Investigating Self-Regulated Learning Measurement Based on Trace Data: A Systematic Literature Review”, *Technology, Knowledge and Learning*, January 2025, `https://hal.science/hal-04907756`.

M. Buron, F. Goasdoué, I. Manolescu, M.-L. Mugnier, “Obi-Wan: Ontology-Based RDF Integration of Heterogeneous Data”, *in: VLDB 2020 - 46th International Conference on Very Large Data Bases*, Tokyo, Japan, August 2020, `https://inria.hal.science/hal-02921434`.

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, “OptiRef: Query Optimization for Knowledge Bases”, *in: The ACM Web Conference (WWW)*, Austin, United States, April 2023, `https://inria.hal.science/hal-04023665`.

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`.

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`.

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, “myCADI: my Contextual Anomaly Detection using Isolation”, *in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM ’24)*, Boise, ID, United States, October 2024, `https://hal.science/hal-04743207`.

Articles in referred journals and book chapters

[1] A. Boulahmel, F. Djelil, G. Smits, "Investigating Self-Regulated Learning Measurement Based on Trace Data: A Systematic Literature Review", *Technology, Knowledge and Learning*, January 2025, `https://hal.science/hal-04907756`.

Publications in Conferences and Workshops

[2] A. Aly, O. Pivert, V. Thion, "A Flexible Framework for Transposition-Aware Querying of a Musical Score Database (Best Paper Award RCIS 2025)", *in: Research Challenges in Information Science (RCIS'25)*, Seville, Spain, May 2025, `https://hal.science/hal-05011084`.

[3] A. Aly, O. Pivert, V. Thion, "Fuzzy Retrieval of Musical Scores Based on Melodic Patterns", *in: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'25)*, Reims, France, July 2025, `https://hal.science/hal-05050369`.

[4] A. Aly, O. Pivert, V. Thion, "Maelis4Skrid: an Approximate Query Engine for an Online Graph-Based Musical Score Library", *in: Companion Proceedings of the ACM Web Conference 2025, WWW 2025*, Sydney, Australia, April 2025, `https://hal.science/hal-04904525`.

[5] A. Aly, O. Pivert, V. Thion, "Where Graphs Meet Fuzzy Logic - A DBMS-Centered Engine for Polyphonic Music Matching in Score Databases", *in: EDBT 2026 - 29th International Conference on Extending Database Technology*, Tampere, Finland, March 2026, `https://hal.science/hal-05485122`.

[6] A. Drissi, L. d'Orazio, M. Damigos, "Scalable Structural Similarity Analysis of JSON documents Using MapReduce", *in: IEEE International Conference on Fuzzy Systems*, Reims (51), France, July 2025, `https://hal.science/hal-05071674`.

[7] O. Shahbaz Khan, A. Duane, H. Hasnan, N. Le Blavec, P. Ouvrard, J. Verdon, L. d'Orazio, C. Thierry, B. Þ. Jónsson, "Multi-Dimensional Exploration of Media Collection Metadata", *in: MMM 2025 - International Conference on Multimedia Modeling*, Nara, Japan, January 2025, `https://hal.science/hal-05136933`.

[8] P.-A.-T. Tran, L. d'Orazio, T.-C. Phan, L. Gruenwald, "Energy and Performance Evaluation of Serverless and Serverful Models on Spark for Database Join Operations", *in: International Conference on Database and Expert Systems Applications (DEXA)*, Bangkok, Thailand, August 2025, `https://hal.science/hal-05136844`.

[9] R. Wrembel, J.-P. Kasprzyk, R. Billen, S. Bimonte, L. d'Orazio, D. Sacharidis, P. Skrzypczyński, "On Integrating Robotic Data with GIS Tools in a Cloud Environment (Application Paper)", *in: Data Analytics solutions for Real-LIfe APplications (DARLI-AP@EDBT)*, *1*, p. 7, Barcelone, Spain, March 2025, `https://hal.science/hal-05288001`.