# SHAMAN : Symbolic and Human-centric view of dAta MANagement

François Goasdoué

# SHAMAN : Symbolic and Human-centric view of dAta MANagement

*IRISA/SHAMAN*
*Université de Rennes*
*http://www-shaman.irisa.fr*

**François GOASDOUÉ**
*Responsable d'équipe*
*fg@irisa.fr*

## Membres

- Ludivine DUROYON, Doctorat
- François GOASDOUÉ, PR
- Hélène JAUDOIN, MCF
- Trung-Dung LE, Doctorat
- Ludovic LIÉTARD, MCF
- Pierre NERZIC, MCF
- Laurent d'ORAZIO, PR
- Olivier PIVERT, PR
- Daniel ROCACHER, PR
- Amit SHUKLA, Post-doctorat
- Grégory SMITS, MCF
- Virginie THION, MCF
- Thi To Quyen TRAN, Doctorat
- Van Hoang TRAN, Doctorat

## Mots-clés

- Apprentissage automatique
- Bases de données
- Logique floue
- Qualité des données
- Raisonnement automatique
- Représentation des connaissances
- Traitement automatique des langues

## Thématique générale de l'équipe

SHAMAN est une équipe du département «Data and Knowledge Management» de l'Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) de l'Université de Rennes 1. Cette équipe, créée en 2014 et renouvelée en 2019, est une équipe de *Bases de Données* dont les travaux ont de fortes interactions avec le domaine de l'*Intelligence Artificielle*. Ces travaux s'organisent autour de trois axes de recherche.

## Gestion de données à l'aide de connaissances

La gestion de données à l'aide de connaissances peut être vue comme la convergence des travaux sur la gestion de données en Bases de Données et en Intelligence Artificielle. Comme en Bases de Données, il s'agit de gérer *efficacement* de grands volumes de données. Toutefois, les données ne sont pas décrites par des schémas rigides fondés sur des structures mathématiques (tables relationnelles, arbres XML, etc), mais par des modèles conceptuels des domaines d'application appelés *ontologies* et exprimés dans des formalismes issus ou proches de ceux de la Représentation des Connaissances (logiques de description, règles existentielles, etc). L'objectif est de faciliter l'accès à la gestion de données, notamment aux utilisateurs non-informaticiens, en leur montrant les données et en exprimant leurs opérations sur celles-ci au travers de modèles plus intelligibles. Du point de vue de l'Intelligence Artificielle, cette gestion de données s'apparente à la gestion de bases de connaissances contenant *beaucoup* de données décrites par des langages *pragmatiques* (à l'expressivité limitée), afin d'envisager des systèmes de gestion de bases de connaissances dont les performances sont comparables à celles des systèmes de gestion de bases de données. Un point clé est la nécessité de *raisonnement automatique* pour effectuer les tâches de gestion de don-

nées (consistance, interrogation, mise-à-jour). Cette nouvelle forme de gestion de données a connu un essor sans précédent ces dernières années grâce aux *standards RDF et OWL2 du W3C*, qui ont été rapidement adoptés par de nombreuses communautés d'utilisateurs (biologie, journalisme, médecine, science de l'information et des bibliothèques, etc) : RDF est un modèle de données graphes avec un langage d'ontologie peu expressif ; OWL2 repose sur des logiques de description et permet l'utilisation d'ontologies plus riches que celles de RDF, qui en particulier permettent d'exprimer des contraintes d'intégrité.

Dans l'équipe SHAMAN, nous nous intéressons à la conception de systèmes pour RDF et le dialecte QL d'OWL2 fondé sur la logique de description légère DL-lite$_{\mathcal{R}}$, dédiée à la gestion de données volumineuses et sémantiquement riches.

Concernant RDF, nous étudions la *gestion efficace de données RDF*, y compris en présence de mises-à-jour, dans le cadre des systèmes centralisés [24, 11, 10] ou massivement parallèles [8, 22]. Nous nous intéressons également à l'*intégration efficace, en RDF, de sources de données hétérogènes* dans les architectures classiques d'entrepôts de données [16] ou de médiateurs [9]. Enfin, nous travaillons aussi sur des tâches non-standards de raisonnement permettant aux utilisateurs de mieux comprendre les données de leurs systèmes ou comment ces données sont utilisées. Par exemple, nous faisons du *résumé de données RDF* et des visualisations associées [15, 21], ainsi que de la comparaison de bases RDF ou des requêtes qui y accèdent par *apprentissage automatique de leurs points communs* [25, 26].

Concernant OWL2 QL, nous étudions la *gestion efficace de données DL-lite*$_{\mathcal{R}}$ dans les systèmes centralisés [12, 13], y compris lorsque les données sont *inconsistantes* avec les contraintes d'intégrité exprimées (Projet ANR Practical AlGorithms for Ontology-based Data Access). Pour cela, nous adoptons des sémantiques non-standards tolérantes à l'inconsistance, fondées sur les réparations possibles des inconsistances, afin de calculer et de répondre aux requêtes des utilisateurs, que nous sommes aussi capables d'*expliquer* [4]. Enfin, nous étudions la *réparation des inconsistances de données DL-lite*$_{\mathcal{R}}$ en fonction des réponses ou des non réponses aux requêtes posées [3].

Ces travaux ont des applications en *analyse multidimensionnelle de données ouvertes* (Projet DGA Rapid Open-Data INtelligence) et en *fact checking* [5, 23, 18] (Projets ANR ContentCheck et INRIA Project Lab iCoda avec Le Monde et Ouest France).

## Gestion de données flexible, coopérative et guidée par la qualité

Au delà de disposer de systèmes capables de gérer efficacement des données, il est aussi important que ces systèmes soient simples d'utilisation.

L'équipe SHAMAN travaille sur la *flexibilité* des systèmes qui permet de prendre en compte des préférences utilisateur. Pour représenter et raisonner sur ces préférences dans les langages de requête, nous utilisations la *théorie des sous-ensembles flous* [28] afin d'enrichir les algèbres et langages standards de bases de données [6, 30] ainsi que les systèmes associés [34, 29] (projet DGA ODIN).

Nous cherchons également à rendre les systèmes de gestion de données *coopératifs*, c'est-à-dire capables d'interagir avec l'utilisateur afin d'exprimer [36] ou d'affiner ses besoins en information, par exemple quand une requête n'a aucune réponse ou au contraire en a une pléthore [35]. En complément de ces travaux, nous développons des méthodes efficaces de résumé linguistique de données afin que les utilisateurs puissent mieux appréhender les données

d'un système [33, 32].

Enfin, les données manipulées par les utilisateurs souffrent souvent de problèmes de qualité. L'équipe SHAMAN s'intéresse à la représentation des données incertaines et au raisonnement sur de telles données, notamment à l'aide de la théorie des possibilités [7]. Pour quantifier et décrire le caractère imparfait des données, de nombreux travaux sont également réalisés autour de la modélisation de métadonnées de qualité (fraîcheur, exactitude, complétude, etc.) pour décrire les données stockées [31, 19].

Ces travaux ont des applications en gestion et analyse de données ouvertes (projet DGA ODIN) et musicales (projet Mastodon GIOQOSO), ou encore l'analyse du suivi de marchandises dans le transport maritime (projet CREDOC du pôle de compétitivité Images & Réseaux).

## Gestion de données massives

Ces dernières années, les besoins en gestion de données massives (big data) ont conduit à de nouvelles architectures dites nuages (clouds), ainsi qu'au développement associé de systèmes de fichiers à grande échelle comme Google File System (GFS) [20], ou encore d'environnements de traitement parallèle tels que MapReduce [17] et Spark [2]. Ceci a permis de concevoir de nouveaux systèmes de gestion de données massivement parallèles tels que Hive [37] ou encore Flink [14].

Dans l'équipe SHAMAN, nous nous intéressons à l'exécution et l'optimisation de requêtes au sein de ces systèmes. Nous étudions l'interrogation efficace de données sur des fédérations de nuages [27], des grandes jointures floues [39], ou encore du traitement de requêtes sur des données chiffrées au sein des nuages [38]. Le lien avec l'Intelligence Artificielle est double dans ce contexte. D'une part, le passage à l'échelle obtenu nous permet d'appliquer des techniques issues de l'Intelligence Artificielle, par exemple du data mining [1], sur des volumes d'informations plus importants et/ou d'obtenir de meilleures performances. D'autre part, l'Intelligence Artificielle est vue comme une composante de nos solutions, notamment pour l'aide à la décision dans l'optimisation de requêtes par des techniques d'apprentissage automatique [40].

Ces travaux ont des applications dans de nombreux domaines, notamment la santé (projet NSF MOCCAD), la cyber-sécurité (cofinancement Lannion Trégor Communauté et Région Bretagne pour le projet SERBER) ou encore l'environnement (PHC-SIAM AGRIWATCH).

## Références


[1] Sabeur Aridhi, Laurent d'Orazio, Mondher Maddouri, and Engelbert Mephu Nguifo. Density-based data partitioning strategy to approximate large-scale subgraph mining. *Inf. Syst.*, 48 :213–223, 2015.

[2] Michael Armbrust, Reynold S Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K Bradley, Xiangrui Meng, Tomer Kaftan, Michael J Franklin, Ali Ghodsi, and Matei Zaharia. Spark SQL : Relational Data Processing in Spark. In *SIGMOD*, pages 1383–1394, Melbourne, Victoria, Australia, 2015.

[3] Meghyn Bienvenu, Camille Bourgaux, and François Goasdoué. Query-driven repairing of inconsistent dl-lite knowledge bases. In *IJCAI*, pages 957–964, 2016.

[4] Meghyn Bienvenu, Camille Bourgaux, and François Goasdoué. Computing and explaining query answers over inconsistent dl-lite knowledge bases. *J. Artif. Intell. Res.*, 64 :563–644, 2019.

[5] Raphaël Bonaque, Tien Duc Cao, Bogdan Cautis, François Goasdoué, J. Lete-

lier, Ioana Manolescu, O. Mendoza, S. Ribeiro, Xavier Tannier, and Michaël Thomazo. Mixed-instance querying : a lightweight integration architecture for data journalism. *PVLDB*, 9(13) :1513–1516, 2016.

[6] Patrick Bosc and Olivier Pivert. Sqlf : a relational database language for fuzzy querying. *IEEE transactions on Fuzzy Systems*, 3(1) :1–17, 1995.

[7] Patrick Bosc and Olivier Pivert. Modeling and querying uncertain relational databases : A survey of approaches based on the possible worlds semantics. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 18(05) :565–603, 2010.

[8] Francesca Bugiotti, Jesús Camacho-Rodríguez, François Goasdoué, Zoi Kaoudi, Ioana Manolescu, and Stamatis Zampetakis. SPARQL query processing in the cloud. In *Linked Data Management.*, pages 165–192. Chapman and Hall/CRC, 2014.

[9] Maxime Buron, François Goasdoué, Ioana Manolescu, and Marie-Laure Mugnier. Rewriting-Based Query Answering for Semantic Data Integration Systems. In *BDA*, 2018.

[10] Maxime Buron, François Goasdoué, Ioana Manolescu, and Marie-Laure Mugnier. Reformulation-based query answering for RDF graphs with RDFS ontologies. In *ESWC*, pages 19–35, 2019.

[11] Damian Bursztyn, François Goasdoué, and Ioana Manolescu. Optimizing reformulation-based query answering in RDF. In *EDBT*, pages 265–276, 2015.

[12] Damian Bursztyn, François Goasdoué, and Ioana Manolescu. Optimizing FOL reducible query answering : Understanding performance challenges. In *ISWC*, 2016.

[13] Damian Bursztyn, François Goasdoué, and Ioana Manolescu. Teaching an RDBMS about ontological constraints. *PVLDB*, 9(12) :1161–1172, 2016.

[14] Paris Carbone, Asterios Katsifodimos, Stephan Ewen, Volker Markl, Seif Haridi, and Kostas Tzoumas. Apache Flink : Stream and Batch Processing in a Single Engine. *IEEE Data Engineering Bulletin*, 38(4) :28–38, 2015.

[15] Sejla Cebiric, François Goasdoué, and Ioana Manolescu. A framework for efficient representative summarization of RDF graphs. In *ISWC*, 2017.

[16] Dario Colazzo, François Goasdoué, Ioana Manolescu, and Alexandra Roatis. RDF analytics : lenses over semantic graphs. In *WWW*, pages 467–478, 2014.

[17] Jeffrey Dean and Sanjay Ghemawat. MapReduce : simplified data processing on large clusters. *Communications of the ACM*, 51(1) :107–113, 2008.

[18] Ludivine Duroyon, François Goasdoué, and Ioana Manolescu. A linked data model for facts, statements and beliefs. In *WWW Workshop*, pages 988–993, 2019.

[19] Francesco Foscarin, David Fiala, Florent Jacquemard, Philippe Rigaux, and Virginie Thion. Gioqoso, an online Quality Assessment Tool for Music Notation. In *TENOR*, 2018.

[20] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The Google file system. In *Symposium on Operating Systems Principles (SOSP)*, pages 29–43, Bolton Landing, NY, USA, 2003.

[21] François Goasdoué, Pawel Guzewicz, and Ioana Manolescu. Incremental structural summarization of RDF graphs. In *EDBT*, pages 566–569, 2019.

[22] François Goasdoué, Zoi Kaoudi, Ioana Manolescu, Jorge-Arnulfo Quiané-Ruiz,

and Stamatis Zampetakis. Cliquesquare : Flat plans for massively parallel RDF queries. In *IEEE ICDE*, pages 771–782, 2015.

[23] François Goasdoué, Konstantinos Karanasos, Yannis Katsis, Julien Leblay, Ioana Manolescu, and Stamatis Zampetakis. Fact checking and analyzing the web. In *ACM SIGMOD*, pages 997–1000, 2013.

[24] François Goasdoué, Ioana Manolescu, and Alexandra Roatis. Efficient query answering against dynamic RDF databases. In *EDBT*, pages 299–310, 2013.

[25] Sara El Hassad, François Goasdoué, and Hélène Jaudoin. Learning commonalities in RDF. In *ESWC*, pages 502–517, 2017.

[26] Sara El Hassad, François Goasdoué, and Hélène Jaudoin. Learning commonalities in SPARQL. In *ISWC*, pages 278–295, 2017.

[27] Trung-Dung Le, Verena Kantere, and Laurent d'Orazio. Optimizing DICOM data management with NSGA-G. In *DOLAP, EDBT workshop*, 2019.

[28] Olivier Pivert and Patrick Bosc. *Fuzzy preference queries to relational databases*. World Scientific, 2012.

[29] Olivier Pivert, Olfa Slama, and Virginie Thion. An extension of sparql with fuzzy navigational capabilities for querying fuzzy rdf data. In *FUZZ-IEEE*, pages 2409–2416. IEEE, 2016.

[30] Olivier Pivert, Virginie Thion, Hélène Jaudoin, and Grégory Smits. On a fuzzy algebra for querying graph databases. In *IEEE ICTAI*, pages 748–755. IEEE, 2014.

[31] Philippe Rigaux and Virginie Thion. Quality awareness over graph pattern queries. In *IDEAS*, pages 90–97, 2017.

[32] Grégory Smits, Pierre Nerzic, Olivier Pivert, and Marie-Jeanne Lesot. Efficient generation of reliable estimated linguistic summaries. In *FUZZ-IEEE*, pages 1–8. IEEE, 2018.

[33] Grégory Smits and Olivier Pivert. Linguistic and graphical explanation of a cluster-based data structure. In *SUM*, pages 186–200. Springer, 2015.

[34] Grégory Smits, Olivier Pivert, and Thomas Girault. Reqflex : fuzzy queries for everyone. *PVLDB*, 6(12) :1206–1209, 2013.

[35] Grégory Smits, Olivier Pivert, and Allel Hadjali. Fuzzy cardinalities as a basis to cooperative answering. In *Flexible Approaches in Data, Information and Knowledge Management*, pages 261–289. Springer, 2014.

[36] Grégory Smits, Olivier Pivert, Hélène Jaudoin, and François Paulus. Aggrego search : Interactive keyword query construction. In *EDBT*, pages 636–639, 2014.

[37] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. Hive - A Warehousing Solution Over a Map-Reduce Framework. *PVLDB*, 2(2) :1626–1629, 2009.

[38] Hoang Van Tran, Tristan Allard, Laurent d'Orazio, and Amr El Abbadi. Range query processing for monitoring applications over untrustworthy clouds. In *EDBT*, pages 666–669, 2019.

[39] Thi-To-Quyen Tran, Thuong-Cang Phan, Anne Laurent, and Laurent d'Orazio. Improving hamming distance-based fuzzy join in mapreduce using bloom filters. In *FUZZ-IEEE*, pages 1–7, 2018.

[40] Chenxiao Wang, Zachary Arani, Le Gruenwald, and Laurent d'Orazio. A vision of a decisional model for re-optimizing query execution plans based on machine learning techniques. In *DOLAP EDBT workshop*, 2019.