

1 Equipe HADAS - Axe Traitement de Données et de Connaissances à Grande Echelle



1.1 Scientific Presentation

Research group: The HADAS (Heterogeneous Autonomous Distributed data Services) group was founded in october 2005 as a new team for the LIG laboratory. It actually follows the STORM team, directed by M. Adiba until 1996. Over the years, the group proposed an evolution of the scientific vision of a database management systems as a semantic-based infrastructure for managing ubiquitous and heterogeneous data services. The team currently includes 9 permanent research people (2 full professors, 5 Associate Professors and two CNRS Research Scientists), a research engineer and 15 PHD students and post-docs. During this period we hired 4 people. The following table lists the permanent people during the period.

Name	First name	Function	Institution	Arrival date
Adiba	Michel	External collaborator	UJF	Oct 2010
Amer Yahia	Sihem	Research Director	CNRS	Dec 2011
Bobineau	Christophe	Associate Professor	Grenoble INP	Sep 2003
Collet	Christine	Full Professor	Grenoble INP	Sep 1999
Dubl�	Etienne	ITA	CNRS	May 2011
Ibrahim	Noha	Associate Professor	Grenoble INP	Sep 2010
Jouanot	Fabrice	Associate Professor	UJF	Sep 2003
Leroy	Vincent	Associate Professor	UJF	Sep 2012
Rousset	Marie-Christine	Full Professor	UJF	Sep 2005
Termier	Alexandre	Associate Professor	UJF	Sep 2007
Vargas-solar	Genoveva	Research Scientist	CNRS	Jan 2002

Group evolution: Research activities on services have been reinforced in the group with the arrival of Noha Ibrahim, associate professor who joined the team in September 2010. Then, the arrival of Sihem Amer-Yahia, Research Director (Dec 2011) and Vincent Leroy, Associate Professor UJF / Research scientist (Sept 2012) brought new areas of research in the group centered around large scale data processing on the social Web. Etienne Dubl , Research Engineer (Mai 2011) helps us in developing and finalizing some of our prototypes. He also brought expertise in new kind of sensor networks and of file systems.

Research description: Technological changes have reduced the cost of creating, capturing, managing and storing information to a sixth of what it was in 2005. This allowed a scale change in the size and distribution of data, the number of connected devices, and the number of users. Data can be numerical data coming from sensors, scientific data or personal data coming from heterogeneous and largely distributed data sources. They cannot be handled anymore by centralized data management systems with pre-established schemas. Modern data and knowledge management systems requires new data models, new services and algorithms that must be largely distributed and deployed over different types of large scale systems (grids, peer-to-peer networks, sensor networks, web infrastructures). The HADAS group has revisited and extended standard database systems to deal with dynamic and distributed data, to design scalable data mining algorithms, and to combine data and knowledge at large scale for accessing data. Semantics is at the heart of this approach as it is used at all levels of the process of designing or composing data services for handling autonomy, dynamic behavior and heterogeneity of both users and data sources.

The activities of the group during these past years have been centered on the following themes:

- Accessing data in large-scale systems: a first aspect concerns query optimization in distributed and dynamic systems ; a second aspect deals with mining large amounts of data to extract patterns of interest.
- Composing data services in a dynamic way: we investigate models, algorithms and tools for coordinating services with non functional properties (contracts) and for providing access to heterogeneous data coming from services

- Reasoning on data semantics: we investigate different models and algorithms for querying data (or resources) through possibly heterogeneous and distributed ontologies.



We participated to the Optimacs and Ubiquet ANR projects that bridge the gap between data management and (web) services and, also between networks and data management. We have been involved in the Continuum, Dataring and Qualinca projects that exploit the semantics for relating data, devices and services in ubiquitous environments and peer-to-peer data management systems. In the PAGODA project, we collaborate with the LaDAF (Laboratoire d'Anatomie de Grenoble) to enrich anatomical 3D graphical models with ontologies using Semantic Web and Linked Data technologies. We collaborate with industry in several projects. The Datalyse project involves several partners interested in Big Data (Business & Decision, Eolas, supermarkets), its objective is to build a smart warehouse demonstrator for the collection, analysis, integration and enrichment of heterogeneous Big Data of type "Big Data User" (UBD) or from machines, of type "Big Data Monitoring" (MBD). We also have established strong ties with STMicroelectronics, especially in the context of the SoCTrace project for analyzing execution traces with data mining methods. In the framework of the ADEME SoGrid project we collaborate (with more than 10 companies) on event flow management systems for Smartgrids. In addition, we recently started ALICIA, an ANR project on crowdsourcing and are strongly involved in the creation of a GDR on Big Data and Data Sciences.

Results of our research have direct impact on applications dealing with huge amounts of data and resources largely distributed in pervasive environments, such as data spaces, smart grids and smart buildings, hardware and software observation and social networks.

1.2 Scientific and Technological Results



1.2.1 Accessing data in large-scale systems

Accessing data concerns several aspects of large scale systems: number of resources, data volume and data complexity. It basically means using declarative queries that are optimized based on system characteristics. Data mining is another way to query large quantities of data, by extracting interesting patterns from them. Such patterns provide meaningful abstractions of raw data, which are thus more appropriate for data analysis. Globally, the difficulty for evaluating queries efficiently on nowadays applications motivates this work to revisit traditional query optimization techniques. The following presents these two aspects of accessing data in the large. It also focuses on works done on querying the social web.

1- CBR query optimization

Our research contributes to the development of new distributed query optimization techniques. It relies on the adaptation of machine learning, more precisely Case-Based Reasoning(CBR), and pseudo random search space exploration (also exploiting the case base) to produce efficient query execution plans according to application specific optimization objectives expressed over resource consumption (e.g. time, energy, number of messages). The query plan generation considered multiple optimization objectives customizable to application requirements (QoS based Hybrid Query optimization). This research led to the following original contributions:

- A query optimization approach that uses cases generated from the evaluation of similar past queries. A query case comprises: (i) the query (the problem), (ii) the query plan (the solution) and (iii) the measures of computational resources consumed during the query plan execution (the evaluation of the solution).
- A query plan generation process [1] that uses classical query optimization heuristics and makes decisions randomly when information on data is not available (e.g. for ordering joins, selecting algorithms or choosing message exchange protocols). This process also exploits the CBR principle for generating plans for sub-queries, thus accelerating the learning of new cases.
- A Simulation Platform [2] allowing to experiment distributed query optimization and rule-based programs over a set of distributed data-enabled devices hosting virtual machines(VM). A VM integrates a query optimization engine [3] implementing the above techniques.

Main contracts:

- UBIQUEST(2009-2013) ANR- Programme BLANC. Ubiquitous Quest: a declarative approach for integrated network and data management in wireless multi-hop networks.
- OPTIMACS (2009-2012) ANR-Programme ARPEGE. Service composition based framework for optimizing queries.
- CAISES (2012 - 2015) European Union FP7, IRESES program. Observation and industrial management on the cloud.
- SOGrid (2013-2017) ADEME Le réseau électrique de demain.

Key references:

- [1] Lourdes Martinez, Christine Collet, Christophe Bobineau, Etienne Dublé. The QOL approach for optimizing distributed queries without complete knowledge. IDEAS, 91-99, 2012
- [2] Ahmad Ahmad-Kassem, Christophe Bobineau, Christine Collet, Etienne Dublé, Stéphane Grumbach, FudaMa, Lourdes Martinez, Stéphane Ubéda. UBIQUEST, for rapid prototyping of networking applications. IDEAS, 187-192, 2012
- [3] Lourdes Martinez, Christine Collet, Christophe Bobineau and Etienne Dublé. CoBRa for optimizing global queries. BDA, 2013

2- Data mining

Data mining is the automatic extraction of unknown and potentially interesting information from large quantities of data. One of the major fields of data mining consists in discovering patterns occurring frequently (i.e. more than a given threshold) in data.

This research focused on improving frequent pattern mining algorithms, both to make them more scalable and to apply them to real data analysis contexts. Main results are:

- From the scalability point of view, we acquired a strong expertise on exploiting multicore processors for pattern mining [2, 5]. The proposed algorithms are also based on the notion of closed patterns [2, 4, 5], reducing the output size (hence the computation time) without loss of information.
- These works culminated with the proposition of ParaMiner [2], the first parallel and generic algorithm for mining closed patterns.
- From an application point of view, works have been done on the analysis of execution traces, in collaboration with STMicroelectronics. They have improved the way to discover periodic behaviors and their disruption in traces [4], to rewrite a trace with a few significant sequences of events [3], and to automatically discover hotspots of memory contention in a parallel code [1].

Main contracts:

- FUI SoCTrace that aims to develop a set of methods and tools based on traces of execution produced by multi-core embedded applications.

Key references:

- [1] Sofiane Lagraa, Alexandre Termier, Frédéric Pétrot: Data mining MPSoC simulation traces to identify concurrent memory access patterns. DATE 2013: 755-760
- [2] Benjamin Nègrevergne, Alexandre Termier, Marie-Christine Rousset, Jean-Francois Méhaut: ParaMiner: a generic pattern mining algorithm for multi-core architectures, Data Mining and Knowledge Discovery, 2013
- [3] Christiane Kamdem Kengne, Leon Constantin Fopa, Alexandre Termier, Noha Ibrahim, Marie-Christine Rousset, Takashi Washio, Miguel Santana: Efficiently rewriting large multimedia application execution traces with few event sequences. KDD 2013: pp 1348-1356
- [4] Patricia Lopez-Cueva, Aurélie Bertaux, Alexandre Termier, Jean-Francois Méhaut, Miguel Santana: Debugging Embedded Multimedia Application Traces through Periodic Pattern Mining, EMSOFT 2012: pp 13-22
- [5] Trong Dinh Thac Do, Anne Laurent, Alexandre Termier: PGLCM: Efficient Parallel Mining of Closed Frequent Gradual Itemsets, ICDM 2010: pp 138-147

[6] Sofiane Lagraa, Alexandre Termier, Frédéric Pétrot: Scalability Bottlenecks Discovery in MP-SoC Platforms Using Data Mining on Simulation Traces, Design Automation and Test in Europe Conference (DATE), 2014, to appear.

3- Social Web Data access

Research has been done on new exploration problems to find useful user groups in collaborative rating datasets [1,2,4] and useful information in online news [3,5]. Our formulation of exploration as an optimization problem where various dimensions such as similarity, diversity, and coverage are optimized, leads to reductions from well-known problems and adaptations of well-established algorithms. Large-scale user studies have been conducted to verify the effectiveness of our findings. The current research direction is to blend efficient mining with exploration and to develop an evaluation methodology for large-scale information exploration.

Social Web Data access also concerns optimization. Data is stored within data centers, which constitute distributed systems. It is therefore important to optimize communications between the machines to avoid saturating the network equipment. A first work considered the problem of data routing between users of social networks. The key idea was to identify hubs that aggregate data from several sources and reduce the number of messages exchanged [6]. A second work considers the problem of data placement in hierarchical network structures. A reactive algorithm monitors data access patterns to identify locations in which new replicas of data should be deployed to reduce routers saturation [7].

Main contracts:

- CrowdHealth (2014), MASTODONS CNRS project on Crowdsourcing.
- ALICIA (2014-2017), Agence Nationale de la Recherche, France, Programme BLANC.
- Datalyse(2013-2016): Big Data Models and Algorithms, Investissement d'Avenir.
- AGIR Big join (2013-2016): Programme AGIR Grenoble INP and University Joseph Fourier, Modèles et algorithmes pour les jointures de Big Data sur Map-Reduce.

Key references:

- [1] Behrooz Omidvar Tehrani, Sihem Amer-Yahia, Alexandre Termier, Aurélie Bertaux, Eric Gaussier, Marie-Christine Rousset: Towards a Framework for Semantic Exploration of Frequent Patterns. IMMoA 2013: 7-14 (workshops)
- [2] Mikalai Tsytarau, Sihem Amer-Yahia, Themis Palpanas: Efficient sentiment correlation for large-scale demographics. SIGMOD Conference 2013: 253-264
- [3] Sofiane Abbar, Sihem Amer-Yahia, Piotr Indyk, Sepideh Mahabadi: Real-time recommendation of diverse related articles. WWW 2013: 1-12
- [4] Mahashweta Das, Saravanan Thirumuruganathan, Sihem Amer-Yahia, Gautam Das, Cong Yu: Who Tags What? An Analysis Framework. PVLDB (11): 1567-1578 (2012)
- [5] Demo: Sihem Amer-Yahia, Samreen Anjum, Amira Ghenai, Aysha Siddique, Sofiane Abbar, Sam Madden, Adam Marcus, Mohammed El-Haddad: MAQSA: a system for social analytics on news. SIGMOD Conference 2012: 653-656
- [6] Aristides Gionis, Flavio P. Junqueira, Vincent Leroy, Marco Serafini and Ingmar Weber: Piggybacking on social networks. In Proceedings of the 39th International Conference on Very Large Databases (VLDB), pages 409-420, 2013
- [7] Xiao Bai, Arnaud Jégou, Flavio P. Junqueira and Vincent Leroy: DynaSoRe: Efficient In-Memory Store for Social Applications. In Proceedings of the 14th International Middleware Conference (Middleware) pages 425-444, 2013

1.2.2 Composing data services on the fly

Composing services exported by different organizations is a key issue when building large scale and data-intensive systems. Composition must take into account the characteristics of execution environments (e.g., memory and computing, and network capabilities) to dynamically compose services, and then to adapt compositions depending on the availability and evolution of services. Compositions can be

executed on platforms providing unlimited resources through a "pay as U go model", aware of energy consumption or services reputation, provenance, availability, and reliability[2]. These features guide the way compositions are specified and executed for fulfilling given user requirements and preferences[3]. These properties are modeled as non functional aspects and QoS (quality of service) criteria that can provide guarantees to the execution of compositions and to the way results are delivered.

Our research in this topic contributes to the construction of service based data management systems as service compositions. Once data management is delivered as a service, it can have associated non-functional properties. We proposed methodologies, algorithms, languages and tools for designing and executing service compositions with non-functional properties expressed as policies.

We applied our approach for the efficient evaluation of queries as coordinations of services, including data and computing services. This lead to the following results:

- an Hybrid query model for expressing queries as data service coordinations based on workflows. The approach uses the abstract state machines (ASM) formalism for defining the model[1].
- a query language HSQL (Hybrid Services Query Languages) associated to the hybrid query model and the language MQLiST (Mashup Query Spatio Temporal Language) for integrating hybrid query results in a mashup. Both are extensions of SQL[5].
- an algorithm BP GYO for generating the query workflow that implements a query expressed in HSQL[4].
- an hybrid query evaluation engine HYPATIA[1].
- an Active Policy model and language for specifying the QoS properties to be associated to service compositions modeled as workflows; and enforcement actions when they are not verified[5,6].

Main contracts:

- OPTIMACS (2009-2012) ANR-Programme ARPEGE. Service composition based framework for optimizing queries.
- CLEVER (2011-2013) STICAMSUD program U. de la República, Uruguay, UFRN Brazil, France. Environment virtual observatory on cloud.
- SWANS (2014-2016) CNRS STiC-AMSUD Program U. de la República, Uruguay, UFRN Brazil, France.
- Keystone - COST ICT Program (2013-2014): Semantic keyword-based search on structured data sources.
- AIWS (2012 - 2014), PEPS CNRS program. Discovering conversations among services by analyzing event logs.
- CAISES (2012 - 2015), European Union FP7, IRESES program (UK, France, Ukrania, China). Observation and industrial management on the cloud.

Key references:

- [1] V. Cuevas-Vicenttin, G. Vargas-Solar, C. Collet, Evaluating Hybrid Queries through Service Coordination in HYPATIA, In Proceedings of the 15th International Conference on Extending Database Technology (EDBT), Berlin, Germany, 2012
- [2] T. Delot, S. Ilarri, M. Thilliez, G. Vargas-Solar, S. Lecomte, Multi-scale query processing in vehicular networks, In Journal of Ambient Intelligence and Humanized Computing, Springer Verlag, ISSN 1868-75137, 2(3), 2011, pp. 213-226
- [3] Genoveva Vargas-Solar, Catarina Ferreira da Silva, Parisa Ghodous, José-Luis Zechinelli-Martini, Moving energy consumption control into the cloud by coordinating services, International Journal of Computing Applications, Special Issue. December 2013.
- [4] Carlos-Manuel Lopez-Enriquez, Genoveva Vargas-Solar, José-Luis Zechinelli-Martini, Christine Collet, Hybrid query generation, LANMR, 2012: pp. 117-128, .
- [5] Valeria de Castro, Martin A. Musicante, Umberto Souza da Costa, Plácido A. de Souza Neto, and Genoveva Vargas-Solar, Supporting Non-Functional Requirements in Services Software Development Process: An MDD Approach, In Proceedings of the 40th International Conference on

Current Trends in Theory and Practice of Computer Science, LNCS Springer Verlag, High Tatras, Slovakia, January, 2014.

[6] Javier A. Espinosa-Oviedo, Genoveva Vargas-Solar, José-Luis Zechinelli-Martini, Christine Collet. Policy driven services coordination for building social networks based applications. In Proc. of the 8th Int. Conference on Services Computing (SCC'11), Work-in-Progress Track, Washington, DC, USA, July 2011.

1.2.3 Reasoning on data semantics

This research is focused on combining reasoning and data management for efficiently querying and linking Web data through ontologies. Ontologies are very useful in many applications to express domain-specific knowledge over data that may be incomplete, uncertain or even inconsistent because coming from autonomous data sources distributed over the Web.

The proposed approach relies on recent complexity results showing that the expressive power of ontologies must be limited for making tractable reasoning on data enriched with ontologies. In particular, (several fragments of) the DL-Lite description logic we have been studied in the decentralized setting of P2P semantic networks. [1] designs a novel setting for robust module-based data management allowing to re-use a part of a reference ontology-based data system as an independent module while guaranteeing that it evolves safely w.r.t both the reference schema and its associated data. We are investigating how it applies to extract modules from the knowledge base on anatomy of My Corporis Fabrica. [2,3] proposed a novel model of trust based on alignments between taxonomies for guiding the query answering process in P2P semantic networks. Finally, [4] provides a novel method that is systematic and mathematically well-founded for discovering mappings between taxonomies of classes.

Main contracts:

- CONTINUUM (2008-2011), ANR - Programme Réseaux du futur et services. It addresses the problem of service continuity within the long-term vision of ambient intelligence.
- DATARING (2008-2011), ANR - Programme Réseaux du futur et services on the problem of P2P data sharing for online communities.
- PAGODA (2013-2016) ANR 12 JS02 007 01, Programme JCJC. The aim of this project is to develop practical algorithms for ontology-based data access.
- QUALINCA (2012-2016): ANR-2012-CORD-012. The aim of this project is to develop methods and algorithms for improving the quality and interoperability of large documentary catalogs.

Key references:

- [1] Robust Module-based Data Management. Francois Goasdou and Marie-Christine Rousset. IEEE Transactions on Knowledge and Data Engineering , Volume 25, Issue 3, March 2013, pages 648-661.
- [2] Alignment-based trust for resource finding in semantic P2P networks. Manuel Atencia, Jerome Euzenat, Giuseppe Pirro and Marie-Christine Rousset. Proceedings of ISWC 2011 (10th International Semantic Web Conference).
- [3] Trust in Networks of Ontologies and Alignments. Manuel Atencia, Mustafa Al Bakri and Marie-Christine Rousset. Knowledge and Information Systems, Volume 37, number 3, December 2013, 28 pages.
- [4] Discovery of Probabilistic Mappings between Taxonomies: Principles and Experiments Remi Tournaire, Jean-Marc Petit, Marie-Christine Rousset, and Alexandre Termier. Journal of Data Semantics (JoDS), Volume 15, pages 66-101.
- [5] Web Data Management, Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, Pierre Senellart, book published by Cambridge University Press.

1.2.4 Publications

The following table synthesizes the scientific publications of the group.



	2009	2010	2011	2012	2013	2014	Total
International peer reviewed journal [ACL]	3	3	1	4	4	1	16
International peer-reviewed conference proceedings [ACT]	9	16	4	18	17	2	66
Short communications [COM] and posters [AFF] in conferences and work-shops	2	0	0	0	2	0	4
Scientific books and chapter [OS]	1	2	1	0	0	0	4
National peer-reviewed conference proceedings [ACTN]	2	0	0	0	0	0	2
Book or Proceedings editing [DO]	1	0	0	0	0	0	1
Invited conferences [INV]	5	2	1	5	5	1	19
Doctoral Dissertations and Habilitations Theses [TH]	2	2	2	0	3	3	12
Other Publications [AP]	5	3	5	6	7	2	28
Total	30	28	14	33	38	9	152

The research has been done with PHD and post-doc. The number of doctoral students is rather stable over the last five years. We have 10PhD per year, all supervised by faculty members, resulting in 2 PhD's defense per year.

The numbers of publications in international conferences is significant and is stable for this past five years (14 per year) but the number of international publications per scientific member globally decreased. From the point view of research dissemination and recognition we have a very good impact with a lot of keynotes talks, tutorials and conferences programs participation among which are the well-known conferences in the domain of data management.



1.3 Visibility and attractivity

1.3.1 Rayonnement

Honor

- *Chevalier de l'ordre national du mérite*: : M.-C. Rousset (2011), Ch. Collet (2011)

Nomination

- *Membre Institut Universitaire de France (IUF)*: M.-C. Rousset (2011-2016)
- *Membre du Comité National du CNRS (nommée 2010-2012 dans l'ancienne section 7, élue dans la nouvelle section 6)*: M.-C. Rousset.
- *Conseil scientifique de la chaire d'excellence Smart Grids entre Grenoble INP et ERDF (2012-)*: Ch. Collet.
- *Comité de pilotage ANR, Mod les numériques (2010-2013)*: Ch. Collet.
- *VP adjointe recherche groupe Grenoble INP*: Ch. Collet (April 2007-2012).
- *Conseil scientifique INS2i - CNRS*: Ch. Collet (2010-).
- *Déléguée scientifique du LIG*: M.-C. Rousset.
- *Jury du prix de thèse Gilles Kahn (2010-2012) (prix décerné par Specif et patronné par l'Académie des Sciences)*: Ch. Collet.
- *Membre de la commission de spécialistes Réseau de technologies d'information CONACYT, Mexique*: G. Vargas-Solar.
- *Chargée de mission "International" auprès du Collège Doctoral de l'Université de Grenoble*: M.-C. Rousset.
- *Responsable scientifique et technique du Labex PERSYVAL-lab*: M.-C. Rousset.

Best Paper Awards

- Three prizes at SSSW 12 (Summer School on Ontology Engineering and the Semantic Web): Mustafa Al Bakri
- Best Paper Award track Embedded Software, conference DATE 2014, Sofiane Lagraa (HADAS/LIG and SLS/TIMA)

1.3.2 Contribution to the Scientific Community

Management of Scientific Organisations

- *President of the Extended Database Technology (EDBT) association*: Ch. Collet (2013 -)
- *Member of the Extended Database Technology(EDBT) association*: Ch. Collet (2004 - 2013) in charge of school organization program
- *Member of the IJCAI-09 advisory committee*: M.-C. Rousset (2009)
- *Deputy director of the French Mexican Laboratory in Informatics and Automatic Control (LAFMIA, UMI 3175) (2008 -)*: G. Vargas-Solar
- *Director of Labex PERSYVAL-lab*: M.-C. Rousset(2012-)
- *Membre du conseil d'administration du VLDB Endowment*: S. Amer-Yahia (2010-2013)
- *Membre du conseil exécutif de ACM SIGMOD*: S. Amer-Yahia (2010-2012)
- *Membre de IEEE, ACM (membre sénior) et ACM SIGMOD*: S. Amer-Yahia

Editorial Boards

- *PVLDB, publication of the Very Large Database Endowment*: Ch. Collet (2008-2011)
- *Computacion y sistemas*: G. Vargas-Solar, since 2002
- *ICDIM Journal special issue*: G. Vargas-Solar, since 2005
- *KER Journal special issue*: G. Vargas-Solar, since 2007
- *ActaPress Journal*: G. Vargas-Solar, since 2008
- *Interstices*: M.-C. Rousset
- *ACM Transactions on Internet Technology (TOIT)*: M.-C. Rousset, until 2005
- *AI Communications(AICOM)*: M.-C. Rousset
- *Communications of the ACM*: M.-C. Rousset, since 2009
- *advisory board of 21th International Joint Conference on Artificial Intelligence (IJCAI-09)*: M.-C. Rousset, since 2009
- *Rédacteur en chef de ACM TODS, Jan. 2011-2014*: S. Amer-Yahia
- *Rédacteur en chef de VLDB Journal, Sep. 2009-2015*: S. Amer-Yahia
- *Rédacteur en chef de Information Systems Journal (social search and recommendation and user influence in social media) depuis Juin 2010*: S. Amer-Yahia

Chair of Conferences and Workshops

- *Conference PC Chair*: SIGMOD Industrial 2015, BDA 2015, EDBT 2014, CIKM 2008, S. Amer-Yahia;
Conference SIIE'2012 (4th internationale conference on information systems and economical Intelligence). M.-C. Rousset
- *Chair of 5th International Conference on Information Systems and Economic Intelligence (SIIE 2012)*: M.-C. Rousset
- *Track Chair*: PVLDB 2013, SIGIR 2012, SIGMOD 2011, ICDE 2010, WWW 2010, ICDE 2008, S. Amer-Yahia
- *Industrial Chair*: EDBT/ICDT 2012, VLDB 2009, S. Amer-Yahia
- *Demonstration Chair*: EDBT 2011, S. Amer-Yahia
- *Workshop Chair*: PersDB 2009, XSym 2006, WebDB 2004, S. Amer-Yahia
- *Parallel Data Mining Workshop*, in conjunction with Conference SIAM Data Mining 2011, Mesa, USA: Alexandre Termier

- *Tutorial Chair*: SIGMOD 2009, S. Amer-Yahia ; Tutorial chair of International Joint Conference on Artificial Intelligence 2011 (IJCAI), Barcelona, M.-C. Rousset ;
- *Tutorial Chair of IJCAI-11*: M.-C. Rousset
- *Area Chair of IJCAI-13*: M.-C. Rousset

Organization of Conferences and Workshops

- *Extended Data Base Technology (EDBT) school 2009*, Ch. Collet, T. Delot and G. Vargas-Solar
- *Extended Data Base Technology (EDBT) school 2013*, S. Amer-Yahia, Ch. Collet and G. Vargas-Solar
- *Conférence Gestion de Données, principes, Technologies et Applications (BDA) 2014*, Ch. Bobineau et F. Jouanot
- *Tutorial Parallel Data Mining on Multicores*, International Joint Conference on Artificial Intelligence 2011 (IJCAI): Alexandre Termier

Program committee members

- *International Conference on Distributed Computing Systems (ICDCS)*: Ch. Collet (2007, 2009)
- *Extended Data Base Technologies (EDBT)*: Ch. Collet (2009, 2011, 2012 et 2013 (PHD workshop)), N. Ibrahim (2014), V. Leroy (2014)
- *International Conference on Data Engineering (ICDE)*: Ch. Collet (2009)
- *International Conference on Web Engineering (ICWE)*: Ch. Collet (2010)
- *International Conference on Model & Data Engineering (MEDI)*: Ch. Collet (2011)
- *International Conference on World Wide Web (WWW)*: Ch. Collet (2012 Web Engineering track), M.-C. Rousset (2012)
- *International Conference on Web Information Systems and Technologies (WEBIST)*: Ch. Collet (2012, 2013 et 2014), F. Jouanot (2012)
- *International Conference on Cloud Computing and Services Science (CLOSER)*: Ch. Collet (2012, 2013, 2014)
- *International Workshop on Information Management in Mobile Applications (IMMOA)*: Ch. Collet (2012 et 2013)
- *International Conference on Data Technologies and Applications (DATA)*: Ch. Collet (2012, 2013 et 2014)
- *International Conference on Information and Knowledge Management (CIKM)*: V. Leroy (2013)
- *International Conference on Data Mining (ICDM)*: A. Termier (2009 ... 2013)
- *SIAM International Conference on Data Mining (SDM)*: A. Termier (2009)
- *International workshop on ambient data integration (ADI)*: F. Jouanot (2009)
- *International Workshop on Data and Services Management in Mobile Environments (D2SME)*: Ch. Bobineau (2009)
- *International ACM Conference on Management of Emergent Digital EcoSystems (MEDES)*: G. Vargas-Solar (2009)
- *International Conference on the Applications of Digital Information and Web Technologies (ICADIWT)*: G. Vargas-Solar (2009)
- *International Conference on Parallel and Distributed Systems track on "Web Services" (ICPADS)*: G. Vargas-Solar (2009)
- *IEEE IFIP Conference on e-Business, e-Services, e-Society*: Ch. Bobineau (2009)
- *Database Engineering and Applications Symposium (IDEAS)*: Ch. Collet (2010), Ch. Bobineau (2011, 2012, 2014), F. Jouanot (2010)
- *International Conference on Tools with Artificial Intelligence (ICTAI)*: F. Jouanot (2010)
- *Journal d'Ingénierie des Systèmes d'Information (ISI)*: Ch. Bobineau (2009), F. Jouanot (2010)
- *Journal of ACM Transactions on Computer Systems*: Ch. Bobineau (2010)
- *Journal Technique et Science Informatiques (TSI)*: Ch. Bobineau (2014)
- *International Conference on Web Engineering (ICWE)*: F. Jouanot (2010)
- *International Symposium on Wearable Computers (ISWC)*: F. Jouanot (2011)
- *International Extended Semantic Web Conference (ESWC)*: F. Jouanot (2014)

- *Very Large Data Bases Conference (VLDB)*: Ch. Collet (2011), M.-C. Rousset (2013)
- *International Semantic Web Conference (ISWC)*: M.-C. Rousset (2011)
- *International Conference on Database Theory (ICDT)*: M.-C. Rousset (2011)
- *American Artificial Intelligence Conference (AAAI)*: M.-C. Rousset (2010)
- *European Conference on Artificial Intelligence (ECAI)*: M.-C. Rousset (2010, 2014)
- *International Joint Conference on Artificial Intelligence (IJCAI)*: M.-C. Rousset (2009, 2013), A. Termier (2013)
- *European Semantic Web Conference (ESWC)*: M.-C. Rousset (2009, 2014)
- *Reasoning Web Summer School*: M.-C. Rousset (2009)
- *Bases de Données Avancées (BDA)*: Ch. Collet (2010, 2012, 2013), Ch. Bobineau (2009, 2011), G. Vargas-Solar (2011)
- *Gestion des Données dans les Systèmes d'Information Pervasifs (GEDSIP)*: Ch. Bobineau (2009, 2010), G. Vargas-Solar (2009)
- *Conférence en Recherche d'Information et Applications (CORIA)*: V. Leroy (2014)
- *Conférence Extraction et Gestion de connaissances EGC'13*: A. Termier (2009-2013)
- *International Conference on Computational Systems (ICCS)*: M.-C. Rousset (2014)

Evaluation committee members

- *Vice-presidence comité d'évaluation scientifique "Big Data, décision, simulation, HPC"*, Ch. Collet (2014-).
- *ANR, Comité de pilotage Modèles numériques*, Ch. Collet, 2010-2013.
- *Member of the Specialists Council of the Delegation of Science and Technologies in Puebla, Mexico*., G. Vargas-Solar (2007-).
- *Member of the executive board of the National Network of Information and Communications Technologies, Mexico*., G. Vargas-Solar (2010-).
- *Invited member of the Comité de Sélection of INSA de Lyon*: Ch. Bobineau (2009, 2010, 2013).
- *Invited member of the Comité de Sélection of Université de Valenciennes et du Haut Cambrasis*: Ch. Bobineau (2009, 2010).
- *Invited member of the Comité de Sélection of Université d'Aix-Marseille 3*: Ch. Bobineau (2011).
- *Invited member of the Comité de Sélection of Université de Bordeaux*: Ch. Collet (2014).
- *Invited member of the Comité de Sélection of Université Paris Nord*: A. Termier (2010, 2013).
- *Invited member of the Comité de Sélection of Université de Rennes 1*: A. Termier (2012).
- *Member of the Comité de Sélection of Université Joseph Fourier*: A. Termier (2010).

1.3.3 Public Dissemination

Panels

- *Social Sites: Challenges and Opportunities*, Georgia Koutrika (moderator), Amr Al Abbadi (UCSB), Sihem Amer-Yahia, Laks Lakshmanan (UBC), Raghu Ramakrishnan (Yahoo!Labs), In PersDB, Seattle, Sep. 2011.
- *Does Social Media Make News Generation and Consumption Better?*, Sihem Amer-Yahia (moderator), Krishna Gummadi (MPI), Jimmy Lin (U. Maryland and Twitter), Gilad Lotan (SocialFlow), Marcus Mabry (NY Times and International Herald Tribune), Mor Naaman (Rutgers U.), Catherine Quayle (PBS Need to Know). In the International Conference on Web Logs and Social Media (ICWSM), July 2011.
- *Masses de Données, Big data et Data Management in Cloud*, Ch. Collet, M. Bouzeghoub, A. Laurent, D. Gros-Amblard. Conseil scientifique de l'INS2i; Feb. 2012.



Keynotes

- *Big Data and Smartgrids*, journée thématique centrée sur le traitement et la valorisation des données appliquées au domaine de l'énergie, Minalagic et Tenerdis. Avril 2014, Ch. Collet.
- *Défis du Big Social Data Management*. Les Fondamentales du CNRS. Nov. 2013, S. Amer-Yahia.
- *User Activity Analytics on the Social Web of News*. 18th International Conference on Management of Data, COMAD, Dec 2012, S. Amer-Yahia.
- *Crowd-Sourcing Literature Review in SUNFLOWER*. 1st International Workshop on Crowdsourcing Web Search (CrowdSearch) in conjunction with WWW, Apr. 2012, S. Amer-Yahia.
- *User and Topic Analytics of the Social Web of News*. 5th International Conference on Information Systems and Economic Intelligence (SIIE), Feb. 2012, S. Amer-Yahia.
- *I am structured: Cluster Me, Don't Just Rank me*. 2nd International Workshop on Business intelligence and the WEB (BEWEB) in conjunction with EDBT, Mar. 2011, S. Amer-Yahia.
- *Parallel Data Analysis, Dagstuhl Seminar, 2013, A. Termier*.
- Addressing Data Management on the Cloud: Tackling the Big Data Challenges at *CONIELECOMP 2013, Puebla, Mexique, G. Vargas-Solar*
- Addressing Data Management on the Cloud: Tackling the Big Data Challenges at *ITESM 2013, Academic Leaders Seminars, Puebla, Mexique, G. Vargas-Solar*
- Addressing Data Management on the Cloud: Tackling the Big Data Challenges, *Puebla, at Alberto Mendelzon Workshop 2013 Mexique, G. Vargas-Solar*
- Building data management services in clouds, at *Microsoft LATAM Faculty Summit 2012, Riviera Maya, Mexique, G. Vargas-Solar*
- Reasoning on Web Data Semantics, at *Collège de France, 2012, M.-C. Rousset*
- Cloud and Data Management: Research and Technical Challenges, *Conference UBIMOB 2010, Lyon, G. Vargas-Solar*
- Building the future of LATAM information and communication technologies, at *Microsoft LATAM Faculty Summit 2010, Guarujá, Brazil, G. Vargas-Solar*
- ECLOUDSS : Building E-government Clouds using Distributed Semantic Services. In *Microsoft Research Summit, Buenos Aires, Argentina, 2009, G. Vargas-Solar*
- Services in Clouds : a new perspective for accessing the digital world. In *Conference IETI, Universidad Popular Autonoma de Puebla, 2009, G. Vargas-Solar*
- Observing the environment for building today's and future information systems, *Universidad Juarez Autonoma de Tabasco (2009), G. Vargas-Solar*
- Semana de Juarez, Mujer serpiente: une historia alternativa sobre el origen de la civilización, *Universidad Juarez Autonoma de Tabasco (2009), G. Vargas-Solar*
- Semantic oriented data spaces. In *Invited tutorial at EDBT Summer School on Data and Resource Management in Ambient Computing, Sept. 2009, M.-C. Rousset*
- Polyglot persistence and multi-cloud data management solutions. In *Invited tutorial at EDBT Summer School on Data all around Big, Linked, Open, Sept. 2013, G. Vargas-Solar*



1.3.4 Principal International collaborations

China: The cooperation of HADAS with China started with the Beijing Institute of Automation (Chinese Academy of Science), more specifically the LIAMA French-Chinese laboratory, in 2007. We have obtained an ANR funding for the UBIQUEST project (Programme BLANC 2009-2012) where we have developed Case-Based Reasoning approach for optimizing distributed queries when (almost) no metadata are available. These techniques transparently exploit user-defined rule-based programs to combine network and data management.

The FP7 People CASES project extends this collaboration to the SouthEast University and the Aerospacial Center, both from Nanjing. The topics addressed in this context are service coordination data querying on the Cloud applied to energy consumption control in industrial processes. Within these projects, senior and young researchers have done internships in France and China.

Japan: There are strong ties between A. Termier and Takashi Washio of Osaka University. Since 2011, they have been collaborating with the PhD student C. Kamdem-Kengne and her supervisors N. Ibrahim and M.-C. Rousset on combining optimization methods and pattern mining algorithms in order to find new ways of rewriting execution traces.

Mexico: the group has a long tradition (20 years) in developing cooperation actions between the Mexican and French governments in TICS. The cooperation of HADAS with Mexico includes the most important private and public institutions of that country: three major public research centres CINVESTAV, CICESE, INAOE, private centres like LANIA; and important universities like UDLAP, UATx, ITESM. The main research topics in the cooperation are services based infrastructures for managing distributed data with non-functional properties, services based query processing and flexible data storage services. Collaboration on these topics has been formalized through projects (see below) and PhD students: A. Portila-Flores ¹, V. Cuevas ², J. Espinosa-Oviedo ³, Carlos-Manuel Lopez-Enriquez ⁴ and Juan Carlos Castrejon ⁵. Ch. Collet, and G. Vargas-Solar co-advised these students and with Ch. Bobineau they also advised master students of Mexican institutions. Some of them continued their education as PhD. students in France: Lourdes A. Martinez Medina ⁶.

Since 2008, G. Vargas-Solar is deputy director of the French Mexican Laboratory in Informatics and Automatic Control (LAFMIA, UMI 3175 <http://lafmia.imag.fr>) an international unit of the CNRS. The cooperation of HADAS with Mexico and the LAFMIA has lead to scientific results and to the education of graduate students through co-advising contracts and the organization of thematic schools.

Vietnam: We have developed since 15 years a strong collaboration with the Hanoi University of Science and Technology, including the International Research Institute MICA (UMI CNRS), and the National Universities in Danang and Ho-Chi-Minh City (4 doctors formed). This cooperation concerns adaptable distributed query optimization over stored data and data streams. Collaboration actions concern visits of senior scientists (2-3 times per year) and co-supervision of students.

Brazil: The collaboration with Brazil includes mainly the Universidade Federal Rio Grande do Norte, department DIMAP, equipe FORALL since 2007. This collaboration concerns service based data management on the cloud: semantic integration of data services through mashups, policy based service coordination, data processing on the cloud using map-reduce models, query languages based on service compositions. We privileged applications on meteorology data and energy distribution. These common research activities were funded by different organizations: Microsoft-LACCIR (e-CLOUDSS), the CNRS STICAMSUD program (CLEVER, SWANS) and by co-advising graduate students. In addition, postdoctoral and PhD and senior scientists internships in HADAS were funded by the CAPES and the CNRS.

Uruguay: The collaboration with Uruguay concerns the University of La Republica since 2007. This collaboration concerns the specification of mashup languages including policies for associating quality

¹Double diploma funded by the PROMEP Mexican program and the Foundation Jenkins.

²Funded by the project ANR - ARPEGE OPTIMACS.

³Funded through a CONACyT fellowship in the context of the ECOS-ANUIES project ORCHESTRA.

⁴Funded by the project ANR OPTIMACS, the CONACyT and the Jenkins Foundation

⁵Funded by an excellence fellowship of the doctoral school MSTII.

⁶Funded by the ANR project UBIQUEST.

of service attributes for retrieving and integrating data. The collaboration is done through commun projects (e-CLOUDSS, CLEVER, SWANS), co-advising of graduate students and invitations for lecturing seminars on topics related to the commun projects.


Spain: The collaboration with Spain concerns the region of Madrid and includes Universidad Rey Juan Carlos and Universidad Politecnica de Madrid. The collaboration concerns the specification of service oriented methodologies including non-functional properties. Collaboration actions concern visits of senior scientists, participation in PhD. viva, development of plug-ins for building a framework implementing the methodology π -SODM that we have proposed in the PhD of Placido Souza Neto.

Canada: HADAS group is participating in the LIA *Contrôle et Communication pour le Smart Grid* (CCSG) started in 2013. Canadian partners are *Institut National de la Recherche Scientifique* (INRS), McGill University and Hydro-Québec. Collaboration actions concern visits of senior and junior scientists.

1.4 Social, economical, and cultural impact

Results of our research have direct impact on applications dealing with huge amounts of data and resources in pervasive environments. They include traditional enterprise applications such as mining logs and traces, web applications but also "e-science" applications (in astronomy, biology, earth science, smartgrids, etc.). Environments we consider are wireless sensor networks (e.g. natural environment surveillance, industrial process monitoring), peer-to-peer data sharing, application deployment and maintenance for smart grids, transports, networks, smart cities.

1.4.1 Main Contracts and grants



Webcontent (RNTL, The semantic web framework-2006-2009), 10 partners (CEA LIST, EADS DCS, Thales Research & Technology, France Telecom R & D, ADRIA Développement, Soredab SAS, Exalead, New Phenix, Xyleme, INRIA-GEMO, INRA, INRIA-Mostrare, LIP6, PRISM, INRIA-InSitu, LIG, LIMSI-CNRS, GRIMM, EXMO, PSY-CO), (<http://www.webcontent-project.org/>), 83 300 €. Coordinator: CEA, Scientific lead in LIG: Ch. Collet and M.-C. Rousset. The WebContent project is creating a software platform to accommodate the tools necessary to efficiently exploit and extend the future of the Internet: the Semantic Web. The first targeted domain is the watch, a subpart of intelligence dedicated to warn the decider on the occurrence of an event or the evolution of a situation. It joins several Open Source tools to create the core of a Service Oriented Application and it defines the interface of several services that are available through several partners, either freely or through commercial licences. These services then exchange data in a formalized manner.

OPTIMACS : (ANR, program ARPEGE 2008-2011), 3 partners (LIG, LAMIH, LIRIS), (<http://optimacs.imag.fr>), 227 128 €. Coordinator: Grenoble INP-LIG, Scientific lead in LIG: G. Vargas-Solar.

OPTIMACS (SERVICE COMPOSITION BASED FRAMEWORK FOR OPTIMIZING QUERIES) combines hybrid query processing and services composition, addressing services composition and query processing including adaptive hybrid query optimization according to QoS criteria. OPTIMACS is an original research project that will lead to results with an important expected impact on "modern data and services intensive systems" deployed on networks of heterogeneous devices, the so called ecosystem or dataspace.

DATARING (ANR, Programme Réseaux du futur et services 2008-2011), 3 partners (<http://www.lina.univ-nantes.fr/projets/DataRing/>). 130 549 €. Coordinator: INRIA Nantes, Scientific lead in LIG: M.-C. Rousset.

The DataRing project addresses the problem of P2P data sharing for online communities, by offering a high-level network ring across distributed data source owners. Users may be in high numbers and interested in different kinds of collaboration and sharing their knowledge, ideas, experiences, etc. Data sources can be in high numbers, fairly autonomous, i.e. locally owned and controlled, and highly heterogeneous with different semantics and structures. What we need then

is new, decentralized data management techniques that scale up while addressing the autonomy, dynamic behavior and heterogeneity of both users and data sources.

CONTINUUM (ANR, Programme Réseaux du futur et services 2008-2011), 7 partners (I3S, LIG, SUEZ ENVIRONNEMENT, LYONNAISE DES EAUX, GEMALTO, LUDOTIC, MOBILEGOV), (<http://continuum.unice.fr>), 279 652 €. Coordinator: University of Nice, Scientific lead in LIG: F. Jouanot and M.-C. Rousset.

CONTINUUM (CONTINUE DE SERVICE EN INFORMATIQUE UBIQUITAIRE ET MOBILE) addresses the problem of service continuity within the long-term vision of ambient intelligence. A core problem is to achieve software adaptation to a variety of resources in dynamic and heterogeneous environments with an appropriate balance between system autonomy and human control. Three key scientific issues will be addressed: context management and awareness, semantic heterogeneity, and human control versus system autonomy. The professions related to water management is used as a business application domain.

UBIQUEST (ANR, Programme BLANC 2009-2012), 3 partners (LIG, CITI, LIAMA), 149 099 €. Coordinator: Grenoble INP-LIG, Ch. Collet. Scientific lead in LIG: Ch. Bobineau and Ch. Collet.

UBIQUEST (Ubiquitous Quest: a declarative approach for integrated network and data management in wireless multi-hop networks) aims at integrating network and data management in dynamic ad-hoc networks. this integration will be done by giving a distributed database view of the whole network. Each node stores network and application data in a local database. Messages between nodes are queries or answers. The objective of this integration is the rapid development and deployment of applications and network protocols.

Datalyse (Investissement d'Avenir May 2013 – November 2016), 219 950 €.

The aim of this project is to develop scalable algorithms for data mining and processing in collaboration with INRIA Saclay, LIFL, LIRMM and industrial partners: Eolas and B&D. The project defines 3 use cases, all of them with industrial impact. The first use case, network analysis applied to data centers datasets, aims to provide interactive traffic monitoring interfaces including traffic aggregation over time abnormal traffic identification. The second use case, digital marketing, applied to server and application logs, aims to provide customer-centric statistics and customer engagement analysis using sequence mining. The third use case, linked open data, aims to develop a platform that integrates open data on the city of Grenoble and makes it readily available for the development of various applications.

ALICIA (ANR, February 2014 – January 2017), 90 400 €.

The target of this project is the development of methods for information access and intelligent crowdsourcing in collaboration with Université Paris Sud, LTCI, Xerox, and UPS/IMT. In the context of information access (e.g. search or recommendation), building and maintaining user preference profiles helps applications satisfy diverse preferences. For intelligent crowdsourcing (e.g. data sourcing and micro-task completion), expertise profiles help better assign task to users. In both scenarios, the key challenges are that user preferences and expertise cannot be known in advance; and can rarely be explicitly declared by users in a reliable or stable way. Consequently, preferences and expertise need to be discovered over time via a learning approach. Our project's goal is the study of models and algorithms that rely on adaptive learning techniques to improve the effectiveness, performance, and scalability of user-centric applications.

PAGODA (ANR 12 JS02 007 01, Programme JCJC, January 2013 – December 2016, coordinated by Meghyn Bienvenu), funding for one year post-doctoral researcher.

The aim of this project is to develop practical algorithms for ontology-based data access (OBDA) in collaboration with LRI, LIRMM, and the LADAF (Laboratoire d'anatomie de Grenoble). This project is centered on two challenges: *(i)* Scalability: in contrast with relational database management systems that benefit from decades of research on querying algorithms and optimizations, ontology-based data access is a young area of study, and despite important recent advances, including the identification of interesting tractable ontology languages, much work remains to be done in designing scalable OBDA query answering algorithms. *(ii)* Handling data inconsistencies: In real-world applications involving large amounts of data or multiple data sources, it is very likely that the data will be inconsistent with the ontology, rendering standard querying algorithms useless (as

everything is entailed from a contradiction). Appropriate mechanisms for dealing with inconsistent data are thus crucial to the successful use of OBDA in practice, yet have been little explored thus far.

SoGrid (2013-2017) - ADEME Le réseau électrique de demain, 270 000 €. Scientific lead in LIG: Ch. Collet. Funding for one year post-doctoral researcher and one PhD.

SoGrid aims at confirming the path opened by ERDF in the technological revolution of the Smart Grid. The goal is to develop a global communication chain linking all components as the basis of future Smart Grids. By deploying intelligence all along this communication chain, SoGrid will allow (i) real-time supervision and control of the electrical grid; (ii) integration of decentralized renewable energy production; (iii) anticipation and support of new uses of electricity, in particular electric vehicles; (iv) the possibility to ensure that at every moment the balance between production and consumption, particularly during peak consumption; and (v) Control of consumption by the end-user and better quality of service. HADAS group is developing new optimized distributed and adaptable event stream management techniques taking into account the specificities of the Smart Grids.

CLEVER (2012-2013) CLEVER CLOUD-BASED LATIN-AMERICAN ENVIRONMENTAL VIRTUAL OBSERVATORY. Scientific lead in LIG: Genoveva Vargas-Solar.

The project aims at providing the underlying services that will enable the VO to personalize and manage mashed up services. The result will be a platform where climate reports coming from different providers in LATAM will be mashed up. Resulting mashups will be exported as VO tools for eventually building other mashups.

QUALINCA (2012-2015). 80 600 € QUALINCA is a ANR Contint funded research project looking at developing mechanisms allowing to quantify the quality level of a bibliographical knowledge base, to improve the afore mentioned quality level, to maintain the quality when updating the knowledge base and to exploit the knowledge bases taking into account their quality levels. This project aims to develop mechanisms to: (i) describe the quality of an existing document database; (ii) maintain a given level of quality by controlling updates on such databases; (iii) improve the quality of a database; (iv) exploit these databases according to their level of quality.

SocTrace (2011-2015) FUI-Minalogic, OSEO. 374 700€. Partners: INRIA, LIG, TIMA, STMicroelectronics, Magilem, probayes. Coordinator: STMicroelectronics Scientific lead in LIG: Alexandre Termier.

The SoC-Trace project aims to develop a set of methods and tools based on traces of execution produced by multi-core embedded applications. It will allow developers to optimize and debug these applications more efficiently. Such methods and tools should become a building block for the design of embedded software, in response to the growing needs of analysis and debugging required by the industry. The technological barriers consist of a scaling problem (millions of events stored on gigabytes) and a trace understanding problem related to applications whose complexity is increasing. The project addresses the problem of controlling the volume of traces and of developing new analysis techniques. SocTrace is composed of academic partners with related themes, and several industry partners including STMicroelectronics.

1.4.2 Research Networks (European, National, Regional, Local)

E-CLOUDSS: (BUILDING E-GOVERNMENT CLOUDS USING DISTRIBUTED SEMANTIC SERVICES, Microsoft, 2007-2011, LACCIR, <http://e-cloudss.imag.fr>), 5 partners (CNRS LIG-LAFMIA, Fundacion Universidad de las Américas, Puebla, Mexique, Universidad de la Republica de Uruguay, Uruguay, Universidade Federal do Rio Grande do Norte), 50,000 USD. Coordinator: J.L. Zechinelli Martini, LAFMIA, Scientific lead in LIG: G. Vargas-Solar). The objective of E-CLOUDSS is to propose an infrastructure for mashing up reliable semantic services for building e-government clouds. Mashups represent a new wave for building Web applications. E-CLOUDSS addresses the management (definition and enforcing at execution time) of non functional properties associated to services coordination for building reliable mashups. Effective ways to perform virtual executions is one of the main subjects of study of E-CLOUDSS.



WebIntelligence: (Cluster Régional "Informatique, Signal, Logiciels embarqués" - 2006-2009). The project aims at organizing research on web intelligence in Rhone-Alpes.
suite WebIntelligence (ARC 6 et ARC 7) : Cloud Computing Research Group (<http://cloud.liris.cnrs.fr/wiki/doku.php>). Our participation in the regional cluster addresses topics related to cloud computing, particularly big data collections integration and management on multi-cloud environments guided by service level agreements. We work with the LIRIS lab and the Management School of University Lyon 3. Our common research results are applied to smart energy, intelligent transport and political strategies cases. We actively, organize seminars and collective actions willing to encourage research collaboration synergy among the actors of the region working on cloud computing and big data management.

ORCHESTRA: (ORCHESTRATION TRANSACTIONNELLE DE SERVICES, Program: ECOS-ANUIES 2007-2011), 3 partners (Grenoble INP, Universidad Autonoma de Tlaxcala, Fundacion UNiversidad de las Américas, Puebla, Mexique). Coordinator: Ch. Collet. Missions for Professors (Ch. Collet in 2007 and 2008, and G. Vargas in 2009) and PhD students. The objective of ORCHESTRA is to propose an infrastructure pour building transactional, secure and evolutive service-based applications. The key elements of the project are: (i) the definition of a framework (general solution) of technical services for managing the security, transactional properties and evolution of business services ; and (ii) implementation of the framework an its validation in the development of service-based applications: production chains.

CASES (2012-2015) - European Union FP7, PEOPLE program, UK, France, Ukrania, China. Customized Advisory Services for Energy-efficient Manufacturing Systems. Scientific lead in LIG: Genoveva Vargas-Solar. 156 000 €.

The project aims at teaming up transcontinental researchers in the areas of sustainable manufacturing and information technologies to enrich the knowledge base and achieve research synergies to develop smart design and manufacturing services in terms of energy efficiency. The project integrates the complementary expertise of the European, Chinese and Ukrainian teams to devise ICT-based smart services and standards to address the multi-faceted requirements of global eco-design and sustainable manufacturing planning.

1.4.3 Internal Funding



RED-SHINE: (RELIABLY AND SEMANTICALLY INTEGRATING WEB INFORMATION BY MASHING UP DATA SERVICES, BQR Grenoble INP, 2009). 2 partners (LIG, LAFMIA-UMI 3175) (<http://lafmia.weebly.com/>), 20 000€ - one PhD grant and 4 months for inviting professors. Coordinator: Grenoble INP-LIG, Scientific lead in LIG: G. Vargas-Solar).

The objective of RED-SHINE is to propose an infrastructure for mashing up services using semantics and thereby integrating information from the Web. RED-SHINE will redefine and extend OQLiST for declaratively defining reliable semantic mashups. RED-SHINE addresses the management (definition and enforcing at execution time) of non functional properties (NF-P) associated to services' coordination for building reliable mashups. The objective of our work will be to propose a language for orthogonally expressing NF-P and ensuring strategies, and to specify execution strategies for adding NF-P to mashups.

DAMOCLES: (MSTIQ project, 2009). 2 partners (LIG, TIMA), 15 000 € - one year postdoc. Coordinator: Grenoble INP-LIG, Scientific lead in LIG: A. Termier).

DAMOCLES (DAta Mining for On Chip Low Energy Systems) aims at developing data mining algorithms for analyzing memory accesses in System-on-Chip processors, in order to optimise data placement and thus reduce energy consumption.

Smart Energy: Grenoble INP (2012-2014), Participants: LIG, G-SCOP, G2ELab, GIPSA-Lab.

This projects aims at federating the scientific communities from Grenoble INP supported laboratories around the development of Smart Grid technologies. The idea is to identify common interests among researchers to propose new research projects.

WalT: Grenoble INP and University Joseph Fourier, Programme AGIR (2013-2015) on Wireless Testbed.

This project, proposed by HADAS and DRAKKAR research groups, aims at developing an easy

configurable testbed composed of embedded computing devices (Raspberry Pi), sensors and network equipments (e.g. commutators, wireless communications). It will be exploited to test new networking protocols or distributed database techniques developed in both research groups.

Big join: Grenoble INP and University Joseph Fourier, Programme AGIR (2013-2015) on Modèles et algorithmes pour les jointures de Big Data sur Map-Reduce, Scientific lead in LIG: S. Amer Yahia).

1.5 Team Organization and life

Full group meeting are organized once every two weeks for discussions about research works, progress concerning research contracts and so on. These meetings are an opportunity for students or permanents to present their works, their difficulties, to exercise presentation of their papers and/or to present interesting research papers they have found. In parallel and with the same frequency, we organize meeting for permanent staff to discuss management aspects.

Every PhD student has to participate to at least one summer school and to assist to at least one major conference during his PhD. They also have to prepare a poster presenting their work that will be presented in very short sessions (two-minutes madness).

Each researcher is responsible of his contracts and manages the associated budget as he/she wants. Every one is fully autonomous in his research work and operation. But major decisions concerning the group are still taken collegiately. Some responsibilities are distributed among permanent people:

- management of offices and keys: N. Ibrahim
- management of computing equipment: Ch. Bobineau
- Website administration: F. Jouanot.

Finally, three full group seminars off-campus have been organized to discuss major points within the group since 2009:

- January 14th and 15th 2009 at the "Pic de la Belle Etoile" hotel in Pinsot.
- April 8th and 9th 2011 at the "Pic de la Belle Etoile" hotel in Pinsot.
- October 22nd and 23rd 2012 at the "Centaure" center in Réaumont.

The last evaluation report on our group read as follows:

- Points à améliorer et risques :
Si l'activité de publication en revue est en hausse, les efforts doivent encore se poursuivre pour amener les différentes thématiques à un niveau comparable. Le nombre de HDR est faible comparativement au nombre de doctorants. Les problématiques abordées dans l'équipe nécessitent une synergie avec les systèmes répartis, le parallélisme et les réseaux. Même si des collaborations sont déjà établies sur ces thématiques avec les équipes du LIG ou à l'extérieur, cet effort doit être poursuivi si l'on veut véritablement contribuer au plus niveau.
- Recommandations :
L'équipe doit augmenter son nombre de HDR pour gérer au mieux ses doctorants et les contrats en cours. La mise à disposition d'un ingénieur permettrait à l'équipe d'aller plus loin dans la réalisation et la valorisation de ses productions logicielles.

Two remarks are in order concerning this report.

- First, the number of journal publications continued to increase with a better repartition between the research themes of the group. The collaboration with groups working on networks, distributed systems and parallel algorithms were strengthened. Common publications or projects attest this.
- Second, two HDR have been defended during this period in 2013 and 2014. The arrival of S. Amer-Yahia (DR CNRS) also helped in increasing the potential of coaching. Etienne Dublé, research engineer part-time joint the group in 2011. He was a great help in developing prototypes. However, this is not enough to have a real valorisation strategy.

1.6 Training through research, educational involvement

Thesis: 10 thesis have been defended; On average, a thesis is done in four years. Also, two HDR have been defended.

- *Pattern mining rock: more faster, better.* Alexandre Termier, HDR UJF, July 2013.
- *Efficient, continuous and reliable Data Management by Coordinating Services.* Genoveva Vargas-Solar, HDR Grenoble INP, May 2014.

PhD student supervision: regular meetings, presentations of papers, "workshop" on talks / papers, participation to schools, poster.

Future: The professional perspectives of students are diverse: 5 former PHD are in industry (researcher or engineer R & D) and 5 have post-doc or academic positions.

Educational involvement

Supervision of Educational Programs:

- M.-C. Rousset: Co-director for the Master of Sciences in Informatics (2009-2013)
- N. Ibrahim: Co-director for the Master of Sciences in Informatics (2014 -)
- Ch. Bobineau:
 - Correspondant Vietnam pour Grenoble INP - Ensimag.
 - Chargé de projet PFIEV (Programme de Formation d'Ingénieurs d'Excellence au Vietnam) pour Grenoble INP

Administrative involvement

Member of the Commission des Utilisateurs des Moyens Informatiques (CUMI-LIG): Ch. Bobineau

Member of the Commission Web (Web-LIG): F. Jouanot

Correspondant visio-conférence sur poste pour le LIG: Ch. Bobineau

1.7 Strategy and Research Project

As said in introduction, the HADAS group evolved all along the current period with the arrival of three persons. Sihem Amer-Yahia Research Director (Dec 2011) and Vincent Leroy Associate Professor UJF / Research scientist (Sept 2012) brought new areas of research around large scale data processing on the social Web. As a consequence, there has been a thematic evolution in the HADAS team, resulting in two research projects, each of them been part of a separate group. The following sections present the Strategy and Research Project of these two groups, HADAS and SLIDE, respectively.

HADAS

HADAS now stands for Heterogeneous and Adaptive DATA management Systems. Its research project is guided by new challenges introduced by the continuous production of huge, distributed and heterogeneous data that require data technologies to support several activities such as capturing, integrating, searching/querying, filtering, indexing, recording, preserving, annotating, etc ... We propose to contribute to the development of new largely distributed, scalable, adaptive and intelligent data and knowledge management infrastructures that include these technologies. During the next period, we will focus on:

- **Management of massive datasets particularly focusing on:**
 - Adaptive and distributed storage and cache for storing large heterogeneous datasets.
 - Indexing data on the fly to facilitate efficient data manipulation.
 - Economy and energy oriented integration of big datasets management: economic cost model.
 - Quality-based continuous data/event stream processing and composition.

- **Adaptive querying systems:**

- Declarative hybrid languages for expressing data (streams) processing.
- Learning-based distributed query optimization for efficient (continuous) query evaluation with scarce metadata.
- Query operators for on-the-fly data reorganization facilitating future data manipulations.
- Service Level Agreement guided optimization of continuous and mobile queries.

Of course these data technologies have to be largely distributed and deployed over different types of architectures (grids, peer-to-peer networks, sensor networks, cloud infrastructure). We will adopt a service-based approach and develop new data models, algorithms and services that will fulfill properties such as efficiency, adaptivity, reliability, robustness, security, confidentiality, and privacy. This vision is nowadays well accepted as we have to consider large and heterogeneous data sets, huge numbers of connected devices with data management capabilities and increasing numbers of users/applications. In such a vision, well-adapted for the internet of objects, security of data and service is a big issue, especially data provenance that could be managed using physical tagging systems.

Sustainable mobility and urban systems like smart cities, energy, clean, safe and efficient technologies like Smart Grids, smart energy, clean technologies and data markets for extracting business value from data, are examples of applications that call for an intelligent, adaptive, efficient and scalable data and knowledge management infrastructure. The smart grids domain we choose to explore is very promising as the management of data (where to put the data, what to do with it, which data to collect, integrate, summarize, how to access it efficiently, ...) is the foundation for developing intelligent metering systems and adaptive supervisory control able to handle huge amount of events and alerts. We will also test the reliability and robustness of our data technologies when collecting healthcare data.

Our agenda falls within the scientific research directions on Information and Communication Technologies given by the H2020 program. We identified two of them for the group: (i) Advanced Cloud Infrastructures and Services, and (ii) Big Data Innovation and take-up and Big Data. It also concerns the societal challenges: Secure, clean and efficient energy.

Members:	Name	First name	Function	Institution	Arrival date
	Adiba	Michel	External collaborator	UJF	Oct 2010
	Bobineau	Christophe	Associate Professor	Grenoble INP	Sep 2003
	Collet	Christine	Full Professor	Grenoble INP	Sep 1999
	Dubl�	Etienne	ITA	CNRS	May 2011
	Vargas-solar	Genoveva	Research Scientist	CNRS	Jan 2002

SLIDE

SLIDE stands for ScaLable Information Discovery and Exploitation. Data today is heterogeneous, ubiquitous, and overwhelming. It is a great challenge to go from data to information, then to knowledge and actionable intelligence. SLIDE's research agenda addresses this challenge and studies problems such as 1) how to efficiently organize and mine information from data; 2) how to better involve humans in data procurement and application evaluation; 3) how to efficiently discover knowledge from data in multiple sources; 4) how to reason over data semantics to enable new access methods. Research within the team is organized into three axes:

- **Data acquisition and enrichment:** This axis addresses challenges behind how to optimize data acquisition via crowd data sourcing, how to subsequently prepare raw data for further processing by cleaning it and enriching it with techniques such as Web data linkage. The outcome is a reusable framework for acquiring and pre-processing data in a scalable fashion by addressing the following three sub-axes:
 - Big data preparation.
 - Web data linkage.
 - Crowd data sourcing.
- **Large-scale data mining:** This axes addresses the question of extracting value from pre-processed datasets in order to provide recommendations and advanced search. It covers the development of scalable algorithms (parallel and distributed) on state-of-the-art manycore architectures to mine

and process large data volumes at scale. The three sub-axes are:

- Advanced pattern mining.
- Distributed join algorithms.
- Social media and health analytics.

- **Information exploration:** This axis aims to design and implement novel ways of navigating in the space of information extracted from raw datasets. It covers the development of logic-based approaches for ontology-based data access and of optimization-based approaches for exploring the large space of patterns discovered via mining. It also develops principled approaches that include crowdsourced validation for the evaluation of these newly developed exploration methods.
 - Ontology-based data access.
 - Interactive pattern exploration.
 - Principled exploration evaluation.



SLIDE aims at leveraging data availability today to conduct data-driven research and develop algorithms and infrastructures for large-scale analytics, data linkage and ontology-based data access, crowd data sourcing and crowdsourced application evaluation. The first step is data acquisition from different domains including user demographics, behavioral and opinion data on the social Web, data from the semantic Web, user data on health and well-being, user and program traces and data center monitoring. Research within SLIDE targets two kinds of end-users: experts who are interested in exploring and extracting value from large datasets and novice users interested in search and recommendation. The data analytics axis (D for Discovery in SLIDE) covers advanced pattern mining including generic mining algorithms, mining on parallel infrastructures such as MapReduce and many core processors and social media analytics. The result of this axis are models and algorithms that combine data mining with multi-dimensional indexing to discover a variety of information of interest from raw data in a scalable manner. On the other hand, expert-facing applications enable new data exploration approaches such as interactive mining. This same axis encompasses a data preparation framework composed of an algebra and algorithms for sanitizing and transforming large data volumes into ready-to-be exploited data. It also covers a crowd data sourcing framework that optimizes online data acquisition. Finally, extensions to Datalog are being developed with the goal to express and infer linkage between heterogeneous data sources. The information exploitation axis (E for Exploitation in SLIDE) covers the development of distributed data access methods that combine data partitioning and data placement techniques with traditional join algorithms to design faster data processing on distributed parallel infrastructures. It also covers ontology-based data access algorithms that allow analysts to explore large data volumes through high-level concepts. User-facing applications developed within SLIDE rely on novel search and recommendation algorithms ranging from searching for relevant and diverse results, to defining and implementing novel semantics for recommendation that include social networks and different user similarity functions. A large number of applications are evaluated using traditional information retrieval and learning techniques but also using crowdsourcing. One research axis to be developed in SLIDE is the design and implementation of models and algorithms for crowd data sourcing and crowd-sourced application evaluation. More particularly, worker-to-task assignment algorithms that account for human factors such as worker skill, are being developed with the team.

Members:

Name	First name	Function	Institution	Arrival date
Amer Yahia	Sihem	Research Director	CNRS	Dec 2011
Dubl�	Etienne	ITA	CNRS	May 2011
Ibrahim	Noha	Associate Professor	Grenoble INP	Sep 2010
Jouanot	Fabrice	Associate Professor	UJF	Sep 2003
Leroy	Vincent	Associate Professor	UJF	Sep 2012
Rousset	Marie-christine	Full Professor	UJF	Sep 2005
Termier	Alexandre	Associate Professor	UJF	Sep 2007

1.8 Self assessment

This section introduces the self assessment of the two groups HADAS and SLIDE.

HADAS

Our main strength is our experience for designing and developing components at the heart of data management systems. HADAS has been one of the pioneer groups proposing the unbundling of database systems and considering future database system as a large-scale semantic-based infrastructure for data and resources management. It raised many fundamental questions which have been the basis of fundamental research published in top journals and conferences. HADAS has been very attractive as shown with the recruiting of several experts this past years.

A strong positive point of our activity is the numerous and fruitful collaborations that we have. They are local with our implication in the Smart Energy project and the AMIQUAL-4-HOME platform. At the national level, we participated to several ANR projects and we are involved in the ADEME SoGrid project. SoGrid is an industrial project with ambitions world led by ERDF and STMicroelectronics in a consortium of 10 partners (with a budget of 27M€). We have been recently very active for the creation of the future GDR on Big Data and Data Sciences. At the international level we have numerous collaborations. We have strong relationships for years with Mexico (G. Vargas is the deputy director of the LAFMIA lab) that had lead to scientific results (publications and prototypes) and to the education of graduate students through co-advising contracts. We also have collaborations with Vietnam. During this period we developed exchanges with Brazil, Uruguay (research network), Spain and and China (CASE network).

From the point of view of contracts, the group is very active. We have an average of 120K€ as inputs per year. This means around 40K€ per (equivalent) permanent researcher per year. Contracts allow us to hire PhD students, post-doc and engineers. During the period we also got 2 grants from the french research Ministry, including an excellence grant from the UJF.

Managing and developing collaborations, PhD supervision, contracts is time consuming and is a potential cause of difficulty. With the emergence of the SLIDE group, the main problem of the HADAS group will still remain, namely a lack of critical mass in terms of permanent staff even if we plan to hire two post-doc for the two incoming years. These problems will probably hinder the activities of the group in terms of support for contractual work, and to a lesser extent in the supervision of doctoral students.

The research themes of the group concerns two mains domains of the 4th revolution of computer sciences: big data, and cloud computing. The Big data phenomenon is a real opportunity to validate our component-based data systems vision as in this domain, data technologies have to (i) consider large and heterogeneous data sets, huge numbers of connected devices with data management capabilities and several numbers of users/applications, and (ii) face the challenges of scalability, heterogeneity, distribution, efficiency, quality, security and adaptivity. All of these aspects will be considered in our researches.

Also the smart grids domain we choose to explore is very promising as the management of data (where to put the data, what to do with it, which data to collect, integrate, summarize, how to access it efficiently, ...) is the foundation for developing intelligent metering systems and adaptive supervisory control able to handle huge amount of events and alerts.

SLIDE

Strengths. SLIDE is a unique group that combines efficient large-scale data processing techniques with pattern mining algorithms to extract value from large amounts of data. Its members have expertise at the intersection of knowledge representation and reasoning, data mining, and data management. Members of the SLIDE team are regularly invited to give keynotes in top conferences and have multiple publications in top rated conferences and journals in their fields.

SLIDE is involved in 3 ANR projects, one FUI and one investissement d'avenir. SLIDE members have collaborations with 6 teams within LIG, all of which are represented by co-advised students (SLIDE currently has 12 PhD students) via common projects, and with other laboratories in Grenoble (TIMA, LJK and Laboratoire d'Anatomie de Grenoble).

SLIDE members have numerous national and international collaborations. International collaborations are with the University of British Columbia on large-scale information discovery and exploration, with Yahoo! Labs, Barcelona on open-sourcing a stream mining platform, with the University of Washington in Tacoma and the University of Texas in Arlington on crowd data sourcing, with Osaka University on pattern mining optimization for trace analysis, and finally with Tsukuba University on building an academic crowdsourcing network worldwide.

Weaknesses. Data-driven research requires to seamlessly combine different expertise domains that are not all carried by SLIDE members, in particular statistics, visualization and machine learning. However, within the LIG environment and more generally, in Grenoble, members of the SLIDE team have been naturally collaborating with many researchers and practitioners in these areas.

Opportunities. SLIDE members have great expertise in working with large data amounts and in collaborating with industry. Two of its members were employed by Yahoo! Labs before joining LIG and will continue to use their experience in establishing itself as one of the few groups in France with this expertise and an ability to work with large amounts of real datasets. SLIDE has also several members involved in a long-lasting collaboration with STMicroelectronics, a world-class semi-conductor designer located in Grenoble. This collaboration has been the source of major advances in the area of data mining for execution trace analysis. Finally, SLIDE members have shown a recent interest in medicine, health and wellbeing. Current projects include a collaboration with the University hospital in Grenoble. These areas open new opportunities in knowledge representation and linked data and in mining social media.

Threats. SLIDE is a recently created team with ambitious members who may get involved in too many projects. One of the most important aspects in the team is to maintain international collaborations while affirming itself in France. This may cause over-committment.

The main threat for data-driven research is the unavailability of data or the unwillingness of industrial partners to share their datasets. That is why SLIDE members are establishing long-lasting collaborations with multiple industrial partners.

Lastly, another threat is the need for a fair amount of engineering to prepare and pre-process raw datasets and also build demonstrations to “keep industrial partners interested”. That is why SLIDE members make sure to hire research engineers in their projects in order to make sure research time is not spent in engineering.