IN PARTNERSHIP WITH:
CNRS

Université des sciences et techniques du Languedoc (Montpellier 2)

Activity Report 2015

Project-Team ZENITH

Scientific Data Management

IN COLLABORATION WITH: Laboratoire d'informatique, de robotique et de microélectronique de Montpellier (LIRMM)

RESEARCH CENTER
Sophia Antipolis - Méditerranée

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team ZENITH

Creation of the Team: 2011 January 01, updated into Project-Team: 2012 January 01

Keywords:

Computer Science and Digital Science:

1. - Architectures, systems and networks
1.1. - Architectures
1.1.6. - Cloud
1.1.7. - Peer to peer
3. - Data and knowledge
3.3. - Data and knowledge analysis
3.3.2. - Data mining
3.3.3. - Big data analysis
3.5. - Social networks
3.5.2. - Recommendation systems
4. - Security and privacy
4.8. - Privacy-enhancing technologies
5. - Interaction, multimedia and robotics
5.4. - Computer vision
5.4.3. - Content retrieval

Other Research Topics and Application Domains:

1. - Life sciences
1.1. - Biology
1.1.9. - Bioinformatics
1.2. - Ecology
1.2.1. - Biodiversity
6. - IT and telecom
6.5. - Information systems

1. Members

Research Scientists

Reza Akbarinia [Inria, Researcher]
Alexis Joly [Inria, Researcher]
Florent Masseglia [Inria, Researcher, HdR]
Didier Parigot [Inria, Researcher, HdR]
Patrick Valduriez [Team leader, Inria, Senior Researcher, HdR]

Faculty Members

Sarah Cohen-Boulakia [University Paris Sud (on leave at Inria)]
Esther Pacitti [Associate Team Leader, Univ. Montpellier 2, Professor, HdR]
Dennis Shasha [Inria international chair, NYU, Professor, HdR]

Engineers

Julien Champ [Inria and Inra, Pl@ntNet and ARCAD projects]

Julien Diener [Inria, Xdata project unti June 2015]
Dimitri Dupuis [Inria, Scifloware project]
Boyan Kolev [Inria, FP7 CoherentPaaS project]
Pierre Larmande [IRD, collaborator]
Miguel Liroz-Gistau [Inria, Datascale project]
Jean-Christophe Lombardo [Inria]
Oleksandra Levchenko [Inria, FP7 CoherentPaaS project, since Sept 2015]
Antoine Affouard [Inria, PIA Floris'tic project]

PhD Students

Carlyna Bondiombouy [Congo fellowship]
Valentin Leveau [INA CIFRE]
Ji Liu [Inria-MSR]
Saber Salah [Inria, Hemera action]
David Fernandez [Univ. Nice, Triton e-lab]
Daniel Gaspar [LNCC, Rio de Janeiro, since Sept 2015]
Sakina Maboubi [Univ. Montpellier, Averroes, since Oct 2015]
Vitorl Silva [UFRJ, Rio de Janeiro, since Sept 2015]
Djamel-Edine Yagoubi [Univ. Montpellier, Averroes]
Mehdi Zitouni [University of Tunis, Tunisia]

Post-Doctoral Fellows

Maximilien Servajean [Inria, FP7 CoherentPaaS project]
Mohamed Reda Bouadjenek [Inria, Bigdatanet associated team, until June 2015]
Rim Moussa [Inria, Datascale project, until June 2015]

2. Overall Objectives

2.1. Overall Objectives

Modern science such as agronomy, bio-informatics, astronomy and environmental science must deal with overwhelming amounts of experimental data produced through empirical observation and simulation. Such data must be processed (cleaned, transformed, analyzed) in all kinds of ways in order to draw new conclusions, prove scientific theories and produce knowledge. However, constant progress in scientific observational instruments (e.g. satellites, sensors, large hadron collider) and simulation tools (that foster in silico experimentation, as opposed to traditional in situ or in vivo experimentation) creates a huge data overload. For example, climate modeling data are growing so fast that they will lead to collections of hundreds of exabytes (10^{18} bytes) expected by 2020.

Scientific data is also very complex, in particular because of heterogeneous methods used for producing data, the uncertainty of captured data, the inherently multi-scale nature (spatial scale, temporal scale) of many sciences and the growing use of imaging (e.g. satellite images), resulting in data with hundreds of attributes, dimensions or descriptors. Processing and analyzing such massive sets of complex scientific data is therefore a major challenge since solutions must combine new data management techniques with large-scale parallelism in cluster, grid or cloud environments.

Furthermore, modern science research is a highly collaborative process, involving scientists from different disciplines (e.g. biologists, soil scientists, and geologists working on an environmental project), in some cases from different organizations distributed over different countries. Each discipline or organization tends to produce and manage its own data, in specific formats, with its own processes. Thus, integrating distributed data and processes gets difficult as the amounts of heterogeneous data grow.

Despite their variety, we can identify common features of scientific data: big data; manipulated through complex, distributed workflows; typically complex, e.g. multidimensional or graph-based; with uncertainty in the data values, e.g., to reflect data capture or observation; important metadata about experiments and their provenance; and mostly append-only (with rare updates).

Generic data management solutions (e.g. relational DBMS) which have proved effective in many application domains (e.g. business transactions) are not efficient for dealing with scientific data, thereby forcing scientists to build ad-hoc solutions which are labor-intensive and cannot scale. In particular, relational DBMSs have been lately criticized for their "one size fits all" approach. Although they have been able to integrate support for all kinds of data (e.g., multimedia objects, XML documents and new functions), this has resulted in a loss of performance and flexibility for applications with specific requirements because they provide both "too much" and "too little". Therefore, it has been argued that more specialized DBMS engines are needed. For instance, column-oriented DBMSs, which store column data together rather than rows in traditional row-oriented relational DBMSs, have been shown to perform more than an order of magnitude better on decision-support workloads. The "one size does not fit all" counter-argument generally applies to cloud data management as well. Cloud data can be very large, unstructured (e.g. text-based) or semi-structured, and typically append-only (with rare updates). Though cloud users and application developers may be in high numbers, DBMS experts wouldn't. Therefore, current cloud data management solutions have traded consistency for scalability, simplicity and flexibility. As alternative to relational DBMS (which use the standard SQL language), these solutions have been quoted as Not Only SQL (NoSQL) by the database research community.

The three main challenges of scientific data management can be summarized by: (1) scale (big data, big applications); (2) complexity (uncertain, multi-scale data with lots of dimensions), (3) heterogeneity (in particular, data semantics heterogeneity). The overall goal of Zenith is to address these challenges, by proposing innovative solutions with significant advantages in terms of scalability, functionality, ease of use, and performance. To produce generic results, these solutions are in terms of architectures, models and algorithms that can be implemented in terms of components or services in specific computing environments, e.g. grid, cloud. To maximize impact, a good balance between conceptual aspects (e.g. algorithms) and practical aspects (e.g. software development) is necessary. We design and validate our solutions by working closely with scientific application partners (CIRAD, INRA, IRD, etc.). To further validate our solutions and extend the scope of our results, we also want to foster industrial collaborations, even in non scientific applications, provided that they exhibit similar challenges.

3. Research Program

3.1. Data Management

Data management is concerned with the storage, organization, retrieval and manipulation of data of all kinds, from small and simple to very large and complex. It has become a major domain of computer science, with a large international research community and a strong industry. Continuous technology transfer from research to industry has led to the development of powerful DBMSs, now at the heart of any information system, and of advanced data management capabilities in many kinds of software products (application servers, document systems, search engines, directories, etc.).

The fundamental principle behind data management is *data independence*, which enables applications and users to deal with the data at a high conceptual level while ignoring implementation details. The relational model, by resting on a strong theory (set theory and first-order logic) to provide data independence, has revolutionized data management. The major innovation of relational DBMS has been to allow data manipulation through queries expressed in a high-level (declarative) language such as SQL. Queries can then be automatically translated into optimized query plans that take advantage of underlying access methods and indices. Many other advanced capabilities have been made possible by data independence : data and metadata modeling, schema management, consistency through integrity rules and triggers, transaction support, etc.

This data independence principle has also enabled DBMS to continuously integrate new advanced capabilities such as object and XML support and to adapt to all kinds of hardware/software platforms from very small smart devices (smart phone, PDA, smart card, etc.) to very large computers (multiprocessor, cluster, etc.) in distributed environments.

Following the invention of the relational model, research in data management has continued with the elaboration of strong database theory (query languages, schema normalization, complexity of data management algorithms, transaction theory, etc.) and the design and implementation of DBMS. For a long time, the focus was on providing advanced database capabilities with good performance, for both transaction processing and decision support applications. And the main objective was to support all these capabilities within a single DBMS.

The problems of scientific data management (massive scale, complexity and heterogeneity) go well beyond the traditional context of DBMS. To address them, we capitalize on scientific foundations in closely related domains: distributed data management, cloud data management, big data, uncertain data management, metadata integration, data mining and content-based information retrieval.

3.2. Distributed Data Management

To deal with the massive scale of scientific data, we exploit large-scale distributed systems, with the objective of making distribution transparent to the users and applications. Thus, we capitalize on the principles of large-scale distributed systems such as clusters, peer-to-peer (P2P) and cloud, to address issues in data integration, scientific workflows, recommendation, query processing and data analysis.

Data management in distributed systems has been traditionally achieved by distributed database systems which enable users to transparently access and update several databases in a network using a high-level query language (e.g. SQL) [11]. Transparency is achieved through a global schema which hides the local databases' heterogeneity. In its simplest form, a distributed database system is a centralized server that supports a global schema and implements distributed database techniques (query processing, transaction management, consistency management, etc.). This approach has proved to be effective for applications that can benefit from centralized control and full-fledge database capabilities, e.g. information systems. However, it cannot scale up to more than tens of databases. Data integration systems, e.g. price comparators such as KelKoo, extend the distributed database approach to access data sources on the Internet with a simpler query language in read-only mode.

Parallel database systems extend the distributed database approach to improve performance (transaction throughput or query response time) by exploiting database partitioning using a multiprocessor or cluster system. Although data integration systems and parallel database systems can scale up to hundreds of data sources or database partitions, they still rely on a centralized global schema and strong assumptions about the network.

Scientific workflow management systems (SWfMS) such as Kepler (http://kepler-project.org) and Taverna (http://www.taverna.org.uk) allow scientists to describe and execute complex scientific procedures and activities, by automating data derivation processes, and supporting various functions such as provenance management, queries, reuse, etc. Some workflow activities may access or produce huge amounts of distributed data and demand high performance computing (HPC) environments with highly distributed data sources and computing resources. However, combining SWfMS with HPC to improve throughput and performance remains a difficult challenge. In particular, existing workflow development and computing environments have limited support for data parallelism patterns. Such limitation makes complex the automation and ability to perform efficient parallel execution on large sets of data, which may significantly slow down the execution of a workflow.

In contrast, peer-to-peer (P2P) systems [9] adopt a completely decentralized approach to data sharing. By distributing data storage and processing across autonomous peers in the network, they can scale without the need for powerful servers. Popular examples of P2P systems such as Gnutella and BitTorrent have millions of users sharing petabytes of data over the Internet. Although very useful, these systems are quite simple (e.g.

file sharing), support limited functions (e.g. keyword search) and use simple techniques (e.g. resource location by flooding) which have performance problems. To deal with the dynamic behavior of peers that can join and leave the system at any time, they rely on the fact that popular data get massively duplicated.

Initial research on P2P systems has focused on improving the performance of query routing in the unstructured systems which rely on flooding, whereby peers forward messages to their neighbors. This work led to structured solutions based on Distributed Hash Tables (DHT), e.g. CHORD and Pastry, or hybrid solutions with super-peers that index subsets of peers. Another approach is to exploit gossiping protocols, also known as epidemic protocols. Gossiping has been initially proposed to maintain the mutual consistency of replicated data by spreading replica updates to all nodes over the network. It has since been successfully used in P2P networks for data dissemination. Basic gossiping is simple. Each peer has a complete view of the network (i.e., a list of all peers' addresses) and chooses a node at random to spread the request. The main advantage of gossiping is robustness over node failures since, with very high probability, the request is eventually propagated to all nodes in the network. In large P2P networks, however, the basic gossiping model does not scale as maintaining the complete view of the network at each node would generate very heavy communication traffic. A solution to scalable gossiping is by having each peer with only a partial view of the network, e.g. a list of tens of neighbor peers. To gossip a request, a peer chooses at random a peer in its partial view to send it the request. In addition, the peers involved in a gossip exchange their partial views to reflect network changes in their own views. Thus, by continuously refreshing their partial views, nodes can self-organize into randomized overlays which scale up very well.

We claim that a P2P solution is the right solution to support the collaborative nature of scientific applications as it provides scalability, dynamicity, autonomy and decentralized control. Peers can be the participants or organizations involved in collaboration and may share data and applications while keeping full control over their (local) data sources.

But for very-large scale scientific data analysis or to execute very large data-intensive workflow activities (activities that manipulate huge amounts of data), we believe cloud computing (see next section), is the right approach as it can provide virtually infinite computing, storage and networking resources. However, current cloud architectures are proprietary, ad-hoc, and may deprive users of the control of their own data. Thus, we postulate that a hybrid P2P/cloud architecture is more appropriate for scientific data management, by combining the bests of both approaches. In particular, it will enable the clean integration of the users' own computational resources with different clouds.

3.3. Cloud Data Management

Cloud computing encompasses on demand, reliable services provided over the Internet (typically represented as a cloud) with easy access to virtually infinite computing, storage and networking resources. Through very simple Web interfaces and at small incremental cost, users can outsource complex tasks, such as data storage, system administration, or application deployment, to very large data centers operated by cloud providers. Thus, the complexity of managing the software/hardware infrastructure gets shifted from the users' organization to the cloud provider. From a technical point of view, the grand challenge is to support in a cost-effective way the very large scale of the infrastructure which has to manage lots of users and resources with high quality of service.

Cloud customers could move all or part of their information technology (IT) services to the cloud, with the following main benefits:

- **Cost.** The cost for the customer can be greatly reduced since the IT infrastructure does not need to be owned and managed; billing is only based only on resource consumption. For the cloud provider, using a consolidated infrastructure and sharing costs for multiple customers reduces the cost of ownership and operation.
- **Ease of access and use.** The cloud hides the complexity of the IT infrastructure and makes location and distribution transparent. Thus, customers can have access to IT services anytime, and from anywhere with an Internet connection.

- **Quality of Service (QoS).** The operation of the IT infrastructure by a specialized provider that has extensive experience in running very large infrastructures (including its own infrastructure) increases QoS.
- **Elasticity.** The ability to scale resources out, up and down dynamically to accommodate changing conditions is a major advantage. In particular, it makes it easy for customers to deal with sudden increases in loads by simply creating more virtual machines.

However, cloud computing has some drawbacks and not all applications are good candidates for being "cloudified". The major concern is w.r.t. data security and privacy, and trust in the provider (which may use no so trustful providers to operate). One earlier criticism of cloud computing was that customers get locked in proprietary clouds. It is true that most clouds are proprietary and there are no standards for cloud interoperability. But this is changing with open source cloud software such as Hadoop, an Apache project implementing Google's major cloud services such as Google File System and MapReduce, and Eucalyptus, an open source cloud software infrastructure, which are attracting much interest from research and industry.

There is much more variety in cloud data than in scientific data since there are many different kinds of customers (individuals, SME, large corporations, etc.). However, we can identify common features. Cloud data can be very large, unstructured (e.g. text-based) or semi-structured, and typically append-only (with rare updates). And cloud users and application developers may be in high numbers, but not DBMS experts.

3.4. Big Data

Big data has become a buzz word, with different meanings depending on your perspective, e.g. 100 terabytes is big for a transaction processing system, but small for a web search engine. It is also a moving target, as shown by two landmarks in DBMS products: the Teradata database machine in the 1980's and the Oracle Exadata database machine in 2010.

Although big data has been around for a long time, it is now more important than ever. We can see overwhelming amounts of data generated by all kinds of devices, networks and programs, e.g. sensors, mobile devices, internet, social networks, computer simulations, satellites, radiotelescopes, etc. Storage capacity has doubled every 3 years since 1980 with prices steadily going down (e.g. 1 Gigabyte for: 1M$ in 1982, 1K$ in 1995, 0.12$ in 2011), making it affordable to keep more data. And massive data can produce high-value information and knowledge, which is critical for data analysis, decision support, forecasting, business intelligence, research, (data-intensive) science, etc.

The problem of big data has three main dimensions, quoted as the three big V's:

- Volume: refers to massive amounts of data, making it hard to store, manage, and analyze (big analytics);
- Velocity: refers to continuous data streams being produced, making it hard to perform online processing and analysis;
- Variety: refers to different data formats, different semantics, uncertain data, multiscale data, etc., making it hard to integrate and analyze.

There are also other V's like: validity (is the data correct and accurate?); veracity (are the results meaningful?); volatility (how long do you need to store this data?).

Current big data management (NoSQL) solutions have been designed for the cloud, as cloud and big data are synergistic. They typically trade consistency for scalability, simplicity and flexibility. They use a radically different architecture than RDBMS, by exploiting (rather than embedding) a distributed file system such as Google File System (GFS) or Hadoop Distributed File System (HDFS), to store and manage data in a highly fault-tolerant manner. They tend to rely on a more specific data model, e.g. key-value store such as Google Bigtable, Hadoop Hbase or Apache CouchDB) with a simple set of operators easy to use from a programming language. For instance, to address the requirements of social network applications, new solutions rely on a graph data model and graph-based operators. User-defined functions also allow for more specific data processing. MapReduce is a good example of generic parallel data processing framework, on top of a

distributed file system (GFS or HDFS). It supports a simple data model (sets of (key, value) pairs), which allows user-defined functions (map and reduce). Although quite successful among developers, it is relatively low-level and rigid, leading to custom user code that is hard to maintain and reuse. In Zenith, we exploit or extend MapReduce and NoSQL technologies to fit our needs for scientific workflow management and scalable data analysis.

3.5. Uncertain Data Management

Data uncertainty is present in many scientific applications. For instance, in the monitoring of plant contamination by INRA teams, sensors generate periodically data which may be uncertain. Instead of ignoring (or correcting) uncertainty, which may generate major errors, we need to manage it rigorously and provide support for querying.

To deal with uncertainty, there are several approaches, e.g. probabilistic, possibilistic, fuzzy logic, etc. The *probabilistic approach* is often used by scientists to model the behavior of their underlying environments. However, in many scientific applications, data management and uncertain query processing are not integrated, i.e., the queries are usually answered using ad-hoc methods after doing manual or semi-automatic statistical treatment on the data which are retrieved from a database. In Zenith, we aim at integrating scientific data management and query processing within one system. This should allow scientists to issue their queries in a query language without thinking about the probabilistic treatment which should be done in background in order to answer the queries. There are two important issues which any PDBMS should address: 1) how to represent a probabilistic database, i.e., data model; 2) how to answer queries using the chosen representation, i.e., query evaluation.

One of the problems on which we focus is *scalable query processing* over uncertain data. A naive solution for evaluating probabilistic queries is to enumerate all possible worlds, i.e., all possible instances of the database, execute the query in each world, and return the possible answers together with their cumulative probabilities. However, this solution can not scale up due to the exponential number of possible worlds which a probabilistic database may have. Thus, the problem is quite challenging, particularly due to the exponential number of possibilities that should be considered for evaluating queries. In addition, most of our underlying scientific applications are not centralized; the scientists share part of their data in a *P2P* manner. This distribution of data makes very complicated the processing of probabilistic queries. To develop efficient query processing techniques for distributed scientific applications, we can take advantage of two main distributed technologies: *P2P* and *Cloud*. Our research experience in P2P systems has proved us that we can propose scalable solutions for many data management problems. In addition, we can use the cloud parallel solutions, e.g. MapReduce, to parallelize the task of query processing, when possible, and answer queries of scientists in reasonable execution times. Another challenge for supporting scientific applications is uncertain data integration. In addition to managing the uncertain data for each user, we need to integrate uncertain data from different sources. This requires revisiting traditional data integration in major ways and dealing with the problems of uncertain mediated schema generation and uncertain schema mapping.

3.6. Big data Integration

Nowadays, scientists can rely on web 2.0 tools to quickly share their data and/or knowledge (e.g. ontologies of the domain knowledge). Therefore, when performing a given study, a scientist would typically need to access and integrate data from many data sources (including public databases). To make high numbers of scientific data sources easily accessible to community members, it is necessary to identify semantic correspondences between metadata structures or models of the related data sources. The main underlying task is called matching, which is the process of discovering semantic correspondences between metadata structures such as database schema and ontologies. Ontology is a formal and explicit description of a shared conceptualization in terms of concepts (i.e., classes, properties and relations). For example, the matching may be used to align gene ontologies or anatomical metadata structures.

To understand a data source content, metadata (data that describe the data) is crucial. Metadata can be initially provided by the data publisher to describe the data structure (e.g. schema), data semantics based on ontologies (that provide a formal representation of the domain knowledge) and other useful information about data provenance (publisher, tools, methods, etc.). Scientific metadata is very heterogeneous, in particular because of the great autonomy of the underlying data sources, which leads to a large variety of models and formats. The high heterogeneity makes the matching problem very challenging. Furthermore, the number of ontologies and their size grow fastly, and so does their diversity and heterogeneity. As a result, schema/ontology matching has become a prominent and challenging topic.

3.7. Data Mining

Data mining provides methods to discover new and useful patterns from very large sets of data. These patterns may take different forms, depending on the end-user's request, such as:

- **Frequent itemsets and association rules**. In this case, the data is usually a table with a high number of rows and the algorithm extracts correlations between column values. This problem was first motivated by commercial and marketing purposes (*e.g.* discovering frequent correlations between items bought in a shop, which could help selling more). A typical example of frequent itemset from a sensor network in a smart building would say that "in 20% rooms, the door is closed, the room is empty, and lights are on."
- **Frequent sequential pattern extraction.** This problem is very similar to frequent itemset mining, but in this case, the order between events has to be considered. Let us consider the smart-building example again. A frequent sequence, in this case, could say that "in 40% of rooms, lights are on at time i, the room is empty at time i+j and the door is closed at time i+j+k". Discovering frequent sequences has become a crucial need in marketing, but also in security (detecting network intrusions for instance) in usage analysis (web usage is one of the main applications) and any domain where data arrive in a specific order (usually given by timestamps).
- **Clustering**. The goal of clustering algorithms is to group together data that have similar characteristics, while ensuring that dissimilar data will not be in the same cluster. In our example of smart buildings, we would find clusters of rooms, where offices will be in one category and copy machine rooms in another one because of their characteristics (hours of people presence, number of times lights are turned on and off, etc.).

One of the main problems for data mining methods has been to deal with data streams. Actually, data mining methods have first been designed for very large data sets where complex algorithms of artificial intelligence were not able to complete within reasonable time responses because of data size. The problem was thus to find a good trade-off between response time and results relevance. The patterns described above well match this trade-off since they both provide interesting knowledge for data analysts and allow algorithm having good time complexity on the number of records. Itemset mining algorithms, for instance, depend more on the number of columns (for a sensor it would be the number of possible items such as temperature, presence, status of lights, etc.) than the number of lines (number of sensors in the network). However, with the ever growing size of data and their production rate, a new kind of data source has recently emerged as data streams. A data stream is a sequence of events arriving at high rate. By "high rate", we usually admit that traditional data mining methods reach their limits and cannot complete in real-time, given the data size. In order to extract knowledge from such streams, a new trade-off had to be found and the data mining community has investigated approximation methods that could allow to maintain a good quality of results for the above patterns extraction.

For scientific data, data mining now has to deal with new and challenging characteristics. First, scientific data is often associated to a level of uncertainty (typically, sensed values have to be associated to the probability that this value is correct or not). Second, scientific data might be extremely large and need cloud computing solutions for their storage and analysis. Eventually, we will have to deal with high dimension and heterogeneous data.

3.8. Content-based Information Retrieval

Today's technologies for searching information in scientific data mainly rely on relational DBMS or text-based indexing methods. However, content-based information retrieval has progressed much in the last decade and is now considered as one of the most promising for future search engines. Rather than restricting search to the use of metadata, content-based methods attempt to index, search and browse digital objects by means of signatures describing their actual content. Such methods have been intensively studied in the multimedia community to allow searching the massive amount or raw multimedia documents created every day (e.g. 99% of web data are audio-visual content with very sparse metadata). Successful and scalable content-based methods have been proposed for searching objects in large image collections or detecting copies in huge video archives. Besides multimedia contents, content-based information retrieval methods recently started to be studied on more diverse data such as medical images, 3D models or even molecular data. Potential applications in scientific data management are numerous. First of all, to allow searching the huge collections of scientific images (earth observation, medical images, botanical images, biology images, etc.) but also to browse large datasets of experimental data (e.g. multisensor data, molecular data or instrumental data). Despite recent progress, scalability remains a major issue, involving complex algorithms (such as similarity search, clustering or supervised retrieval), in high dimensional spaces (up to millions of dimensions) with complex metrics (Lp, Kernels, sets intersections, edit distances, etc.). Most of these algorithms have linear, quadratic or even cubic complexities so that their use at large scale is not affordable without consistent breakthrough. In Zenith, we plan to investigate the following challenges:

- **High-dimensional similarity search**. Whereas many indexing methods were designed in the last 20 years to efficiently retrieve multidimensional data with relatively small dimensions, high-dimensional data have been more challenging due to the well-known dimensionality curse. Only recently have some methods appeared that allow approximate Nearest Neighbors queries in sub-linear time. In particular, Locality Sensitive Hashing methods which offer new theoretical insights in high-dimensional Euclidean spaces and proved the interest of random projections. But there are still some challenging issues that need to be solved including efficient similarity search in any kernel or metric spaces, efficient construction of knn-graphs or relational similarity queries.
- **Large-scale supervised retrieval**. Supervised retrieval aims at retrieving relevant objects in a dataset by providing some positive and/or negative training samples. To solve such a task, there has been a focused interest on using Support Vector Machines (SVM) that offer the possibility to construct generalized, non-linear predictors in high-dimensional spaces using small training sets. The prediction time complexity of these methods is usually linear in dataset size. Allowing hyperplane similarity queries in sub-linear time is for example a challenging research issue. A symmetric problem in supervised retrieval consists in retrieving the most relevant object categories that might contain a given query object, providing huge labeled datasets (up to millions of classes and billions of objects) and very few objects per category (from 1 to 100 objects). SVM methods that are formulated as quadratic programming with cubic training time complexity and quadratic space complexity are clearly not usable. Promising solutions to such problems include hybrid supervised-unsupervised methods and supervised hashing methods.
- **Distributed content-based retrieval**. Distributed content-based retrieval methods appeared recently as a promising solution to manage masses of data distributed over large networks, particularly when the data cannot be centralized for privacy or cost reasons (which is often the case in scientific social networks, e.g. botanist social networks). However, current methods are limited to very simple similarity search paradigms. In Zenith, we will consider more advanced distributed content-based retrieval and mining methods such as k-nn graphs construction, large-scale supervised retrieval or multi-source clustering.

4. Application Domains

4.1. Data-intensive Scientific Applications

The application domains covered by Zenith are very wide and diverse, as they concern data-intensive scientific applications, i.e., most scientific applications. Since the interaction with scientists is crucial to identify and tackle data management problems, we are dealing primarily with application domains for which Montpellier has an excellent track record, i.e., agronomy, environmental science, life science, with scientific partners like INRA, IRD, CIRAD and IRSTEA. However, we are also addressing other scientific domains (e.g. astronomy, oil extraction) through our international collaborations (e.g. in Brazil).

Let us briefly illustrate some representative examples of scientific applications on which we have been working on.

- **Management of astronomical catalogs**. An example of data-intensive scientific applications is the management of astronomical catalogs generated by the Dark Energy Survey (DES) project on which we are collaborating with researchers from Brazil. In this project, huge tables with billions of tuples and hundreds of attributes (corresponding to dimensions, mainly double precision real numbers) store the collected sky data. Data are appended to the catalog database as new observations are performed and the resulting database size is estimated to reach 100TB very soon. Scientists around the globe can query the database with queries that may contain a considerable number of attributes. The volume of data that this application holds poses important challenges for data management. In particular, efficient solutions are needed to partition and distribute the data in several servers. An efficient partitioning scheme should try to minimize the number of fragments accessed in the execution of a query, thus reducing the overhead associated to handle the distributed execution.
- **Personal health data analysis and privacy** The "Quantified Self" movement has gained a large popularity these past few years. Today, it is possible to acquire data on many domains related to personal data. For instance, one can collect data on her daily activities, habits or health. It is also possible to measure performances in sports. This can be done thanks to sensors, communicating devices or even connected glasses (as currently being developped by companies such as Google, for instance). Obviously, such data, once acquired, can lead to valuable knowledge for these domains. For people having a specific disease, it might be important to know if they belong to a specific category that needs particular care. For an individual, it can be interesting to find a category that corresponds to her performances in a specific sport and then adapt her training with an adequate program. Meanwhile, for privacy reasons, people will be reluctant to share their personal data and make them public. Therefore, it is important to provide them solutions that can extract such knowledge from everybody's data, while guaranteeing that their private data won't be disclosed to anyone.
- **Botanical data sharing**. Botanical data is highly decentralized and heterogeneous. Each actor has its own expertise domain, hosts its own data, and describes them in a specific format. Furthermore, botanical data is complex. A single plant's observation might include many structured and unstructured tags, several images of different organs, some empirical measurements and a few other contextual data (time, location, author, etc.). A noticeable consequence is that simply identifying plant species is often a very difficult task; even for the botanists themselves (the so-called taxonomic gap). Botanical data sharing should thus speed up the integration of raw observation data, while providing users an easy and efficient access to integrated data. This requires to deal with social-based data integration and sharing, massive data analysis and scalable content-based information retrieval. We address this application in the context of the French initiative Pl@ntNet, with CIRAD and IRD.
- **Biology data integration and analysis**.

  Biology and its applications, from medicine to agronomy and ecology, are now producing massive data, which is revolutionizing the way life scientists work. For instance, using plant phenotyping platforms such as PhenoDyn, PhenoPsis and PhenoArch at INRA Montpellier, quantitative genetic methods allow to identify genes involved in phenotypic variation in response to environmental conditions. These methods produce large amounts of data at different time intervals (minutes to

days), at different sites and at different scales ranging from small tissue samples until the entire plant. Analyzing such big data creates new challenges for data management and data integration.

These application examples illustrate the diversity of requirements and issues which we are addressing with our scientific application partners. To further validate our solutions and extend the scope of our results, we also want to foster industrial collaborations, even in non scientific applications, provided that they exhibit similar challenges.

5. Highlights of the Year

5.1. Highlights of the Year

The Pl@ntNet application, co-developed by Zenith, exceeded 1M downloads in October 2015.

6. New Software and Platforms

6.1. Hadoop_g5k

Participants: Reza Akbarinia, Miguel Liroz-Gistau, Patrick Valduriez.

URL: https://www.grid5000.fr/mediawiki/index.php/Hadoop_On_Execo

Apache Hadoop provides an open-source framework for reliable, scalable, parallel computing. It can be deployed and used in large-scale platforms such as Grid 5000. However, its configuration and management is very difficult, specially under the dynamic nature of clusters. Therefore, we built Hadoop_g5k (Hadoop easy deployment in clusters), a tool that makes it easier to manage Hadoop clusters and prepare reproducible experiments. Hadoop_g5k offers a set of scripts to be used in command-line interfaces and a Python interface. It is actually used by Grid5000 users, and helps them saving much time when doing their experiments with MapReduce.

6.2. LogMagnet

Participants: Julien Diener, Florent Masseglia.

URL: https://team.inria.fr/zenith/software/LogMagnet

LogMagnet is a software for analyzing streaming data, and in particular log data. Log data usually arrive in the form of lines containing activities of human or machines. In the case of human activities, it may be the behavior on a Web site or the usage of an application. In the case of machines, such log may contain the activities of software and hardware components (say, for each node of a computing cluster, the calls to system functions or some hardware alerts). Analyzing such data is often difficult and crucial in the meanwhile. LogMagnet allows to summarize this data, and to provide a first analysis as a clustering. This summary may also be exploited as easily as the original data.

6.3. MultiSite-Rec

Participants: Mohamed Reda Bouadjenek, Florent Masseglia, Esther Pacitti.

Recommender systems are used as a mean to supply users with content that may be of interest to them. They have become a popular research topic, where many aspects and dimensions have been studied to make them more accurate and effective. In practice, recommender systems suffer from cold-start problems. However, users use many online services, which can provide information about their interest and the content of items (e.g. Google search engine, Facebook, Twitter, etc). These services may be valuable data sources, which supply information to help a recommender system in modeling users and items' preferences, and thus, make the recommender system more precise. Moreover, these data sources are distributed, and geographically distant from each other, which raise many research problems and challenges to design a distributed recommendation algorithm. MultiSite-Rec is a distributed collaborative filtering algorithm, which exploits and combine these multiple and heterogeneous data sources to improve the recommendation quality.

6.4. ThePlantGame: crowdsourced plants identification

Participants: Maximilien Servajean [contact], Alexis Joly, Julien Champ.

URL: http://theplantgame.com/

The Plant Game is a participatory game whose purpose is the production of large masses of taxonomic data to improve our knowledge of biodiversity. The interest of the game is twofold: (i) train and progress in botany while having fun, and (ii) participate to a large citizen sciences project in biodiversity. The game relies on consistent scientific contributions compared to classical crowdsourcing models and algorithms that are not scalable to classification problems with thousands of complex classes such as plant species. The most remarkable one is the active training of the users based on innovative sub-task creation and assignment processes that are adaptive to the increasing skills of the user. The first public version of the game was released in July 2015. Nowadays, about 1000 players are registered and produce on average about 35 new validated plant observations per day. The accuracy of the produced taxonnomic tags is about 94%, which is quite impressive considering the fact that a majority of users are beginners when they start playing.

6.5. Pl@ntNet

Participants: Julien Champ, Hervé Goëau, Alexis Joly.

URL: http://goo.gl/CpSrr3

Pl@ntNet is an image sharing and retrieval application for the identification of plants. It is developed in the context of the Floris'tic project that involves four French research organisations (Inria, Cirad, INRA, IRD) and Tela Botanica social network. The key feature of the iOS and Android front ends is to help identifying plant species from photographs, through a server-side visual search engine based on several results of ZENITH team on content-based information retrieval. Since its first release in March 2013 on the apple store, the application was downloaded by around 1M users in more than 170 countries (between 2,500 and 10,000 active users daily with peaks occurring during the week-ends). The collaborative training set that allows the content-based identification is continuously enriched by the users of the application and the members of Tela Botanica social network. At the time of writing, it includes about 200K images covering more than 5000 French plant species about 4/5 of the whole French flora (this is actually the widest identification tool built anytime).

6.6. Snoop & SnoopIm

Participants: Alexis Joly, Julien Champ, Jean-Christophe Lombardo.

URL: http://otmedia.lirmm.fr/

Snoop is a generalist C++ library dedicated to high-dimensional data management and efficient similarity search. Its main features are dimension reduction, high-dimensional feature vectors hashing, approximate k-nearest neighbors search and Hamming embedding. Snoop is a refactoring of a previous library called PMH developed jointly with the French National Institute of Audiovisual. It is based on the joined research work of Alexis Joly and Olivier Buisson. SnoopIm is a content-based image search engine built on top of Snoop and allowing to retrieve small visual patterns or objects in large collections of pictures. The software is being experimented/used in several contexts including a logo retrieval application set up in collaboration with INA (DigInPix: http://diginpix.ina.fr), a whale's individuals matching application set up in collaboration with CetaMada NGO (IdentyWhale, to be publicly released soon), a hieroglyphs recognition application currently under development in collaboration with the Egyptology department of Montpellier University Paul-Valéry.

6.7. SciFloware

Participants: Dimitri Dupuis, Didier Parigot.

URL: http://www-sop.inria.fr/members/Didier.Parigot/pmwiki/Scifloware

SciFloware is an action of technology development (ADT Inria) with the goal of developing a middleware for the execution of scientific workflows in a distributed and parallel way. It capitalizes on our experience with SON and an innovative algebraic approach to the management of scientific workflows. SciFloware provides a development environment and a runtime environment for scientific workflows, interoperable with existing systems. We validate SciFloware with workflows for analyzing biological data provided by our partners CIRAD, INRA and IRD.

6.8. CloudMdsQL Compiler

Participants: Carlyna Bondiombouy, Boyan Kolev, Oleksandra Levchenko, Patrick Valduriez.

URL: http://cloudmdsql.gforge.inria.fr

The CloudMdsQL (Cloud Multi-datastore Query Language) compiler transforms queries expressed in a common SQL-like query language into an optimized query execution plan to be executed over multiple cloud data stores (SQL, NoSQL, HDFS, etc.) through a query engine. The compiler/optimizer is implemented in C++ and uses the Boost.Spirit framework for parsing context-free grammars. CloudMdsQL is being validated on relational, document and graph data stores in the context of the CoherentPaaS European project.

6.9. Chiaroscuro

Participants: Tristan Allard, Florent Masseglia, Esther Pacitti.

URL: http://people.irisa.fr/Tristan.Allard/chiaroscuro/

Chiaroscuro is a software developped in the context of a research contract with EDF. It aims at clustering time series with privacy preserving guarantees. It is a distributed system, working in a P2P environment. It is used by the team for experiments and by EDF as a proof-of-concept. Chiaroscuro is the first software for that purpose. It is written in Java. The distributed algorithm implemented in Chiaroscuro has been filed by EDF in a patent (with Inria and University of Montpellier)

6.10. FP-Hadoop

Participants: Reza Akbarinia, Miguel Liroz, Patrick Valduriez.

https://gforge.inria.fr/plugins/mediawiki/wiki/fp-hadoop

FP-Hadoop is an extension of Hadoop that efficiently deals with the problem of data skew in MapReduce jobs. In FP-Hadoop, there is a new phase, called intermediate reduce (IR), in which blocks of intermediate values, constructed dynamically, are processed by intermediate reduce workers in parallel, by using a scheduling strategy.

7. New Results

7.1. Big Data Integration

7.1.1. CloudMdsQL, a query language for heterogeneous data stores

Participants: Carlyna Bondiombouy, Boyan Kolev, Oleksandra Levchenko, Patrick Valduriez.

The blooming of different cloud data management infrastructures, specialized for different kinds of data and tasks, has led to a wide diversification of DBMS interfaces and the loss of a common programming paradigm. The CoherentPaaS European project addresses this problem, by providing a common programming language and holistic coherence across different cloud data stores.

In this context, we have started the design of a Cloud Multi-datastore Query Language (CloudMdsQL), and its query engine. CloudMdsQL is a functional SQL-like language, capable of querying multiple heterogeneous data stores, e.g. relational, NoSQL or HDFS) [19], [31]. The major innovation is that a CloudMdsQL query can exploit the full power of the local data stores, by simply allowing some local data store native queries to be called as functions, and at the same time be optimized. Our experimental validation, with three data stores (graph, document and relational) and representative queries, shows that CloudMdsQL satisfies the five important requirements for a cloud multidatastore query language. In [32], we extend CloudMdsQL to allowing the ad-hoc usage of user defined map/filter/reduce operators in combination with traditional SQL statements, to integrate relational data and big data stored in HDFS and accessed by a data processing framework like Spark.

7.1.2. Semantic Data Integration using Bio-Ontologies

Participant: Pierre Larmande.

The AgroPortal project [49] aims at developing and supporting a reference ontology repository for the agronomic domain. The ontology portal features ontology hosting, search, versioning, visualization, comment, with services for semantically annotating data with the ontologies, as well as storing and exploiting ontology alignments and data annotations. All of these within a fully semantic web compliant infrastructure. The main objective of this project is to enable straightforward use of agronomic related ontologies, avoiding data managers and researchers the burden to deal with complex knowledge engineering issues to annotate the research data. Thus, we specifically pay attention to the requirements of the agronomic community and the specificities of the crop domain. AgroPortal will offer a robust and stable platform that we anticipate will be highly valued by the community.

7.1.3. Access and Integration of Molecular Biology Data

Participants: Sarah Cohen-Boulakia, Patrick Valduriez.

The volumes of molecular biology data available on the web are constantly increasing. Accessing and integrating these data is crucial for making progress in biology. In [26], we provide all the necessary pointers to identify the reference databases capable of providing bioinformatic data for molecular biology. We also discuss the problems posed by the exploitation of these very highly heterogeneous and distributed data. Finally, in order to guide a prospective user on the choice of one of these systems, we provide an overview of the systems that provide unified access to these data.

7.2. Distributed Indexing and Searching

7.2.1. Diversified and Distributed Recommendation for Scientific Data

Participants: Esther Pacitti, Maximilien Servajean.

Recommendation is becoming a popular mechanism to help users find relevant information in large-scale data (scientific data, web). To avoid redundancy in the results, recommendation diversification has been proposed, with the objective of identifying items that are dissimilar, but nonetheless relevant to the user's interests.

We propose a new diversified search and recommendation solution suited for scientific data (i.e., plant phenotyping, botanical data) [22] We first define an original profile diversification scoring function that enables to address the problem of returning redundant items, and enhances the quality of diversification. Through experimental evaluation using two benchmarks, we showed that our scoring function gives the best compromise between diversity and relevancy. Next, to implement our new scoring function, we propose a basic Top-k threshold-based algorithm that exploits a candidate list to achieve diversification and several techniques to improve performance. First, we simplify the scoring model to reduce its computational complexity. Second, we propose two techniques to reduce the number of items in the candidate list, and thus the number of diversified scores to compute. Third, we propose different indexing scores that take into account the diversification of items and an adaptive indexing approach to reduce the number of accesses in the index dynamically based on the queries workload. The experimentation results show that our techniques yield a major reduction of response time, up to 12 times compared to a baseline greedy diversification algorithm.

We also address distributed and diversified recommendation in the context of P2P and multisite cloud [23]. We propose a new scoring function (usefulness) to cluster relevant users over a distributed overlay. Our experimental evaluation using different datasets shows major gains in recall (order of 3 times) compared with state-of-the-art solutions.

7.3. Scientific Workflows

7.3.1. Scientific Workflows: combining data analysis and simulation

Participant: Sarah Cohen-Boulakia.

While scientific workflows are increasingly popular in the bioinformatics community in some emerging application domains such as ecology, the need for data analysis is combined with the need to model complex multi-scale biological systems, possibly involving multiple simulation steps. This requires the scientific workflow to deal with retro-action to understand and predict the relationships between structure and function of these complex systems. OpenAlea (openalea.gforge.inria.fr) developped by the EPI Virtual plants is the only scientific workflow system able to uniformly address the problem, which made it successful in the scientific community.

For the first time, we proposed a conceptualisation of OpenAlea in [42]. We introduce the concept of higher-order dataflows as a means to uniformly combine classical data analysis with modeling and simulation. We provide for the first time the description of the OpenAlea system involving an original combination of features. We illustrate the demonstration on a high-throughput workflow in phenotyping, phenomics, and environmental control designed to study the interplay between plant architecture and climatic change. Ongoing work include deploying OpenAlea on a Grid technology using the SciFloware middleware.

7.3.2. Processing Scientific Workflows in Multi-site cloud

Participants: Ji Liu, Esther Pacitti, Patrick Valduriez.

As the scale of the data increases, scientific workflow management systems (SWfMSs) need to support workflow execution in High Performance Computing (HPC) environments. Because of various benefits, cloud emerges as an appropriate infrastructure for workflow execution. However, it is difficult to execute some scientific workflows in one cloud site because of geographical distribution of scientists, data and computing resources. Therefore, a scientific workflow often needs to be partitioned and executed in a multisite environment.

In [21], we define a multisite cloud architecture that is composed of traditional clouds, e.g., a pay-per-use cloud service such as Amazon EC2, private data-centers, e.g. a cloud of a scientific organization like Inria, COPPE or LNCC, and client desktop machines that have authorized access to the data-centers. We can model this architecture as a distributed system on the Internet, each site having its own computer cluster, data and programs. An important requirement is to provide distribution transparency for advanced services (i.e., workflow management, data analysis), to ease their scalability and elasticity. Current solutions for multisite clouds typically rely on application specific overlays that map the output of one task at a site to the input of another in a pipeline fashion. Instead, we define fully distributed services for data storage, intersite data movement and task scheduling.

7.3.3. Data-centric Iteration in Dynamic Workflows

Participant: Patrick Valduriez.

Dynamic workflows are scientific workflows supporting computational science simulations, typically using dynamic processes based on runtime scientific data analyses. They require the ability of adapting the workflow, at runtime, based on user input and dynamic steering. Supporting data-centric iteration is an important step towards dynamic workflows because user interaction with workflows is iterative. However, current support for iteration in scientific workflows is static and does not allow for changing data at runtime.

In [17], we propose a solution based on algebraic operators and a dynamic execution model to enable workflow adaptation based on user input and dynamic steering. We introduce the concept of iteration lineage that makes provenance data management consistent with dynamic iterative workflow changes. Lineage enables scientists to interact with workflow data and configuration at runtime through an API that triggers steering. We evaluate our approach using a novel and real large-scale workflow for uncertainty quantification on a 640-core cluster. The results show impressive execution time savings from 2.5 to 24 days, compared to non-iterative workflow execution. We verify that the maximum overhead introduced by our iterative model is less than 5% of execution time. Also, our proposed steering algorithms are very efficient and run in less than 1 millisecond, in the worst-case scenario.

7.3.4. Analyzing Related Raw Data Files through Dataflows

Participant: Patrick Valduriez.

Computer simulations may ingest and generate high numbers of raw data files. Most of these files follow a de facto standard format established by the application domain, e.g., FITS for astronomy. Although these formats are supported by a variety of programming languages, libraries and programs, analyzing thousands or millions of files requires developing specific programs. DBMS are not suited for this, because they require loading the raw data and structuring it, which gets heavy at large-scale. Systems like NoDB, RAW and FastBit, have been proposed to index and query raw data files without the overhead of using a DBMS. However, they focus on analyzing one single large file instead of several related files. In this case, when related files are produced and required for analysis, the relationship among elements within file contents must be managed manually, with specific programs to access raw data. Thus, this data management may be time-consuming and error-prone. When computer simulations are managed by a SWfMS, they can take advantage of provenance data to relate and analyze raw data files produced during workflow execution. However, SWfMS register provenance at a coarse grain, with limited analysis on elements from raw data files. When the SWfMS is dataflow-aware, it can register provenance data and the relationships among elements of raw data files altogether in a database which is useful to access the contents of a large number of files. In [24], we propose a dataflow approach for analyzing element data from several related raw data files. Our approach is complementary to the existing single raw data file analysis approaches. We validate our approach with the Montage workflow from astronomy and a workflow from Oil and Gas domain as I/O intensive case studies.

7.4. Scalable Query Processing

7.4.1. Scalable Query Processing with Big Data

Participants: Reza Akbarinia, Miguel Liroz, Patrick Valduriez.

The popular MapReduce parallel processing framework is inefficient in case of data skew, which makes the reduce side done by a few worker nodes.

In [28], [20], we propose FP-Hadoop, which makes the reduce side of MapReduce more parallel. We extend the MapReduce programming model to allow the collaboration of reduce workers on processing the values of an intermediate key, without affecting the correctness of the final results. In FP-Hadoop, the reduce function is replaced by two functions: intermediate reduce and final reduce. There are three phases, each phase corresponding to one of the functions: map, intermediate reduce and final reduce phases. In the intermediate reduce phase, the function, which usually includes the main load of reducing in MapReduce jobs, is executed by reduce workers in a collaborative way, even if all values belong to only one intermediate key. This allows performing a big part of the reducing work by using the computing resources of all workers, even in case of highly skewed data. We implemented a prototype of FP-Hadoop by modifying Hadoop's code, and conducted extensive experiments over synthetic and real datasets. The results show that FP-Hadoop makes MapReduce job processing much faster and more parallel, and can efficiently deal with skewed data. We achieve excellent performance gains compared to native Hadoop, e.g. more than 10 times in reduce time and 5 times in total execution time.

7.5. Data Stream Mining

7.5.1. Summarizing Uncertain Data Streams

Participants: Reza Akbarinia, Florent Masseglia.

Probabilistic data management has shown growing interest to deal with uncertain data. In [29], we focus on probabilistic time series with high volumes of data, which requires efficient compression techniques. To date, most of the work on probabilistic data reduction uses synopses that minimize the error of representation wrt. the original data. However, in most cases, the compressed data will be meaningless for usual queries involving aggregation operators such as SUM or AVG. We propose *PHA* (Probabilistic Histogram Aggregation), a compression technique whose objective is to minimize the error of such queries over compressed probabilistic data. We incorporate the aggregation operator given by the end-user directly in the compression technique, and obtain much lower error in the long term. We also adopt a global error aware strategy in order to manage large sets of probabilistic time series, where the available memory is carefully balanced between the series, according to their individual variability.

7.6. Scalable Data Analysis

7.6.1. Parallel Mining of Maximally Informative k-Itemsets in Big Data

Participants: Saber Salah, Reza Akbarinia, Florent Masseglia.

The discovery of informative itemsets is a fundamental building block in data analytics and information retrieval. While the problem has been widely studied, only few solutions scale. This is particularly the case when i) the data set is massive, and/or ii) the length K of the informative itemset to be discovered is high. In [45], we address the problem of parallel mining of maximally informative k-itemsets (miki) based on joint entropy. We propose PHIKS (Parallel Highly Informative K-itemSets) a highly scalable, parallel mining algorithm. PHIKS renders the mining process of large scale databases (up to terabytes of data) succinct and effective. Its mining process is made up of only two compact, yet efficient parallel jobs. PHIKS uses a clever heuristic approach to efficiently estimates the joint entropies of miki having different sizes with very low upper bound error rate, which dramatically reduces the runtime process. PHIKS has been extensively evaluated using massive, real-world data sets. Our experimental results confirm the effectiveness of our approach by the significant scale-up obtained with high featuresets length and hundreds of millions of objects.

7.6.2. Frequent Itemset Mining in Massively Distributed Environments

Participants: Saber Salah, Reza Akbarinia, Florent Masseglia.

While the problem of Frequent itemset mining (FIM) has been thoroughly studied, few solutions scale. This is mainly the case when i) the amount of data tends to be very large and/or ii) the minimum support (MinSup) threshold is very low. In [46], we study the effectiveness and leverage specific data placement strategies for improving parallel FIM (PFIM) performance in MapReduce, a highly distributed computation framework. By offering a clever data placement and an optimal organization of the extraction algorithms, we show that the itemset discovery effectiveness does not only depend on the deployed algorithms. We propose ODPR (Optimal Data-Process Relationship), a solution for fast mining of frequent itemsets in MapReduce. Our method allows discovering itemsets from massive datasets, where standard solutions do not scale.

In [44], we propose a highly scalable PFIM algorithm, namely Parallel Absolute Top Down (PATD). PATD renders the mining process of very large databases (up to Terabytes) simple and compact. Its mining process is made up of only one parallel job, which dramatically reduces the mining runtime, communication cost and energy power consumption overhead, in a distributed computational platform. Based on a clever and efficient data partitioning strategy, namely Item Based Data Partitioning (IBDP), PATD mines each data partition independently, relying on an absolute minimum support (AM inSup) instead of a relative one. Through an extensive experimental evaluation using real-world data sets, we show that PATD is significantly more efficient and scalable than alternative approaches.

7.6.3. Scalable Mining of Closed Frequent Itemsets

Participants: Mehdi Zitouni, Reza Akbarinia, Florent Masseglia.

Mining big datasets poses a number of challenges which are not easily addressed by traditional mining methods, since both memory and computational requirements are hard to satisfy. One solution is to take advantage of parallel frameworks, such as MapReduce, using ordinary machines. In [48], we address the issue of mining closed frequent itemsets (CFI) from big datasets in such environments. We introduce a new parallel algorithm, called CloPN , for CFI mining. One important feature of CloPN is to use a prime number based approach to transform the data into numerical form, and then to mine closed frequent itemsets by using only multiplication and division operations. We carried out exhaustive experiments over big real world datasets to assess the performance of CloPN. The results show that our algorithm is very efficient in CFI mining from large real world datasets with up to 53 million articles.

7.6.4. Chiaroscuro

Participants: Tristan Allard, Florent Masseglia, Esther Pacitti.

The advent of on-body/at-home sensors connected to personal devices leads to the generation of fine grain highly sensitive personal data at an unprecendent rate. However, despite the promises of large scale analytics there are obvious privacy concerns that prevent individuals to share their personnal data. In [30], we propose Chiaroscuro, a complete solution for clustering personal data with strong privacy guarantees. The execution sequence produced by Chiaroscuro is massively distributed on personal devices, coping with arbitrary connections and disconnections. Chiaroscuro builds on our novel data structure, called Diptych, which allows the participating devices to collaborate privately by combining encryption with differential privacy. Our solution yields a high clustering quality while minimizing the impact of the differentially private perturbation. Our study show that Chiaroscuro is both correct and secure.

7.6.5. Large-scale Recognition of Visual and Audio Entities

Participants: Valentin Leveau, Alexis Joly, Patrick Valduriez.

We improved our work on the retrieval of visual identities by introducing a supervised classification layer on top of the large-scale instance-based matching layer. We introduce a new match kernel based on the inverse rank of the Shared Nearest Neighbors (SNN) combined with local geometric constraints [40]. To avoid overfiting and reduce processing costs, the dimensionality of the resulting over-complete representation is further reduced by hierarchically pooling the raw consistent matches according to their spatial position in the training images. The final image representation is obtained by concatenating the resulting feature vectors at several resolutions. Learning from these representations using a logistic regression classifier is shown to provide excellent fine-grained classification performance. In [38], we transpose our new SNN match kernel to the case of audio contents (applied to bird sounds recognition). Thus, the spatial pooling of geometrically consistent visual matches is replaced by a temporal pooling of temporally consistent audio matches. The resulting classification system obtained the second best results at the LifeCLEF bird identification challenge 2015 [36], the largest challenge of this kind ever organized (1000 bird species, 33K audio recordings).

7.6.6. Crowd-sourced Biodiversity Data Production through Pl@ntNet

Participants: Alexis Joly, Julien Champ, Jean-Christophe Lombardo, Antoine Affouard.

Initiated in the context of a citizen sciences project with botanists of the AMAP laboratory and the Tela Botanica social network, Pl@ntNet [18] is an innovative collaborative plateform focused on image-based plant identification as a mean to enlist new contributors and boost the production of biodiversity data and knowledge. Since 2010, several hundreds of thousands of geo-tagged and dated plant photographs were collected and revised by tens of thousands of novice, amateur and expert botanists. A content-based identification tool, available as both web and mobile applications, is synchronized with the growing data and allows any user to query or enrich the system with new observations. As a concrete new result, the cumulative number of downloads of the iPhone or Android app did reach 1M in October 2015. One of the main novelty in 2015 was the introduction of deep learning technologies in order to improve classification performance as well as the quality and speed of the content-based image retrieval.

A comparative study that we conducted in the context of the LifeCLEF [1] plant identification challenge did actually confirm that deep convolutional neural networks definitely outperforms the best fine-grained classification models on the aggregation of hand-crafted visual features [33]. Thus, we integrated this technology in the Pl@ntNet platform and exploited it in two ways: (i) for extracting more relevant (local and global) visual features to be indexed and searched within our efficient content-based indexing and retrieval framework (SnoopIm software) (ii) for reranking the species returned by the content-based search engine so as to increase the average reciprocal rank of the correct species while keeping a good level of interpretability of the returned results.

7.6.7. Crowd-sourced Biodiversity Data Production through LifeCLE

Participants: Alexis Joly, Julien Champ, Jean-Christophe Lombardo, Antoine Affouard.

We continued sharing the data produced by the Pl@ntNet platform with the international research community through the animation of the LifeCLEF research platform and the set-up of three new challenges, one related to plant images, one to bird sounds and one to fish videos. More than 200 research groups registered to at least one of the challenges and about 20 of them crossed the finish lines by running their system on the final test data. A synthesis of the results is published in the LifeCLEF 2015 overview paper [37] and more detailed analyses are provided in technical reports for the plant task [35] and the bird task [36]. We also report on an experimental study aimed at evaluating how state-of-art computer vision systems perform in identifying plants compared to human expertise [15],. A subset of the evaluation dataset used within LifeCLEF 2014 plant identification challenge was shared with volunteers of diverse expertise, ranging from leading experts of the targeted flora to inexperienced test subjects. In total, 16 human runs were collected and evaluated comparatively to the 27 machine-based runs of LifeCLEF challenge. The main outcome of the experiment was that machines are still far from outperforming the best expert botanists but they are clearly competing with some experienced botanists specialists of other floras.

8. Bilateral Contracts and Grants with Industry

8.1. Microsoft (2013-2017)

Participants: Ji Liu, Esther Pacitti, Patrick Valduriez.

This joint project is on advanced data storage and processing for cloud workflows with the Kerdata team in the context of the Joint Inria – Microsoft Research Centre. The project addresses the problem of advanced data storage and processing for supporting scientific workflows in the cloud. The goal is to design and implement a framework for the efficient processing of scientific workflows in clouds. The validation h will be performed using synthetic benchmarks and real-life applications from bioinformatics: first on the Grid5000 platform in a preliminary phase, then on the Microsoft Azure cloud environment.

8.2. Triton I-lab (2014-2016)

Participants: David Fernandez, Houssem-Eddine Chihoud, Didier Parigot.

Triton is a new common lab. (i-lab) created between Zenith and Beepeers (http://beepeers.com) to work on a platform for developing social networks in mobile/Web environments. The main objective of this project is to design and implement a new architecture for beepeers applications to move to the scale. This new architecture will build on our SON middleware and new NoSQL database technologies, especially graph databases.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. Labex NUMEV, Montpellier

URL: http://www.lirmm.fr/numev

[1] www.lifeclef.org

We are participating in the Laboratory of Excellence (labex) NUMEV (Digital and Hardware Solutions, Modelling for the Environment and Life Sciences) headed by University of Montpellier 2 in partnership with CNRS, University of Montpellier 1, and Inria. NUMEV seeks to harmonize the approaches of hard sciences and life and environmental sciences in order to pave the way for an emerging interdisciplinary group with an international profile. The NUMEV project is decomposed in four complementary research themes: Modeling, Algorithms and computation, Scientific data (processing, integration, security), Model-Systems and measurements. Florent Masseglia co-heads (with Pascal Poncelet) the theme on scientific data.

9.1.2. Institut de Biologie Computationnelle (IBC), Montpellier

URL: http://www.ibc-montpellier.fr

IBC is a 5 year project with a funding of 2Meuros by the MENRT (PIA program) to develop innovative methods and software to integrate and analyze biological data at large scale in health, agronomy and environment. Patrick Valduriez heads the workpackage on integration of biological data and knowledge.

9.2. National Initiatives

9.2.1. PIA (Projets Investissements d'Avenir

9.2.1.1. Datascale (2013-2015), 250Keuros

Participants: Reza Akbarinia, Florent Masseglia, Saber Salah, Patrick Valduriez.

The Datascale project is a PIA on big data with Bull (leader), CEA, ActiveEon SAS, Armadillo, Twenga, IPGP, Xedix and Inria (Zenith) . The goal of the project is to develop the essential technologies for big data, including efficient data management, software architecture and database architecture, and demonstrate their scalability with representative applications. In this project, the Zenith team works on data mining with Hadoop MapReduce.

9.2.1.2. Xdata (2013-2015), 125Keuros

Participants: Julien Diener, Patrick Valduriez.

The X-data project is a PIA with Data Publica (leader), Orange, La Poste, EDF, Cinequant, Hurence and Inria (Indes, Planete and Zenith) . The goal of the project is to develop a big data plaftform with various tools and services to integrate open data and partners's private data for analyzing the location, density and consuming of individuals and organizations in terms of energy and services. In this project, the Zenith team heads the workpackage on data integration.

9.2.1.3. PIA Floris'Tic (2015-2018), 430Keuro.

Participants: Julien Champ, Alexis Joly.

Floris'tic is a PIA aimed at promoting the scientific and technical culture of plant sciences through innovative pedagogic methods, including participatory initiatives and the use of IT tools such as the one built within the Pl@ntNet project. A. Joly heads the work package on the development of the IT tools. This is a joint project with the AMAP laboratory and the TelaBotanica social network.

9.2.2. Others

9.2.2.1. CIFRE INA/Inria (2013-2016), 100Keuros

Participants: Alexis Joly, Valentin Leveau, Patrick Valduriez.

This CIFRE contract with INA allows funding a 3-years PhD (Valentin Leveau). This PhD addresses research challenges related to large-scale supervised content-based retrieval in distributed environments.

9.2.2.2. CNRS INS2I Mastodons (2013-2015), 90Keuros

Participants: Alexis Joly, Florent Masseglia, Esther Pacitti [leader], Patrick Valduriez.

This project deals with the problems of big data in the context of life science, where masses of data are being produced, e.g. by Next Generation Sequencing technologies or plant phenotyping platforms. In this project, Zenith addresses the specific problems of large-scale data analysis and data sharing.

9.3. European Initiatives

9.3.1. FP7 Projects

9.3.1.1. CoherentPaaS

Participants: Carlyna Bondiombouy, Boyan Kolev, Oleksandra Levchenko, Patrick Valduriez.

Project title: A Coherent and Rich Platform as a Service with a Common Programming Model
Instrument: Integrated Project
Duration: 2013 - 2016
Total funding: 5 Meuros (Zenith: 500Keuros)
Coordinator: U. Madrid, Spain
Partner: FORTH (Greece), ICCS (Greece), INESC (Portugal) and the companies MonetDB (Netherlands), QuartetFS (France), Sparsity (Spain), Neurocom (Greece), Portugal Telecom (Portugal).
Inria contact: Patrick Valduriez

Accessing and managing large amounts of data is becoming a major obstacle to developing new cloud applications and services with correct semantics, requiring tremendous programming effort and expertise. CoherentPaaS addresses this issue in the cloud PaaS landscape by developing a PaaS that incorporates a rich and diverse set of cloud data management technologies, including NoSQL data stores, such as key-value data stores and graph databases, SQL data stores, such as in-memory and column-oriented databases, hybrid systems, such as SQL engines on top on key-value data stores, and complex event processing data management systems. It uses a common query language to unify the programming models of all systems under a single paradigm and provides holistic coherence across data stores using a scalable, transactional management system. CoherentPaaS will dramatically reduce the effort required to build and the quality of the resulting cloud applications using multiple cloud data management technologies via a single query language, a uniform programming model, and ACID-based global transactional semantics. CoherentPaaS will design and build a working prototype and will validate the proposed technology with real-life use cases. In this project, Zenith is in charge of designing the CloudMdsQL language and implementing its compiler/optimizer and query engine.

9.3.1.2. HPC4E

Participants: Reza Akbarinia, Florent Masseglia, Esther Pacitti, Patrick Valduriez.

Project title: High Performance Computing for Energy
Instrument: H2020
Duration: 2015 - 2017
Total funding: 2 Meuros
Coordinator: Barcelona Supercomputing Center (BSC), Spain
Partner: Europe: Inria, Lancaster University, Centro de Investigaciones Energéticas Medioambientales y Tecnológicas, Repsol S.A., Iberdrola Renovables Energía S.A., Total S.A. Brazil: COPPE/Universidade Federal de Rio de Janeiro, LNCC, Instituto Tecnológico de Aeronáutica (ITA), Universidade Federal do Rio Grande do Sul, Universidade Federal de Pernambuco, PETROBRAS.
Inria contact: Patrick Valduriez

The main objective is to develop beyond-the-state-of-the-art high performance simulation tools that can help the energy industry to respond future energy demands and also to carbon-related environmental issues using the state-of-the-art HPC systems. The project also aims at improving the usage of energy using HPC tools by acting at many levels of the energy chain for different energy sources. Another objective is to improve the cooperation between energy industries from EU and Brazil. The project includes relevant energy industral partners from Brazil (PETROBRAS) and EU (REPSOL and TOTAL as O &G industries), which will benefit from the project's results. A last objective is to improve the cooperation between the leading research centres in EU and Brazil in HPC applied to energy industry. This includes sharing supercomputing infrastructures between Brazil and EU. The cross-fertilization between energy-related problems and other scientific fields will be beneficial at both sides of the Atlantic. In this project, Zenith is working on Big Data management and analysis of numerical simulations.

9.4. International Initiatives

9.4.1. Inria Associate Teams

9.4.1.1. MUSIC

Title: MUltiSite Cloud (MUSIC) data management
Inria principal investigator: Esther Pacitti
International Partner):

Laboratorio Nacional de Computaçao Cientifica, Petropolis (Brazil) - Fabio Porto

Universidade Federal do Rio de Janeiro (Brazil) - Alvaro Coutinho and Marta Mattoso

Universidade Federal Fluminense, Niteroi (Brazil) - Daniel Oliveira

Centro Federal de Educa cao Tecnologica, Rio de Janeiro (Brazil) - Eduardo Ogasawara

Duration: 2014 - 2016
See also: https://team.inria.fr/zenith/projects/international-projects/music/

The cloud has become a good match for managing big data since it provides unlimited computing, storage and network resources on demand. By centralizing all data in a large-scale data-center, the cloud significantly simplifies the task of system administration. But for scientific data, where different organizations may have their own data-centers, a distributed (multisite) cloud model where each site is visible from outside, is needed. The main objective of this research and scientific collaboration is to develop a multisite cloud architecture for managing and analyzing scientific data, including support for heterogeneous data; distributed scientific workflows, and complex big data analysis. The resulting architecture will enable scalable data management infrastructures that can be used to host a variety of scientific applications that benefit from computing, storage, and networking resources that span multiple data-centers.

9.4.1.2. BIGDATANET

Title: A hybrid P2P/cloud for big data
Inria principal investigator: Patrick Valduriez
International Partner : University of California at Santa Barbara (USA) - Amr El Abbadi and Divy Agrawal
Duration: 2013 - 2015
See also: https://team.inria.fr/zenith/projects/international-projects/bigdatanet/

The main objective of this research and scientific collaboration is to develop a hybrid architecture of a computational platform that leverages the cloud computing and the P2P computing paradigms. The resulting architecture will enable scalable data management and data analysis infrastructures that can be used to host a variety of next-generation applications that benefit from computing, storage, and networking resources that exist not only at the network core (i.e., data-centers) but also at the network edge (i.e., machines at the user level as well as machines available in CDNs – content distribution networks hosted in ISPs).

9.4.2. Inria International Partners

9.4.2.1. Informal International Partners

We have regular scientific relationships with research laboratories in

- North America: Univ. of Waterloo (Tamer Özsu).
- Asia: National Univ. of Singapore (Beng Chin Ooi, Stéphane Bressan), Wonkwang University, Korea (Kwangjin Park)
- Europe: Univ. of Amsterdam (Hamideh Afsarmanesh), Univ. of Madrid (Ricardo Jiménez-Periz), UPC Barcelona (Josep Lluis Larriba Pey), HES-SO (Henning Müller), University of Catania (Concetto Spampinatto), The Open University (Stefan Rüger)
- North Africa: Univ. of Tunis (Sadok Ben-Yahia)

9.4.3. Inria International Labs

The Bigdatanet associated team takes part of the Inria@SiliconValley lab.

9.4.4. Participation In other International Programs

We are involved in the following international actions:

- CNPq-Inria project Hoscar (HPC and data management, 2012-2015) with LNCC (Fabio Porto), UFC, UFRGS (Philippe Navaux), UFRJ (Alvaro Coutinho, Marta Mattoso) to work on data management in high performance computing environments.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

Marta Mattoso (UFRJ, Brazil) gave a seminar on “‘Exploratory Analysis of Raw Data Files through Dataflows” in March.

9.5.2. Visits to International Teams

Maximilien Servajean visited UCSB in june, in the context of the Bigdatanet associated team.

10. Dissemination

10.1. Scientific Animation

Participation in the editorial board of scientific journals:

- VLDB Journal: P. Valduriez.
- Journal of Transactions on Large Scale Data and Knowledge Centered Systems, R. Akbarinia.
- Distributed and Parallel Databases, Kluwer Academic Publishers: E. Pacitti, P. Valduriez.
- Internet and Databases: Web Information Systems, Kluwer Academic Publishers: P. Valduriez.
- Journal of Information and Data Management, Brazilian Computer Society Special Interest Group on Databases: P. Valduriez.
- Book series “Data Centric Systems and Applications” (Springer): P. Valduriez.
- Ingénierie des Systèmes d’Information, Hermès : P. Valduriez.
- Journal of Data Semantics (Springer): S. Cohen-Boulakia
- Multimedia Tools and Applications: A. Joly

Participation to the organization of conferences and workshops:

- Alexis Joly was the main chair of the LifeCLEF 2015 international workshop [2] dedicated to multimedia biodiversity data management (Toulouse, September 8-10, part of CLEF 2015 conference)
- Alexis Joly was in the organizing committee of the 2nd International Workshop on Environmental Multimedia Retrieval [3], in conjunction with ICMR 2015, 23 June, Shangai
- Alexis Joly was in the organizing committee of the Floris’tic national workshop hold during the "salon de l’écologie" [4], Montpellier, France, 5-7 November

[2] http://www.imageclef.org/lifeclef/2015
[3] http://mklab.iti.gr/emr2015/
[4] http://www.salon-ecologie.com/congres-ecolotech-montpellier/programme-du-congres/

Participation in conference program committees :

- ACM SIGMOD Conf. 2015: R. Akbarinia, S. Cohen-Boulakia, 2016: R. Akbarinia, S. Cohen-Boulakia
- IEEE Int. Conf. on Data Engineering (ICDE) 2015: S. Cohen-Boulakia, 2016: R. Akbarinia, E. Pacitti, P. Valduriez (area chair)
- ADBIS 2015 - East-European Conference on Advances in Databases and Information Systems: R. Akbarinia, P. Valduriez (PC chair)
- VLDB 2015 : P. Valduriez (sponsor co-chair, and best paper award committee)
- Int. Conf. on Extending DataBase Technologies (EDBT), 2015: E. Pacitti
- BPM 2015 (Business Process Management (BPM) 2015: S. Cohen-Boulakia
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: A. Joly
- ACM Multimedia conference (ACMMM), 2015: A. Joly
- International Conference and Labs of the Evaluation Forum (CLEF), 2015: A. Joly
- IEEE International Conference on Image Processing (ICIP), 2015: A. Joly
- ACM International Conference on Multimedia Retrieval (ICMR), 2015: A. Joly
- IEEE Int. Workshop on Environmental Acoustic Data Mining (EADM), 2015: A. Joly
- Workshop on Theory and Practice of Provenance (TAPP) 2014 et 2015: S. Cohen-Boulakia
- ICDT workshop on Algorithms and Systems for MapReduce and Beyond (BeyondMR) 2015: S. Cohen-Boulakia
- IEEE Int. Conf. on Data Mining, 2015: F. Masseglia (area chair)
- ACM Symposium on Applied Computing, 2015: F. Masseglia
- Int. Conf. on Data Management Technologies and Applications, 2015: F. Masseglia
- Pacific-Asia Conf. on Knowledge Discovery and Data Mining, 2015: F. Masseglia
- European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2015: F. Masseglia
- IEEE Int. Conf. on Data Science and Advanced Analytics, 2015: F. Masseglia
- Workshop on Big Data and Data Mining Challenges on IoT and Pervasive Systems, 2015: E. Pacitti
- 9th BRESCI - Brazilian e-Science Workshop, 2015: E. Pacitti

Reviewing in international journals :

- Distributed and Parallel Databases: R. Akbarinia
- ACM Transactions on Database Systems (TODS): A. Joly
- IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI): A. Joly
- IEEE Transactions on Knowledge and Data Engineering: R. Akbarinia
- Information Sciences: A. Joly
- Ecological Informatics: A. Joly
- Multimedia Tools and Applications Journal (MTAP): A. Joly
- Multimedia Systems: A. Joly
- Transactions on Information Forensics & Security: A. Joly
- International Journal of Computer Vision: A. Joly
- Transactions on Image Processing: A. Joly
- ACM Trans. on Database Systems: E. Pacitti
- Knowledge and Information Systems (KAIS): F. Masseglia
- Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS): F. Masseglia
- IEEE Transactions on Knowledge and Data Engineering (TKDE): F. Masseglia

Other activities (national):

- Zenith participated in the Hemera and Grid'5000 communities by: contributing to the definition of "Storage 5000" (we help defining the financial model for teams to have their data in Grid5000) ; participating to Hemera meetings (Florent Masseglia gave a talk on "Pattern Mining from Big Data, meeting of Sept. 18, Nantes); and contributing to the prospective document of Grid5000.
- P. Valduriez is the scientific manager for the Latin America zone at Inria Direction des Relations Internationales (DRI), Member of the Scientific Committee at Agence Nationale de la Recherche (ANR) - Défi 7 Information and communication society and Member of the Scientific Committee of the BDA conference.
- A. Joly gave several invited talks in national events: *industrial computer vision days 2015* [5], Montpellier, 15 October, *ICAR-AMAP images days 2015* [6], September 17 He defended his « Habilitation à Diriger les Recherches (HDR) » entitled *Large-scale Content-based Information Retrieval* in June.
- F. Masseglia gave an invited talk at the CNRS national seminar on IST about "Publication Data Analytics" (Meudon, 10 November), and an invited talk to the DSI service of IRD on "Scientific Data Mining" (Montpellier, 18 September). Florent also participates in the Class'Code PIA project dedicated to teaching computational thinking for professionals of education (head of the working group on the definition of fundamental notions).
- S. Cohen-Boulakia gave invited tutorials at the DigiCosme Spring School 2015 on Data Management (Saclay, May 2015) and at the summer school Cumulo-Numbio : Cloud computing for the life sciences (Aussois, June 2015).

Other activities (international):

- P. Valduriez gave several invited talks in international events: "Data-intensive HPC: opportunities and challenges", Workshop on Big Data and Extreme-scale Computing (BDEC), Barcelona, january 2015; "Cloud & big data: opportunities and risks for developing countries" at AFRICOMM 2015, Cotonou (Bénin), december 2015; "Big Data Management in Zenith" at Fundação Getulio Vargas, Rio de Janeiro, december 2015.
- A. Joly gave an invited talk at the ERMITES 2015 international summer school [7], 21-22 April 2015. He is the main chair of the LifeCLEF research platform and was the co-organizer of two the challenges run within the lab in 2015 (the plant identification challenge and the bird identification challenge).
- S. Cohen-Boulakia gave invited talks at Universities of Pennsylvania, April, and Humboldt, Berlin, December.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Most permanent members of Zenith teach at the Licence and Master degree levels at UM2.

Reza Akbarinia:

Master Research: New approaches for data storage, 9h, level M2, Faculty of Science, UM

Licence: Computing Tools, 36h, Level L3, Faculty of Science, UM

Florent Masseglia:

Science Popularization: 4 Ph.D students, from 3 different doctoral schools are having a 30h doctoral module under Florent Masseglia's supervision.

[5] http://www.captronic.fr/Image-vision-industrielle.html
[6] http://amap-collaboratif.cirad.fr/ecipp/?page_id=901
[7] http://glotin.univ-tln.fr/ERMITES15/

Esther Pacitti:

IG3: Database design, physical organization, 54h, level L3, Polytech'Montpellier, UM2

IG4: Networks, 42h, level M1, Polytech' Montpellier, UM2

IG4: Object-relational databases, 32h, level M1, Polytech' Montpellier, UM2

IG5: Distributed systems, virtualization, 27h, level M2, Polytech' Montpellier, UM2

Industry internship committee, 50h, level M2, Polytech' Montpellier

Master Research: Large scale data management, 4,5h, level M2, Faculty of Science, UM2

Didier Parigot:

Master Research: Large scale data management, 6h, level M2, Faculty of Science, UM2

Patrick Valduriez:

Master Research: Large scale data management, 12h, level M2, Faculty of Science, UM2

Professional: Distributed Information Systems, 50h, level M2, Capgemini Institut

Professional: XML, 30h, level M2, Orsys Formation

Alexis Joly:

Master Research: Large-scale Content-based Visual Information Retrieval, 3h, level M2, Faculty of Science, UM2

10.2.2. Supervision

- PhD in progress: Mehdi Zitouni Closed Pattern Mining in a Massively Distributed Environment started Sept. 2014, Univ. Tunis, Advisor: Florent Masseglia, co-advisor: Reza Akbarinia
- PhD in progress : Ji Liu, Scientific Worflows in Multisite Cloud, started oct. 2013, Univ. Montpellier, Advisors: Esther Pacitti and Patrick Valduriez
- PhD in progress : Saber Salah, Optimizing a Cloud for Data Mining Primitives, started nov. 2012, Univ. Montpellie, Advisor: Florent Masseglia, co-advisor: Reza Akbarinia
- PhD in progress : Valentin Leveau, Supervised content-based information retrieval in big multimedia data, started April 2013, Univ. Montpellier, Advisor: Patrick Valduriez, co-advisor: Alexis Joly and Olivier Buisson
- PhD in progress : Djamel-Edine Yagoubi, Indexing Time Series in a Massively Distributed Environment, started October 2014, Univ. Montpellier, Advisors: Florent Masseglia and Patrick Valduriez, co-advisor: Reza Akbarinia
- PhD in progress : Sakina Mahboubi, Privacy Preserving Query Processing in Clouds, started October 2015, Univ. Montpellier, Advisor: Patrick Valduriez, co-advisor: Reza Akbarinia

10.2.3. Juries

Members of the team participated to the following Ph.D. committees:

- E. Pacitti: Sébastien Monnet (HDR, Univ. Paris 6), Luis Daniel Ibanez (Univ. Nantes), Maeva Antoine (Univ. Nice), Julien Lacroix (Univ. Aix-Marseille).
- P. Valduriez: Gerson Sunyé (HDR, Univ. Nantes, chair), Amin Mesmoudi (Univ. Lyon 1) Stamatis Zampetakis (Univ. Paris Sud).

10.3. Popularization

Today, one of the main questions in science popularization and computer science teaching at school in France is: "how to scale?". Some of our recent actions in the domaine are mainly oriented towards this goal of scaling, in particular with the Class'Code PIA project (https://pixees.fr/?page_id=1980, F. Masseglia is member of the project).

F. Masseglia has coordinated a national network of colleagues for promoting code learning. After several training sessions given to professionals of education F. Masseglia has written a complete guide for setting up a similar training session (https://pixees.fr/?p=5713).

"La main à la pâte" is leading the writing of a school book on computer science teaching involving Inria (Gilles Dowek, Pierre-Yves Oudeyer, Florent Masseglia and Didier Roy), "France-IOI" and the University of Lorraine. Florent Masseglia has tested the pedagogical sequences of the book in a classroom near Montpellier.

Zenith participated to the following events in Montpellier:

- F. Masseglia co-organized and co-animated the Inria's stand at "La fête de la science" (Montpellier), held by Genopolys (a science village).
- F. Masseglia organized and animated several "Kid&Code" studios in the greater metropolitan area of Montpellier, involving the network of media libraries and the network of extracurricular activities. He also has proposed a two sessions training to activity leaders of "Les petits débrouillards". The goal is to accompany the activity leaders and help them until they are autonomous
- F. Masseglia is member of the project selection committee for "La fête de la science" in Montpellier
- F. Masseglia is co-author of a paper in the Scratch conference on code learning for kids [50]
- M. Servajean and A. Joly co-animated a stand at "La fête de la science" (Montpellier), held by the LIRMM laboratory
- M. Servajean, A. Joly and J. Champ collaborated to a particatory workshop at "Le salon de l'écologie", Montpellier

F. Masseglia is scientific vice-editor of interstices (https://interstices.info), and a member of the scientific committee for the edition of "Datagramme", the game from Inria on science popularization. He has participated to the two days training sessions on "science popularization" given by Claude Vadel in Paris, on Dec. 4 and 5, 2014.

A. Joly and J. Champ participated to the set-up of a Pl@ntNet demo within the French pavillon at the Universal Exposition hold in Milan (about 2M visitors on the French pavillon).

As a member of the organizing committee of the Floris'tic project, A. Joly participated to several popularization and educational actions in collaboration with Tela Botanica NGO (cities, parks, schools, etc.)

F. Masseglia is a member of the management board of "Les Petits Débrouillards" in Languedoc-Roussillon. He also is the scientific responsible for schools visits in the Lirmm Laboratory.

D. Shasha gave several talks: "the changing nature of invention in computer science" (Morgenstern seminar) and "Group Testing to Describe Causality in Gene Networks" at Inria Sophia-Antipolis in april, and "Statistics is easy" at IBC, Montpellier in march.

P. Valduriez gave a talk on "Integrating Big Data and Relational Data with CloudMdsQL" at the DGA seminar "Traitement de l'information multimodale et Big Data", Arcueil, in october.

11. Bibliography

Major publications by the team in recent years

[1] R. AKBARINIA, F. MASSEGLIA. *Fast and Exact Mining of Probabilistic Data Streams*, in "PKDD'2013: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases", Prague, Czech Republic, Lecture Notes in Computer Science, Springer, 2013, pp. 493-508 [*DOI :* 10.1007/978-3-642-40988-2_32], http://hal.inria.fr/lirmm-00838618

[2] R. AKBARINIA, P. VALDURIEZ, G. VERGER. *Efficient Evaluation of SUM Queries Over Probabilistic Data*, in "IEEE Transactions on Knowledge and Data Engineering", 2013, vol. 25, n° 4, pp. 764-775, http://hal.inria.fr/lirmm-00652293

[3] T. ALLARD, G. HÉBRAIL, F. MASSEGLIA, E. PACITTI. *Chiaroscuro: Transparency and Privacy for Massive Personal Time-Series Clustering*, in "34th International ACM Conference on Management of Data (ACM SIGMOD)", Melbourne, Australia, A. SIGMOD (editor), May 2015 [*DOI :* 10.1145/2723372.2749453], https://hal.inria.fr/hal-01136686

[4] N. AYAT, R. AKBARINIA, H. AFSARMANESH, P. VALDURIEZ. *Entity Resolution for Distributed Probabilistic Data*, in "Distributed and Parallel Databases", 2013, vol. 31, n° 4, pp. 509-542, http://hal.inria.fr/lirmm-00879631

[5] A. JOLY, H. GOEAU, P. BONNET, V. BAKIC, J. BARBE, S. SELMI, I. YAHIAOUI, J. CARRÉ, E. MOUYSSET, J.-F. MOLINO, N. BOUJEMAA, D. BARTHÉLÉMY. *Interactive plant identification based on social image data*, in "Ecological Informatics", 2013 [*DOI :* 10.1016/J.ECOINF.2013.07.006], http://www.sciencedirect.com/science/article/pii/S157495411300071X

[6] B. KOLEV, P. VALDURIEZ, C. BONDIOMBOUY, R. JIMENEZ-PERIS, R. PAU, J. O. PEREIRA. *CloudMdsQL: Querying Heterogeneous Cloud Data Stores with a Common Language*, in "Distributed and Parallel Databases", 2015, 41 p. , To appear, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01184016

[7] P. LETESSIER, O. BUISSON, A. JOLY. *Scalable Mining of Small Visual Objects*, in "Proceedings of the 20th ACM International Conference on Multimedia", New York, NY, USA, MM '12, ACM, 2012, pp. 599–608, http://doi.acm.org/10.1145/2393347.2393431

[8] E. OGASAWARA, D. DE OLIVEIRA, P. VALDURIEZ, D. DIAS, F. PORTO, M. MATTOSO. *An Algebraic Approach for Data-Centric Scientific Workflows*, in "Proceedings of VLDB", 2011, vol. 4, n° 11, pp. 1328-1339, http://hal.inria.fr/hal-00640431/en

[9] E. PACITTI, R. AKBARINIA, M. EL DICK. *P2P Techniques for Decentralized Applications*, Morgan & Claypool Publishers, 2012, 104 p. , http://hal.inria.fr/lirmm-00748635

[10] S. SALAH, R. AKBARINIA, F. MASSEGLIA. *Fast Parallel Mining of Maximally Informative k-Itemsets in Big Data*, in "IEEE International Conference on Data Mining (ICDM)", Atlantic city, United States, August 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01187275

[11] T. M. ÖZSU, P. VALDURIEZ. *Principles of Distributed Database Systems, third edition*, Springer, 2011, 845 p. , http://hal.inria.fr/hal-00640392/en

Publications of the year

Doctoral Dissertations and Habilitation Theses

[12] S. COHEN-BOULAKIA. *Data Integration in the Life Sciences: Scientific Workflows, Provenance, and Ranking*, Université Paris-Sud, June 2015, Habilitation à diriger des recherches, https://hal.archives-ouvertes.fr/tel-01245229

[13] A. JOLY. *Large-scale Content-based Visual Information Retrieval*, Université de Montpellier, May 2015, Habilitation à diriger des recherches, https://hal.inria.fr/hal-01182797

Articles in International Peer-Reviewed Journals

[14] M. R. BOUADJENEK, H. HACID, M. BOUZEGHOUB. *Social Networks and Information Retrieval, How Are They Converging? A Survey, a Taxonomy and an Analysis of Social Information Retrieval Approaches and Platforms*, in "Information Systems", March 2016, vol. 56, pp. 1-18 [*DOI :* 10.1016/J.IS.2015.07.008], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01174843

[15] P. BONNET, A. JOLY, H. GOËAU, J. CHAMP, C. VIGNAU, J.-F. MOLINO, D. BARTHÉLÉMY, N. BOUJEMAA. *Plant identification: Man vs. Machine*, in "Multimedia Tools and Applications", June 2015, 19 p. [*DOI :* 10.1007/S11042-015-2607-4], https://hal.inria.fr/hal-01182778

[16] B. BRANCOTTE, B. YANG, G. BLIN, S. COHEN-BOULAKIA, A. DENISE, S. HAMEL. *Rank aggregation with ties: Experiments and Analysis*, in "Proceedings of the VLDB Endowment (PVLDB)", August 2015, 12 p. , https://hal.archives-ouvertes.fr/hal-01165336

[17] J. DIAS, G. GUERRA, F. ROCHINHA, A. COUTINHO, P. VALDURIEZ, M. MATTOSO. *Data-Centric Iteration in Dynamic Workflows*, in "Future Generation Computer Systems", 2015, vol. 46, pp. 114-126 [*DOI :* 10.1016/J.FUTURE.2014.10.021], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01073638

[18] A. JOLY, P. BONNET, H. GOËAU, J. BARBE, S. SELMI, J. CHAMP, S. DUFOUR-KOWALSKI, A. AFFOUARD, J. CARRÉ, J.-F. MOLINO, N. BOUJEMAA, D. BARTHÉLÉMY. *A look inside the Pl@ntNet experience*, in "Multimedia Systems", 2015, 16 p. [*DOI :* 10.1007/S00530-015-0462-9], https://hal.inria.fr/hal-01182775

[19] B. KOLEV, P. VALDURIEZ, C. BONDIOMBOUY, R. JIMENEZ-PERIS, R. PAU, J. O. PEREIRA. *CloudMdsQL: Querying Heterogeneous Cloud Data Stores with a Common Language*, in "Distributed and Parallel Databases", 2015, pp. 1-41, forthcoming [*DOI :* 10.1007/S10619-015-7185-Y], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01184016

[20] M. LIROZ-GISTAU, R. AKBARINIA, P. VALDURIEZ. *FP-Hadoop: Efficient Execution of Parallel Jobs Over Skewed Data*, in "Proceedings of the VLDB Endowment (PVLDB)", 2015, vol. 8, n^{o} 12, pp. 1856-1867, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01162362

[21] J. LIU, E. PACITTI, P. VALDURIEZ, M. MATTOSO. *A Survey of Data-Intensive Scientific Workflow Management*, in "The Journal of Grid Computing", 2015, vol. 13, 44 p. [*DOI :* 10.1007/S10723-015-9329-8], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01144760

[22] M. SERVAJEAN, R. AKBARINIA, E. PACITTI, S. AMER-YAHIA. *Profile Diversity for Query Processing using User Recommendations*, in "Information Systems", March 2015, vol. 48, pp. 44-63 [*DOI :* 10.1016/J.IS.2014.09.001], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01079523

[23] M. SERVAJEAN, E. PACITTI, M. LIROZ-GISTAU, S. AMER-YAHIA, A. EL ABBADI. *Increasing Coverage in Distributed Search and Recommendation with Profile Diversity*, in "Transactions on Large-Scale Data- and Knowledge-Centered Systems", August 2015, vol. LNCS, forthcoming, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01177817

[24] V. SILVA SOUZA, O. DANIEL DE, P. VALDURIEZ, M. MATTOSO. *Analyzing Related Raw Data Files through Dataflows*, in "Concurrency and Computation: Practice and Experience", 2015, 16 p. , forthcoming, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01181231

[25] J. STARLINGER, S. COHEN-BOULAKIA, S. KHANNA, S. DAVIDSON, U. LESER. *Effective and Efficient Similarity Search in Scientific Workflow Repositories*, in "Future Generation Computer Systems", September 2015, 79 p. [*DOI :* 10.1016/J.FUTURE.2015.06.012], https://hal.archives-ouvertes.fr/hal-01170597

Articles in National Peer-Reviewed Journals

[26] S. COHEN-BOULAKIA, P. VALDURIEZ. *Querying and Managing bioinformatic data for molecular biology*, in "Les Techniques de l'Ingenieur", June 2015, 30 p. , https://hal.inria.fr/hal-01160897

Invited Conferences

[27] P. VALDURIEZ. *Data-intensive HPC: opportunities and challenges*, in "BDEC'2015: Big Data and Extreme-scale Computing", Barcelone, Spain, Barcelona Supercomputing Center, January 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01184018

International Conferences with Proceedings

[28] R. AKBARINIA, M. LIROZ-GISTAU, D. AGRAWAL, P. VALDURIEZ. *An Efficient Solution for Processing Skewed MapReduce Jobs*, in "Globe'2015: 8th International Conference on Data Management in Cloud, Grid and P2P Systems", Valencia, Spain, September 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01162359

[29] R. AKBARINIA, F. MASSEGLIA. *Aggregation-Aware Compression of Probabilistic Streaming Time Series*, in "MLDM'2015: International Conference on Machine Learning and Data Mining", Hamburg, Germany, July 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01162366

[30] T. ALLARD, G. HÉBRAIL, F. MASSEGLIA, E. PACITTI. *Chiaroscuro: Transparency and Privacy for Massive Personal Time-Series Clustering*, in "SIGMOD: Conference on Management of Data", Melbourne, Australia, A. SIGMOD (editor), May 2015 [*DOI :* 10.1145/2723372.2749453], https://hal.inria.fr/hal-01136686

[31] C. BONDIOMBOUY. *Query Processing in Cloud Multistore Systems*, in "BDA'2015 : Gestion de données – principes, technologies et applications", Île de Porquerolles, France, September 2015, forthcoming, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01181253

[32] C. BONDIOMBOUY, B. KOLEV, O. LEVCHENKO, P. VALDURIEZ. *Integrating Big Data and Relational Data with a Functional SQL-like Query Language*, in "DEXA'2015: 26th International Conference on Database and Expert Systems Applications", Valencia, Spain, Q. CHEN, A. HAMEURLAIN, F. TOUMANI, R. WAGNER, H. DECKER (editors), September 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01181242

[33] J. CHAMP, T. LORIEUL, M. SERVAJEAN, A. JOLY. *A comparative study of fine-grained classification methods in the context of the LifeCLEF plant identification challenge 2015*, in "CLEF 2015", Toulouse, France, CEUR-WS (editor), September 2015, vol. 1391, https://hal.inria.fr/hal-01182788

[34] M. GOLESTAN FAR, S. SANNER, M. REDA BOUADJENEK, G. FERRARO, D. HAWKING. *On Term Selection Techniques for Patent Prior Art Search*, in "SIGIR'15: 38th International Conference on Research and Development in Information Retrieval", Santiago, Chile, ACM, August 2015 [*DOI :* 10.1145/2766462.2767801], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01163055

[35] H. GOËAU, P. BONNET, A. JOLY. *LifeCLEF Plant Identification Task 2015*, in "CLEF 2015", Toulouse, France, CEUR-WS (editor), September 2015, vol. 1391, https://hal.inria.fr/hal-01182795

[36] H. GOËAU, H. GLOTIN, W.-P. VELLINGA, R. PLANQUÉ, A. RAUBER, A. JOLY. *LifeCLEF Bird Identification Task 2015*, in "CLEF 2015", Toulouse, France, CEUR-WS (editor), September 2015, vol. 1391, https://hal.inria.fr/hal-01182796

[37] A. JOLY, H. GOËAU, H. GLOTIN, C. SPAMPINATO, P. BONNET, W.-P. VELLINGA, R. PLANQUÉ, A. RAUBER, S. PALAZZO, B. FISHER, H. MÜLLER. *LifeCLEF 2015: Multimedia Life Species Identification Challenges*, in "CLEF 2015", Toulouse, France, September 2015, https://hal.inria.fr/hal-01182782

[38] A. JOLY, V. LEVEAU, J. CHAMP, O. BUISSON. *Shared nearest neighbors match kernel for bird songs identification -LifeCLEF 2015 challenge*, in "CLEF 2015", Toulouse, France, CEUR-WS (editor), September 2015, vol. 1391, https://hal.inria.fr/hal-01182784

[39] P. LETESSIER, N. HERVÉ, A. JOLY, H. NABI, M. DERVAL, O. BUISSON. *DigInPix: Visual Named-Entities Identification in Images and Videos*, in "5th ACM on International Conference on Multimedia Retrieval - ICMR '15", Shanghai, China, ACM, June 2015, pp. 661-664 [*DOI :* 10.1145/2671188.2749369], https://hal.inria.fr/hal-01182780

[40] V. LEVEAU, A. JOLY, O. BUISSON, P. VALDURIEZ. *Kernelizing Spatially Consistent Visual Matches for Fine-Grained Classification*, in "International Conference on Multimedia Retrieval 2015", Shangai, China, June 2015, https://hal.inria.fr/hal-01145988

[41] H. LUSTOSA, F. PORTO, R. COSTA, P. BLANCO, P. VALDURIEZ. *Managing Simulation Data with Multidimensional Arrays*, in "SBBD'2015: Simpósio Brasileiro de Banco de Dados", Petropolis, Brazil, Laboratrio Nacional de Computao Cientfica (LNCC) and Centro Federal de Educao Tecnolgica Celso Suckow da Fonseca (CEFET-RJ), Brazil, October 2015, 7 p. , http://hal-lirmm.ccsd.cnrs.fr/lirmm-01184265

[42] C. PRADAL, C. FOURNIER, P. VALDURIEZ, S. COHEN-BOULAKIA. *OpenAlea: Scientific Workflows Combining Data Analysis and Simulation*, in "SSDBM 2015: 27th International Conference on Scientific and Statistical Database Management", San Diego, United States, June 2015 [*DOI :* 10.1145/2791347.2791365], https://hal.archives-ouvertes.fr/hal-01166298

[43] M. REDA BOUADJENEK, S. SANNER, G. FERRARO. *A Study of Query Reformulation for Patent Prior Art Search with Partial Patent Applications*, in "ICAIL'2015: 15th International Conference on Artificial Intelligence and Law", San Diego, United States, June 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01134828

[44] S. SALAH, R. AKBARINIA, F. MASSEGLIA. *Data Partitioning for Fast Mining of Frequent Itemsets in Massively Distributed Environments*, in "DEXA'2015: 26th International Conference on Database and Expert Systems Applications", Valencia, Spain, September 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01169603

[45] S. SALAH, R. AKBARINIA, F. MASSEGLIA. *Fast Parallel Mining of Maximally Informative k-Itemsets in Big Data*, in "IEEE International Conference on Data Mining", Atlantic city, United States, August 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01187275

[46] S. SALAH, R. AKBARINIA, F. MASSEGLIA. *Optimizing the Data-Process Relationship for Fast Mining of Frequent Itemsets in MapReduce*, in "MLDM'2015: International Conference on Machine Learning and Data Mining", Hamburg, Germany, LNCS, July 2015, vol. 9166, pp. 217-231 [*DOI :* 10.1007/978-3-319-21024-7_15], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01171555

[47] A. VENKATESAN, N. E. HASSOUNI, F. PHILLIPE, C. POMMIER, H. QUESNEVILLE, M. RUIZ, P. LARMANDE. *Towards efficient data integration and knowledge management in the Agronomic domain*, in "APIA'15: premiere Conference Applications Pratiques de l'Intelligence Artificielle", Rennes, France, July 2015, https://hal.archives-ouvertes.fr/hal-01176903

[48] M. ZITOUNI, R. AKBARINIA, S. B. YAHIA, F. MASSEGLIA. *A Prime Number Based Approach for Closed Frequent Itemset Mining in Big Data*, in "DEXA'2015: 26th International Conference on Database and Expert Systems Applications", Valencia, Spain, September 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01169606

National Conferences with Proceedings

[49] C. JONQUET, E. DZALÉ-YEUMO, E. ARNAUD, P. LARMANDE. *AgroPortal : a proposition for ontology-based services in the agronomic domain*, in "IN-OVIVE'15: 3ème atelier INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement", Rennes, France, June 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01172232

Conferences without Proceedings

[50] M. DUFLOT, M. QUINSON, F. MASSEGLIA, D. ROY, J. VAUBOURG, T. VIÉVILLE. *When sharing computer science with everyone also helps avoiding digital prejudices*, in "Scratch2015AMS", Amsterdam, Netherlands, August 2015, https://hal.inria.fr/hal-01154767

Scientific Popularization

[51] T. VIEVILLE, S. BOLDO, F. MASSEGLIA, P. BERNHARD. *« Structures : organisation, complexité, dynamique » des mot-clés au sens inattendu*, April 2015, Article de vulgarisation sur pixees.fr, https://hal.inria.fr/hal-01238442

Other Publications

[52] B. BRANCOTTE, B. RANCE, A. DENISE, S. COHEN-BOULAKIA. *Interrogation de bases de données biologiques publiques par reformulation de requêtes et classement des résultats avec ConQuR-Bio*, July 2015, JOBIM (Journées Ouvertes Biologie Informatique Mathématiques), Poster, https://hal.archives-ouvertes.fr/hal-01167840