RESEARCH CENTRE

Inria Branch at the University of Montpellier

IN PARTNERSHIP WITH:

CNRS, Université de Montpellier

2023
ACTIVITY REPORT

Project-Team
ZENITH

Scientific Data Management

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

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Contents

Project-Team ZENITH

Creation of the Project-Team: 2012 January 01

Keywords

Computer sciences and digital sciences

A1.1. – Architectures

A3.1. – Data

A3.3. – Data and knowledge analysis

A3.4.4. – Optimization and learning

A5.4.3. – Content retrieval

A5.7. – Audio modeling and processing

A6.2.6. – Optimization

A9.2. – Machine learning

A9.3. – Signal analysis

Other research topics and application domains

B1.1.11. – Plant Biology

B2.6. – Biological and medical imaging

B3.3. – Geosciences

B3.5. – Agronomy

B3.6. – Ecology

B4. – Energy

B6. – IT and telecom

B6.5. – Information systems

1 Team members, visitors, external collaborators

Research Scientists

- Florent Masseglia [Team leader, INRIA, Senior Researcher, from Apr 2023, HDR]
- Reza Akbarinia [INRIA, Researcher, HDR]
- Christophe Botella [INRIA, ISFP, from Oct 2023]
- Benjamin Bourel [CNRS, Researcher, from Oct 2023]
- Alexis Joly [INRIA, Senior Researcher, HDR]
- Antoine Liutkus [Inria, HDR]
- Diego Marcos Gonzalez [INRIA, Advanced Research Position]
- Maxime Ryckewaert [INRIA, Starting Research Position, from Sep 2023]
- Patrick Valduriez [INRIA, Emeritus, from Apr 2023, HDR]

Faculty Members

- Esther Pacitti [UNIV MONTPELLIER, Professor, HDR]
- Joseph Salmon [UNIV MONTPELLIER, Professor, HDR]

Post-Doctoral Fellows

- Raphael De Freitas Saldanha [INRIA, Post-Doctoral Fellow]
- Benjamin Deneu [INRIA]
- Pallavi Jain [LIRMM, Post-Doctoral Fellow, from Jun 2023]
- Konstantinos-Panagiotis Panousis [INRIA, Post-Doctoral Fellow, from Apr 2023]
- Lukas Picek [INRIA, Post-Doctoral Fellow, from Nov 2023]
- Rebecca Pontes Salles [INRIA, Post-Doctoral Fellow, from Dec 2023]
- Jules Vandeputte [INRIA, Post-Doctoral Fellow, from Nov 2023]

PhD Students

- Ananthu Aniraj [INRIA, from Apr 2023]
- Matteo Contini [IFREMER]
- Guillaume Coulaud [UNIV MONTPELLIER, from Oct 2023]
- Joaquim Estopinan [Inria, INRIA]
- Camille Garcin [Univ Montpellier]
- Cesar Leblanc [INRIA]
- Tanguy Lefort [UNIV MONTPELLIER]
- Ousmane Youme [University Alioune DIOP de Bambey, , from Apr 2023 until Jul 2023]
- Kawtar Zaher [INA, CIFRE, from May 2023]

Technical Staff

- Antoine Affouard [INRIA, Engineer]
- Mathias Chouet [CIRAD, from Apr 2023]
- Baldwin Dumortier [University of Montpelliet, until Sep 2023]
- Maxime Fromholtz [INRIA, Engineer, from Feb 2023]
- Hugo Gresse [INRIA, Engineer, from Jun 2023]
- Benoit Lange [INRIA, Engineer, from Feb 2023]
- Théo Larcher [INRIA, Engineer]
- Pierre Leroy [INRIA, Engineer, from Oct 2023]
- Thomas Paillot [INRIA, from Sep 2023]
- Remi Palard [CIRAD, from Dec 2023]
- Julien Thomazo [LIRMM, from Dec 2023]

Administrative Assistant

- Cathy Desseaux [INRIA, from Sep 2023]

External Collaborators

- Hervé Goëau [CIRAD]
- François Munoz [UGA]
- Christophe Pradal [CIRAD]
- Dennis Shasha [New York University]

2 Overall objectives

Data-intensive science such as agronomy, astronomy, biology and environmental science must deal with overwhelming amounts of experimental data, as produced through empirical observation and simulation. Similarly, digital humanities have been faced for decades with the problem of exploiting vast amounts of digitized cultural and historical data, such as broadcasted radio or TV content. Such data must be processed (cleaned, transformed, analyzed) in order to draw new conclusions, prove scientific theories and eventually produce knowledge. However, constant progress in scientific observational instruments (e.g. satellites, sensors, large hadron collider), simulation tools (that foster in silico experimentation) or digitization of new content by archivists create a huge data overload. For example, climate modeling data has hundreds of exabytes.

Scientific data is very complex, in particular because of the heterogeneous methods, the uncertainty of the captured data, the inherently multiscale nature (spatial, temporal) of many sciences and the growing use of imaging (e.g. molecular imaging), resulting in data with hundreds of dimensions (attributes, features, etc.). Modern science research is also highly collaborative, involving scientists from different disciplines (e.g. biologists, soil scientists, and geologists working on an environmental project), in some cases from different organizations in different countries. Each discipline or organization tends to produce and manage its own data, in specific formats, with its own processes. Thus, integrating such distributed data gets difficult as the amounts of heterogeneous data grow. Finally, a major difficulty is to interpret scientific data. Unlike web data, e.g. web page keywords or user recommendations, which regular users can understand, making sense out of scientific data requires high expertise in the scientific domain.

Furthermore, interpretation errors can have highly negative consequences, e.g. deploying an oil driller under water at a wrong position.

Despite the variety of scientific data, we can identify common features: big data; manipulated through workflows; typically complex, e.g. multidimensional; 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).

The three main challenges of scientific data management can be summarized by: (1) scale (big data, big applications); (2) complexity (uncertain, high-dimensional data), (3) heterogeneity (in particular, data semantics heterogeneity). These challenges are also those of data science, with the goal of making sense out of data by combining data management, machine learning, statistics and other disciplines. 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, we strive to develop architectures, models and algorithms that can be implemented as components or services in specific computing environments, e.g. the cloud. We design and validate our solutions by working closely with our scientific partners in Montpellier such as CIRAD, INRAE and IRD, which provide the scientific expertise to interpret the data. To further validate our solutions and extend the scope of our results, we also foster industrial collaborations, even in non scientific applications, provided that they exhibit similar challenges.

Our approach is to capitalize on the principles of distributed and parallel data management. In particular, we exploit: high-level languages as the basis for data independence and automatic optimization; declarative languages to manipulate data and workflows; and highly distributed and parallel environments such as cluster and cloud for scalability and performance. We also exploit machine learning, probabilities and statistics for high-dimensional data processing, data analytics and data search.

3 Research program

3.1 Distributed 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 database systems, now at the heart of any information system, and of advanced data management capabilities in many kinds of software products (search engines, application servers, document systems, etc.).

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.

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). Transparency is achieved through a global schema which hides the local databases' heterogeneity. In its simplest form, a distributed database system 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.

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.

In contrast, peer-to-peer (P2P) systems 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. P2P systems typically 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. A P2P solution is well-suited 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, we believe cloud computing 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 best of both approaches. In particular, it will enable the clean integration of the users' own computational resources with different clouds.

3.2 Big Data and Parallel Data Management

Big data (like its relative, data science) 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.

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, connected objects (IoT), social networks, computer simulations, satellites, radiotelescopes, etc. Storage capacity has doubled every 3 years since 1980 with prices steadily going down, making it affordable to keep more data around. Furthermore, 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 such as: validity (is the data correct and accurate?); veracity (are the results meaningful?); volatility (how long do you need to store this data?).

Many different big data management solutions have been designed, primarily for the cloud, as cloud and big data are synergistic. They typically trade consistency for scalability, simplicity and flexibility, hence the new term Data-Intensive Scalable Computing (DISC). Examples of DISC systems include data processing frameworks (MapReduce, Spark, Pregel), file systems (GFS, HDFS), NoSQL systems (BigTable, Hbase, MongoDB), NewSQL systems (Spanner, CockroachDB, LeanXcale). In Zenith, we exploit or extend DISC technologies to fit our needs for scientific workflow management and scalable data analysis.

3.3 Data Integration

Scientists can rely on web tools to quickly share their data and/or knowledge. Therefore, when performing a given study, a scientist would typically need to access and integrate data from many data sources (including public databases). Data integration can be either physical or logical. In the former, the source data are integrated and materialized in a data warehouse or data lake. In logical integration, the integrated data are not materialized, but accessed indirectly through a global (or mediated) schema using a data integration system. These two approaches have different trade-offs, e.g. efficient analytics but only on historical data for data warehousing versus real-time access to data sources for data integration systems (e.g. web price comparators).

In both cases, 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 workflow systems are also useful for data integration and data analytics. They allow scientists to describe and execute complex scientific 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. This requires using distributed and parallel execution environments. However, existing workflow management systems have limited support for data parallelism. In Zenith, we use an algebraic approach to describe data-intensive workflows and exploit parallelism.

3.4 Data Analytics

Data analytics refers to a set of techniques to draw conclusions through data examination. It involves data mining, statistics, and data management, and is applied to categorical and continuous data. In the Zenith team, we are interested in both of these data types. Categorical data designates a set of data that can be described as “check boxes”. It can be names, products, items, towns, etc. A common illustration is the market basket data, where each item bought by a client is recorded and the set of items is the basket. The typical data mining problems with this kind of data are:

- **Frequent itemsets and association rules**. In this case, the data is usually a table with a high number of rows and the data mining algorithm extracts correlations between column values. 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 discovery but considering the order between. In the smart building example, a frequent sequence 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 $i+j+k$”.

- **Clustering**. The goal of clustering is to group together similar data while ensuring that dissimilar data will not be in the same cluster. In our example of smart buildings, we could find clusters of rooms, where offices will be in one category and copy machine rooms in another because of their differences (hours of people presence, number of times lights are turned on/off, etc.).

Continuous data are numeric records. A temperature value or a timestamp are examples of such data. They are involved in a widely used type of data known as time series: a series of values, ordered by time, and giving a measure, e.g. coming from a sensor. There is a large number of problems that can apply to this kind of data, including:

- **Indexing and retrieval.** The goal, here, is usually to find, given a query q and a time series dataset D, the records of D that are most similar to q. This may involve any transformation of D by means of an index or an alternative representation for faster execution.

- **Pattern and outlier detection.** The discovery of recurrent patterns or atypical sub-windows in a time series has applications in finance, industrial manufacture or seismology, to name a few. It calls for techniques that avoid pairwise comparisons of all the sub-windows, which would lead to prohibitive response times.

- **Clustering**. The goal is the same as categorical data clustering: group similar time series and separate dissimilar ones.

One main problem in data analytics is to deal with data streams. Existing methods have been designed for very large data sets where complex algorithms from artificial intelligence (AI) were not efficient because of data size. However, we now must deal with data streams, sequences of data events arriving at high rate, where traditional data analytics techniques cannot complete in real-time, given the infinite data size. In order to extract knowledge from data streams, the data mining community has investigated approximation methods that could yield good result quality.

3.5 Machine Learning for High Dimensional Data Processing

High dimensionality is inherent in applications involving images, audio and text as well as in many scientific applications involving raster data or high-throughput data. Because of the *dimensionality curse*, technologies for processing and analyzing such data cannot rely on traditional relational database systems or data mining methods. It rather requires machine learning methods such as dimensionality reduction, representation learning or random projection. The activity of Zenith in this domain focuses on methods for large-scale data processing, in particular in the presence of strong uncertainty and/or ambiguity. Actually, while small datasets are often characterized by a careful collection process, massive amounts of data often come with outliers and spurious items, because it appears impossible to guarantee faultless collection at massive bandwidth. Another source of noise is often the sensor itself, that may be of low quality but of high sampling rate, or even the actual content, e.g. in cultural heritage applications when historical content appears seriously damaged by time. To attack these difficult problems, we focus on the following research topics:

- **Uncertainty estimation**. Items in massive datasets may either be uncertain, e.g. for automatically annotated data as in image analysis, or be more or less severely corrupted by noise, e.g. in noisy audio recordings or in the presence of faulty sensors. In both cases, the concept of *uncertainty* is central for the end-user to exploit the content. In this context, we use probability theory to quantify uncertainty and propose machine learning algorithms that may operate robustly, or at least assess the quality of their output. This vast topic of research is guided by large-scale applications (both data search and data denoising), and our research is oriented towards computationally effective methods.

- **Deep neural networks**. A major breakthrough in machine learning performance has been the advent of deep neural networks, which are characterized by high numbers (millions) of parameters and scalable learning procedures. We are striving towards original architectures and methods that are theoretically grounded and offer state-of-the-art performance for data search and processing. The specific challenges we investigate are: very high dimensionality for static data and very long-term dependency for temporal data, both in the case of possibly strong uncertainty or ambiguity (e.g. hundreds of thousands of classes).

- **Community service**. Research in machine learning is guided by applications. In Zenith, two main communities are targeted: botany, and digital humanities. In both cases, our observation is that significant breakthroughs may be achieved by connecting these communities to machine learning researchers. This may be achieved through wording application-specific problems in classical machine learning parlance. Thus, the team is actively involved in the organization of international evaluation campaigns that allow machine learning researchers to propose new methods while solving important application problems.

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 INRAE, IRD and CIRAD. However, we are also addressing other scientific domains (e.g. astronomy, oil extraction, music processing) through our international collaborations.

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 has hundreds of TB. 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**. Today, it is possible to acquire data on many domains related to personal data. For instance, one can collect data on daily activities, habits or health. It is also possible to measure performance in sports. This can be done thanks to sensors, communicating devices or even connected glasses. 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 some individuals, it can be interesting to find a category that corresponds to their performance in a specific sport and then adapt their 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 with solutions that can extract such knowledge from everybody's data, while guaranteeing that their private data will not 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.

- **Biological 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 HIRROS and PhenoArch at INRAE 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 months), at different sites and at different scales ranging from small tissue samples to the entire plant until whole plant population. Analyzing such big data creates new challenges for data management and data integration, but also for plant modeling. We address this application in the context of the French initiative OpenAlea, with CIRAD and INRAE.

- **Large language models for genomics.** In the context of a collaboration with CNRS - INRAE, we are developing an activity on large language models applied to genomics. In particular , our work focuses on *inverse folding*, i.e., predicting a sequence of amineo acids that are able to generate a given protein structure, with applications in the drug design industry. These models involve training large deep models on several millions of structural data samples. We also investigated explanatory methods for large language models.

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 Social and environmental responsibility

We do consider the ecological impact of our technology, especially large data management.

- We address the (major) problem of energy consumption of our ML models, by introducing energy-based metrics to assess the energy consumption during the training on GPU of our ML models. Furthermore, we want to improve training pipelines that reduce the need for training models from scratch. At inference, network compression methods can reduce the memory footprint and the computational requirements when deploying models.
- In the design of the Pl@ntnet mobile application, we adopt an eco-responsible approach, taking care not to integrate addictive, energy-intensive or non-essential functionalities to uses that promote the preservation of biodiversity and environment.
- To reduce our carbon footprint, we reduce to the minimum the number of long-distance trips, and favor train as much as possible. We also foster journal publications, to avoid traveling. For instance, in 2023, we have 12 journal publications versus 15 conference publications.

6 Highlights of the year

6.1 Awards

Daniel Rosendo's PhD thesis "Methodologies for Reproducible Analysis of Workflows on the Edge-to-Cloud Continuum" [47], supervised by Gabriel Antoniu, Alexandru Costan, and Patrick Valduriez (Inria, France) obtained the Second Place at the Best PhD thesis award of the French Database Conference (BDA), 2023.

6.2 Inria Emeritus

The workshop “Éméritat de Patrick Valduriez” on June, 5th 2023 in Montpellier gathered about 80 people, to discuss past, current and future research in data management.

7 New software, platforms, open data

7.1 New software

7.1.1 Pl@ntNet

Keywords: Plant identification, Deep learning, Citizen science

Functional Description: Pl@ntNet is a participatory platform and information system dedicated to the production of botanical data through deep learning-based plant identification. It includes 3 main front-ends, an Android app (the most advanced and the most used one), an iOs app (being currently re-developed) and a web version. The main feature of the application is to return the ranked list of the most likely species providing an image or an image set of an individual plant. In addition, Pl@ntNet's search engine returns the images of the dataset that are the most similar to the queried observation allowing interactive validation by the users. The back-office running on the server side of the platform is based on Snoop visual search engine (a software developed by ZENITH) and on NewSQL technologies for the data management. The application is distributed in more than 200 countries (30M downloads) and allows identifying about 35K plant species at present time. The platform integrates an open access REST API (my.plantnet.org) that currently accounts for 6500 developers accounts.

Publications: hal-01629195, hal-02937618, hal-03343235, hal-01182775

Contact: Alexis Joly

Participants: Antoine Affouard, Jean-Christophe Lombardo, Pierre Bonnet, Hervé Goëau, Mathias Chouet, Hugo Gresse, Julien Champ, Alexis Joly

7.1.2 ThePlantGame

Keyword: Crowd-sourcing

Functional Description: ThePlantGame is a participatory game whose purpose is the production of big taxonomic data to improve our knowledge of biodiversity. One major contribution 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. Thousands of players are registered and produce on average about tens new validated plant observations per day. The accuracy of the produced taxonnomic tags is very high (about 95%), which is quite impressive considering the fact that a majority of users are beginners when they start playing.

Publication: hal-01629149

Contact: Alexis Joly

Participants: Maximilien Servajean, Alexis Joly

7.1.3 Savime

Name: Simulation And Visualization IN-Memory

Keywords: Data management., Distributed Data Management

Functional Description: SAVIME is a multi-dimensional array DBMS for scientific applications. It supports a novel data model called TARS (Typed ARray Schema), which extends the basic array data model with typed arrays. In TARS, the support of application dependent data characteristics is provided through the definition of TAR objects, ready to be manipulated by TAR operators. This approach provides much flexibility for capturing internal data layouts through mapping functions, which makes data ingestion independent of how simulation data has been produced, thus minimizing ingestion time.

Publication: lirmm-01620376

Contact: Patrick Valduriez

Participants: Hermano Lustosa, Fabio Porto, Patrick Valduriez

Partner: LNCC - Laboratório Nacional de Computação Científica

7.1.4 OpenAlea

Keywords: Bioinformatics, Biology

Functional Description: OpenAlea is an open source project primarily aimed at the plant research community. It is a distributed collaborative effort to develop Python libraries and tools that address the needs of current and future works in Plant Architecture modeling. It includes modules to analyze, visualize and model the functioning and growth of plant architecture. It was formally developed in the Inria VirtualPlants team.

Release Contributions: OpenAlea 2.0 adds to OpenAlea 1.0 a high-level formalism dedicated to the modeling of morphogenesis that makes it possible to use several modeling paradigms (Blackboard, L-systems, Agents, Branching processes, Cellular Automata) expressed with different languages (Python, L-Py, R, Visual Programming, ...) to analyse and simulate shapes and their development.

Publications: hal-01166298, hal-00831811

Contact: Christophe Pradal

Participants: Christian Fournier, Christophe Godin, Christophe Pradal, Frédéric Boudon, Patrick Valduriez, Esther Pacitti, Yann Guédon

Partners: CIRAD, INRAE

7.1.5 Imitates

Name: Indexing and mining Massive Time Series

Keywords: Time Series, Indexing, Nearest Neighbors

Functional Description: Time series indexing is at the center of many scientific works or business needs. The number and size of the series may well explode depending on the concerned domain. These data are still very difficult to handle and, often, a necessary step to handling them is in their indexing. Imitates is a Spark Library that implements two algorithms developed by Zenith. Both algorithms allow indexing massive amounts of time series (billions of series, several terabytes of data).

Publication: lirmm-01886794

Contact: Florent Masseglia

Partners: New York University, Université Paris-Descartes

7.1.6 UMX

Name: open-unmix

Keywords: Source Separation, Audio

Scientific Description: UMX implements state of the art audio/music source separation with deep neural networks (DNNs). It is intended to serve as a reference in the domain. It has been presented in two major scientific communications: An Overview of Lead and Accompaniment Separation in Music (https://hal-lirmm.ccsd.cnrs.fr/lirmm-01766781) and Music Separation with DNNs (Making it work (ISMIR 2018 Tutorial) https://sigsep.github.io/ismir2018_tutorial/index.html#/cover).

Functional Description: UMX implements audio source separation with deep learning, using the Pytorch and Tensorflow frameworks. It comprises the code for both training and testing the separation networks, in a flexible manner. Pre- and post-processing around the actual deep neural nets include sophisticated specific multichannel filtering operations.

Publication: lirmm-01766781

Authors: Antoine Liutkus, Fabian Robert Stoter, Emmanuel Vincent

Contact: Antoine Liutkus

7.1.7 TDB

Keywords: Data assimilation, Big data, Data extraction

Scientific Description: TDB comes as a building block for audio machine learning pipelines. It is a scraping tool that allows large scale data augmentation. Its different components allow building a large dataset of samples composed of related audio tracks, as well as the associated metadata. Each sample comprises a dynamic number of entries.

Functional Description: TDB is composed of two core submodules. First, a data extraction pipeline allows scraping a provider url so as to extract large amounts of audio data. The provider is assumed to offer audio content in a freely-accessible way through a hardcoded specific structure. The software automatically downloads the data locally under a raw data format. To aggregate the raw data set, a list of item ids is used. The item ids will be requested from the provider given a url in parallel fashion. Second, a data transformation pipeline allows transforming the raw data into a dataset that is compatible with machine learning purposes. Each produced subfolder contains a set of audio files corresponding to a predefined set of sources, along with the associated metadata. A working example is provided.

Each component has several submodules, in particular, network handling and audio transcoding. Thus, TDB can be viewed as an extract-transform-load (ETL) pipeline that enables applications such as deep learning on large amounts of audio data, assuming that an adequate data provider url is fed into the software.

Contact: Antoine Liutkus

Participants: Antoine Liutkus, Fabian Robert Stoter

7.1.8 UMX-PRO

Name: Unmixing Platform - PRO

Keywords: Audio signal processing, Source Separation, Deep learning

Scientific Description: UMX-PRO is written in Python using the TensorFlow 2 framework and provides an off-the-shelf solution for music source separation (MSS). MSS consists in extracting different instrumental sounds from a mixture signal. In the scenario considered by UMX-PRO, a mixture signal is decomposed into a pre-defined set of so called targets, such as: (scenario 1) {"vocals", "bass", "drums", "guitar", "other"} or (scenario 2) {"vocals", "accompaniment"}.

The following key design choices were made for UMX-PRO. The software revolves around the training and inference of a deep neural network (DNN), building upon the TensorFlow v2 framework. The DNN implemented in UMX-PRO is based on a BLSTM recurrent network. However, the software has been designed to be easily extended to other kinds of network architectures to allow for research and easy extensions. Given an appropriately formatted database (not part of UMX-PRO), the software trains the network. The database has to be split into train and valid subsets, each one being composed of folders called samples. All samples must contain the same set of audio files, having the same duration: one for each desired target. For instance: {vocals.wav, accompaniment.wav}. The software can handle any number of targets, provided they are all present in all samples. Since the model is trained jointly, a larger number of targets increases the GPU memory usage during training. Once the models have been trained, they can be used for separation of new mixtures through a dedicated end-to-end separation network. Interestingly, this end-to-end network comprises an optional refining step called expectation-maximization that usually improves separation quality.

Functional Description: UMX-PRO implements a full audio separation deep learning pipeline in Tensorflow v2. It provides everything needed to train and use a deep learning model for separating music signals, including network architecture, data pipeline, training code, inference code as well as pre-trained weights. The software comes with full documentation, detailed comments and unit tests.

Authors: Antoine Liutkus, Fabian Robert Stoter

Contact: Antoine Liutkus

7.2 Open data

7.2.1 The GeoLifeCLEF 2023 dataset

The difficulty to measure or predict species community composition at fine spatio-temporal resolution and over large spatial scales severely hampers our ability to understand species assemblages and take appropriate conservation measures. We designed a European scale dataset (GeoLifeCLEF 2023 [48]) covering around ten thousand plant species to calibrate and evaluate Species Distribution Model (SDM) predictions of species composition in space and time at high spatial resolution (ten meters), and their spatial transferability. For model training, we extracted and harmonized five million heterogeneous presence-only records from selected datasets from the Global Biodiversity Information Facility and 6 thousand exhaustive presence-absence surveys both sampled during 2017-2021. We associated species observations to diverse environmental rasters classically used in SDMs, as well as to 10 m resolution

RGB and Near-Infra-Red satellite images and 20 years-time series of climatic variables and satellite point values. The GeoLifeCLEF 2023 dataset is open access and the first benchmark for researchers aiming to improve the prediction of plant species composition at a very fine spatial grain and at continental scale.

8 New results

8.1 Distributed Data and Model Management

8.1.1 Scientific Workflows for Life Science

Participants: Reza Akbarinia, Christophe Botella, Alexis Joly, Florent Masseglia, Esther Pacitti, Christophe Pradal, Patrick Valduriez.

Data driven science requires manipulating large datasets coming from various data sources through complex workflows based on a variety of models and languages. However, current solutions are typically ad-hoc, specialized for particular data, models and workflow systems. Our solution is an open service-based architecture, Life Science Workflow Services (LifeSWS) [14], which provides data analysis workflow services for life sciences. We illustrate our motivations and rationale for the architecture with real use cases from life science. LifeSWS capitalizes on our collaboration with Brazil and NYU in developing major systems for scientific applications.
In this context, we provide different services for plant and crop models with OpenAlea:

- Source code model transformation between programming languages. CyMLTx allows exchanging and reusing crop model components between modeling platforms beyond programming languages using reverse engineering for automatic code-to-code transformation [23].
- Spatio-temporal analysis and modeling of plant architecture. First, we developed a spatio-temporal analysis workflow of strawberry architecture and 3D visualization to study the influence of environmental and genetic cues on strawberry architecture and yield [19]. Then, we developed a functional structural model to simulate the water and solute transport inside the root architecture to determine the respective radial and axial hydraulic conductivities in water deficit conditions [15]. Finally, we developed a SEIR (Susceptible, Exposed, Infectious, Removed) model of two damaging wheat diseases in the context of crop mixture and agroecology [20]. The model suggests that the barrier effect is maximized for an intermediate proportion of companion crop.

8.1.2 Distributed Intelligence on the Edge-to-Cloud Continuum

Participants: Daniel Rosendo, Patrick Valduriez.

The large scale and optimized deployment of learning-based workflows across the Edge-to-Cloud Continuum requires extensive and reproducible experimental analysis of application execution on representative testbeds. In this context, we propose two complementary solutions: KheOps and ProvLight.

- KheOps [39] is a collaborative environment specifically designed to enable cost-effective reproducibility and replicability of Edge-to-Cloud experiments. We illustrate KheOps with a real-life Edge-to-Cloud application. The experimental results show how KheOps helps authors to systematically perform repeatable and reproducible experiments on the Grid5000 and FIT IoT LAB testbeds.

- ProvLight [40, 47] is a tool that enables efficient provenance capture on the IoT/Edge, leveraging simplified data models, data compression and grouping, and lightweight transmission protocols to reduce overheads. Our validation at large scale with synthetic workloads on 64 real-life IoT/Edge devices shows that ProvLight outperforms state-of-the-art systems in terms of in terms of speed to capture and transmit provenance data, and resource (CPU, memory and energy) consumption.

8.1.3 Model Management with Gypscie

Participants: Patrick Valduriez.

The success of machine learning (ML) systems depends on data availability, volume, quality, and efficient computing resources. A problem in this context is to reduce computational costs while maintaining adequate accuracy of the models. To address this problem, we propose a framework to reduce computational costs while maintaining adequate accuracy of ML models [38]. It identifies "subdomains" within the input space and train local models that produce better predictions for samples from that specific subdomain, instead of training a single global model on the full dataset. Our experimental validation on two real-world datasets shows that subset modeling improves the predictive performance compared to a single global model and allows data-efficient training.

8.1.4 GeoAI for Marine Ecosystem Monitoring: a Complete Workflow to Generate Maps from AI Model Predictions

Participants: Matteo Contini, Alexis Joly.

Mapping and monitoring marine ecosystems imply several challenges for data collection and processing: water depth, restricted access to locations, instrumentation costs or weather constraints for sampling, among others. Nowadays, AI and Geographic Information System (GIS) open source software can be combined in new kinds of workflows, to annotate and predict objects directly on georeferenced raster data (e.g. orthomosaics). Here, we describe and share the code of a generic method to train a deep learning model with spatial annotations and use it to directly generate model predictions as spatial features [25]. This workflow has been tested and validated in three use cases related to marine ecosystem monitoring at different geographic scales: (i) segmentation of corals on orthomosaics made of underwater images to automate coral reef habitats mapping, (ii) detection and classification of fishing vessels on remote sensing satellite imagery to estimate a proxy of fishing effort (iii) segmentation of marine species and habitats on underwater images with a simple geolocation. Models have been successfully trained and the models predictions are displayed with maps for the three use cases.

8.1.5 A Global Infrastructure to Exploit the Potential of Digitized Collections

Participants: Alexis Joly.

Tens of millions of images from biological collections have been made available online in the last two decades. In parallel, there has been a dramatic increase in the capabilities of image analysis technologies, in particular using machine learning and computer vision. While image analysis has become mainstream in consumer applications, it is still only used on an artisanal basis in the biological collections community, largely because the image corpora are dispersed. Yet, there is massive untapped potential for novel applications and research if the images of collection objects could be made accessible as a single corpus. Thus, we make the case [18] for building an infrastructure that could support image analysis of collection objects. We show that such an infrastructure is entirely feasible as well as worth the investment.

8.2 Data Analytics

8.2.1 Time Series Prediction

Participants: Reza Akbarinia, Florent Masseglia, Esther Pacitti.

Time series event detection methods are evaluated mainly by standard classification metrics that focus solely on detection accuracy. However, inaccuracy in detecting an event can often result from its preceding or delayed effects reflected in neighboring detections. To address this problem, we proposed SoftED metrics [54], a new set of metrics designed for soft evaluating event detection methods, which enable the evaluation of both detection accuracy and the degree to which their detections represent events. They improve event detection evaluation by associating events and their representative detections, incorporating temporal tolerance in over 36% of experiments compared to the usual classification metrics.
We also continued the implementation of TSPred for time series prediction in association with data preprocessing [44]. TSPred establishes a prediction process that seamlessly integrates nonstationary time series transformations with state-of-the-art statistical and machine learning methods. It is made available as an R-package, which provides functions for defining and conducting time series prediction, including data pre(post)processing, decomposition, modeling, prediction, and accuracy assessment.

8.2.2 Spatial-time Motif Analysis

Participants: Esther Pacitti.

Time series mining aims at discovering motifs, i.e., patterns that occur many times. However, it is difficult to understand and characterize the meaning of the motifs obtained concerning the data domain, and analyzing the quality of the results obtained. STMotif Explorer [26] is a spatial-time motif analysis system that aims to interactively discover and visualize spatial-time motifs in different domains, offering insight to users. STMotif Explorer enables users to use and implement novel spatiotemporal motif detection techniques and then run this across various domains.
We released the GSTSM R package [28], the first tool for mining spatial time-stamped sequences in constrained space and time. It allows users to search for spatio-temporal patterns that are not frequent in the entire database, but are dense in restricted time-space intervals. Thus, it makes it possible to find non-trivial patterns that would not be found using common data mining tools.

8.2.3 Variable-Size Segmentation for Time Series Representation

Participants: Reza Akbarinia, Florent Masseglia.

SAX is one of the most popular time series representations, allowing dimensionality reduction on the classic data mining tasks. It constructs symbolic representations by splitting the time domain into segments of equal size where the mean values of segments represent the time series intervals. This approximation technique is effective for time series having a uniform and balanced distribution over the time domain. However, in the case of time series having high variation over given time intervals, this division into segments of fixed length may lead to significant information loss.
In [17], we propose a new time series representation approach, called ASAX, that by smart selection of variable-size segments allows to significantly reduce the information loss in the time series representation. Particularly, we propose two new representation techniques, called ASAX_EN (based on the entropy) and ASAX_SAE (based on SAE the Sum of Absolute Errors), that allow obtaining a variable-size segmentation of time series with better precision in retrieval tasks thanks to the lower information loss. The experimental results illustrate that our techniques can obtain significant performance gains in terms of precision for similarity search compared to SAX, particularly for datasets with non-uniform distributions.

8.2.4 kNN Matrix Profile for Knowledge Discovery from Time Series

Participants: Reza Akbarinia, Florent Masseglia.

Matrix Profile (MP) has been proposed as a powerful technique for time series analysis, e.g., detecting motifs or anomalies. The definition of MP in the literature is as follows: given a time series T and a subsequence length m, the MP returns for each subsequence included in T its distance to the most similar subsequence (1NN) in the time series. We call 1NN MP this type of MP. Although 1NN MP has been shown useful for knowledge discovery, it has its own drawbacks. Particularly, it does not allow to detect a cluster of discords, e.g., two subsequences that are similar to each other, but dissimilar to all other subsequences. In [24], we propose a new type of matrix profile, called kNN MP, and illustrate its utility for knowledge discovery from time series. Particularly, we define the kNN MP, and propose a fast algorithm to calculate it in time series. We also propose a technique for parallel execution of the proposed algorithm by using multiple cores of an off-the-shelf computer. The experimental evaluation illustrates the efficiency of our solution. For example, the accuracy of anomaly detection can be improved from 37% with 1NN MP to 99% with 10NN MP, for one of the benchmarks of the Yahoo dataset.

8.2.5 Explainability of large language models

Participants: Antoine Liutkus, Ondrej Cifka.

The increasingly widespread adoption of large language models has highlighted the need for improving their explainability. In [42], we introduce context length probing as a novel explanation technique for causal language models, based on tracking the predictions of a model as a function of the length of available context, and allowing to assign differential importance scores to different contexts. The technique is model-agnostic and does not rely on access to model internals beyond computing token-level probabilities. We apply context length probing to large pre-trained language models and offer some initial analyses and insights, including the potential for studying long-range dependencies.

8.3 Machine Learning for Biodiversity and Agroecology

8.3.1 Deep Learning for Agroecology

Participants: Hervé Goëau, Alexis Joly, Shamprikta Mehreen.

One obstacle to generalize new practices in agroecology is that they often require expert skills or costly manual work from field workers. Digital technologies and AI in particular can play a crucial role in removing this barrier.

In [22], we worked on automatically estimating compositions and nutritional values of seed mixes based on Vision Transformers. For this purpose, an original open image dataset has been built containing 4,749 images of seed mixes, covering 11 seed varieties, with which 2 types of deep learning models have been evaluated. The best-performing model, a vision transformer pre-trained with self-supervision allows an estimation of the nutritional value of seed mixtures with a coefficient of determination R2 score of 0.91, which demonstrates the interest of this type of approach for its possible use on a large scale.

In [30], we worked on identifying plant diseases on a large scale targeting multiple crops and diseases. Existing methods usually address the problem with general supervised recognition tasks based on the seen composition of the crop and the disease. However, ignoring the composition of unseen classes during model training can lead to a reduction in model generalization. Therefore, in this work, we propose a new approach that leverages the visual features of crop and disease from the seen composition, using them to learn the features of unseen crop-disease composition classes.

8.3.2 Evaluation of Species Identification and Prediction Algorithms

Participants: Alexis Joly, Herve Goeau, Benjamin Deneu, Titouan Lorieul, Camille Garcin.

We ran a new edition of the LifeCLEF evaluation campaign [32, 31] with the involvement of hundreds of data scientists and research teams worldwide. Overall, the study shows that the field continues to progress year after year, and that, although the challenges that are most closely related to common tasks, such as multi-class classification based on images, are able to profit from the most recent advances in computer vision, certain problems are still wide open, such as the prediction of species as a function of location (as part of the GeoLifeCLEF challenge [48]). In terms of the methods used, the results show that convolutional neural networks (CNNs) are still a very powerful method for image and sound processing. In 4 of the 5 challenges, the best results were obtained using CNNs. Only the PlantCLEF challenge [29] obtained much better results (for the identification of plants from images) with the use of foundation vision transformer models such as EVA. Complementary to vision-based models, Natural Language Processing (NLP) models were also used successfully, in particular hybrid models such as CLIP that efficiently learn visual concepts from natural language supervision. We believe that this principle of combining different modalities in the training of deep learning models will be a key to future progress in AI for biodiversity.

8.3.3 New features in the Pl@ntNet platform

Participants: Alexis Joly, Benjamin Deneu, Jean-Christophe Lombardo, Antoine Affouard.

Pl@ntnet is a citizen observatory that relies on AI technologies to help people identify plants with their smartphones. Over the past few years, Pl@ntNet has become one of the largest plant biodiversity observatories in the world with several million contributors. A set of new features were developed in 2023.

First, we migrated the management of taxa towards POWO (Plant of the World Online), an online database and resource managed by the Royal Botanic Gardens, Kew (UK). It aims to provide comprehensive and up-to-date information on plant names, taxonomy, distribution, and other relevant botanical data. One of the main advantages is that it allows us to manage a unique global flora that can be filtered by biogeographical region (see here for instance). This migration required a major development effort, as it affected all the platform's components (database, data access layer, AI identification engine, mobile and web front-end).

Second, we changed the AI model for species recognition to DinoV2, a self-supervised vision transformer co-developed by Inria and Meta AI. We first tested and evaluated it on the PlantCLEF 2023 dataset to compare its performance with other state-of-the-art models such as BEIT. We then extended its architecture to address Pl@ntNet's multi-task problem (view type classification, rejection and species identification) and then integrated it into the training workflow of the platform. We benchmarked and validated it on the platform's test bench which allowed to confirm that it provides a performance gain on most of our evaluation datasets. Finally, we tested it in real-world conditions on a test server for a few weeks before putting it into production.

Third, we have developed a new feature to identify all plants in images of vegetation plots (typically canopy quadrats). It is based on the tiling principle of the classic Pl@ntNet model and returns a list of the species present, together with statistics on their spatial coverage. This functionality has been integrated into the Pl@ntNet API, as well as into a beta version of the web front end currently being tested.

8.3.4 Forecasting animal migrations

Participants: Antoine Liutkus, Ondrej Cifka.

In the context of an ongoing collaboration with ecologists from CEFE (UMR5175), the MultiNode project (NUMEV) aims at exploiting state of the art forecasting methods for better modeling and understanding animal migrating behaviors.

The movement of animals is a central component of their behavioral strategies. Statistical tools for movement data analysis, however, have long been limited, and in particular, unable to account for past movement information except in a very simplified way. In [49], we propose MoveFormer, a new step-based model of movement capable of learning directly from full animal trajectories. While inspired by the classical step-selection framework and previous work on the quantification of uncertainty in movement predictions, MoveFormer also builds upon recent developments in deep learning, such as the Transformer architecture, allowing it to incorporate long temporal contexts.

9 Partnerships and cooperations

9.1 International research visitors

9.1.1 Visits of international scientists

International visits to the team

Dennis Shasha

Status: researcher

Institution of origin: New York University

Country: USA

Dates: 17 april - 11 june

Context of the visit: AI3P project (MUSE iSite)

Mobility program/type of mobility: research stay, lectures

Fabio Porto

Status: researcher

Institution of origin: Laboratório Nacional de Computação Científica (LNCC)

Country: Brazil

Dates: 28 may - 6 june, 21 oct - 30 oct

Context of the visit: HPDaSc associated team

Mobility program/type of mobility: research stay, lectures

9.1.2 Visits to international teams

Esther Pacitti

Visited institution: LNCC, Universidade Federal do Rio de Janeiro (UFRJ), Universidade Federal Fluminense (UFF), Centro Federal de Educação Tecnológica (CEFET)

Country: Brazil

Dates: 21 march - 9 may, 9 july - 20 august

Context of the visit: HPDaSc associated team

Mobility program/type of mobility: research stay

Patrick Valduriez

Visited institution: LNCC, UFRJ, UFF, CEFET, USP

Country: Brazil

Dates: 21 march - 9 may, 9 july - 20 august

Context of the visit: HPDaSc associated team

Mobility program/type of mobility: research stay

Visited institution: University of Waterloo

Country: Canada

Dates: 1-5 sept

Context of the visit: VLDB conference

Mobility program/type of mobility: lecture

Christophe Pradal

Visited institution: Wageningen University

Country: The Nederlands

Dates: 28 - 30 august

Context of the visit: PhD jury, lecture

Mobility program/type of mobility: research stay

Visited institution: University of Florida

Country: USA

Dates: 26 nov - 3 december

Context of the visit: International consortium building

Mobility program/type of mobility: research stay

9.2 European initiatives

9.2.1 Horizon Europe

B3 B3 project on cordis.europa.eu

Title: Biodiversity Building Blocks for policy

Duration: From March 1, 2023 to August 31, 2026

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- UNIVERSITATEA OVIDIUS DIN CONSTANTA (OVIDIUS UNIVERSITY OF CONSTANTA), Romania
- MARTIN-LUTHER-UNIVERSITAT HALLE-WITTENBERG (MLU), Germany

- Global Biodiversity Information Facility (GBIFS), Denmark
- EIGEN VERMOGEN VAN HET INSTITUUT VOOR NATUUR- EN BOSONDERZOEK (EV INBO), Belgium
- LA TROBE UNIVERSITY (LTU), Australia
- JUSTUS-LIEBIG-UNIVERSITAET GIESSEN (JLU), Germany
- UNIVERSIDADE DE AVEIRO (UAveiro), Portugal
- SOUTH AFRICAN NATIONAL BIODIVERSITY INSTITUTE (SANBI), South Africa
- AGENTSCHAP PLANTENTUIN MEISE (AGENCE JARDIN BOTANIQUE DE MEISE), Belgium
- ALMA MATER STUDIORUM - UNIVERSITA DI BOLOGNA (UNIBO), Italy
- PENSOFT PUBLISHERS (PENSOFT), Bulgaria
- STELLENBOSCH UNIVERSITY (SU UNIVERSITY OF STELLENBOSCH), South Africa

Inria contact: Alexis Joly

Coordinator: AGENTSCHAP PLANTENTUIN MEISE

Summary: The world is changing rapidly; climate change, land use change, pollution and natural resource exploitation are creating a global crisis for biodiversity whose magnitude and dynamics are hard to quantify. Decision makers at all levels need up-to-date information from which to evaluate policy options. For this reason rapid, reliable, repeatable monitoring of biodiversity data is needed at all scales from local to global. Only by leveraging large volumes of data, advanced modeling techniques and powerful computing tools can we hope to synthesize these data within timescales that are relevant to policy.

Data on biodiversity come from a diverse range of sources, citizen scientists, museums, herbaria and researchers are all major contributors, but increasingly new technologies are being deployed, such as automatic sensors, camera traps, eDNA and satellite tracking. Integrating these data is a major challenge, but is necessary if we are to create dependable information on biodiversity change. B3 will use the concept of data cubes to simplify and standardize access to biodiversity data using the Essential Biodiversity Variables framework. These cubes will be used, in conjunction with other environmental data and scenarios, as the basis for models and indicators of past, current and future biodiversity.

The overarching goal of the project is to provide easy access to tools in a cloud computing environment, in real-time and on-demand, with state of the art prediction models of biodiversity, that will output models and indicators of biodiversity status and change. The project envisages a future where primary biodiversity data are seamlessly integrated into monitoring and forecasting such that policy and management can proactively respond to problems while at the same time reduce the costs of monitoring and management, and the negative impacts of biodiversity change.

GUARDEN GUARDEN project on cordis.europa.eu

Title: safeGUARDing biodivErsity aNd critical ecosystem services across sectors and scales

Duration: From November 1, 2022 to October 31, 2025

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- PARC NATIONAL DE PORT-CROS (CONSERVATOIRE BOTANIQUE NATIONAL MEDITERRANEEN DE PORQUEROLLES), France
- STICHTING NATURALIS BIODIVERSITY CENTER (NATURALIS), Netherlands
- MINISTRY OF AGRICULTURE, RURAL DEVELOPMENT AND ENVIRONMENT OF CYPRUS, Cyprus

- PLYMOUTH MARINE LABORATORY LIMITED (PML), United Kingdom
- UNIVERSITY OF ANTANANARIVO, Madagascar
- CHAROKOPEIO PANEPISTIMIO (HAROKOPIO UNIVERSITY OF ATHENS (HUA)), Greece
- AGENCIA ESTATAL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS (CSIC), Spain
- DRAXIS ENVIRONMENTAL SA (DRAXIS), Greece
- EBOS TECHNOLOGIES LIMITED (eBOS), Cyprus
- CENTRE DE COOPERATION INTERNATIONALE EN RECHERCHE AGRONOMIQUE POUR LEDEVELOPPEMENT - C.I.R.A.D. EPIC (CIRAD), France
- AGENTSCHAP PLANTENTUIN MEISE (AGENCE JARDIN BOTANIQUE DE MEISE), Belgium
- ENVECO ANONYMI ETAIRIA PROSTASIAS KAI DIAHIRISIS PERIVALLONTOS A.E. (ENVECO S.A. ENVIRONMENTAL PROTECTION AND MANAGEMENT), Greece
- AREA METROPOLITANA DE BARCELONA (AMB), Spain
- FREDERICK UNIVERSITY FU (FREDERICK UNIVERSITY FU), Cyprus
- EREVNITIKO PANEPISTIMIAKO INSTITOUTO SYSTIMATON EPIKOINONION KAI YPOLOGISTON (RESEARCH UNIVERSITY INSTITUTE OF COMMUNICATION AND COMPUTER SYSTEMS), Greece

Inria contact: Alexis Joly

Coordinator: CIRAD

Summary: GUARDEN's main mission is to safeguard biodiversity and its contributions to people by bringing them at the forefront of policy and decision-making. This will be achieved through the development of user-oriented Decision Support Applications (DSAs), and leveraging on Multi-Stakeholder Partnerships (MSPs). They will take into account policy and management objectives and priorities across sectors and scales, build consensus to tackle data gaps, analytical uncertainties or conflicting objectives, and assess options to implement adaptive transformative change. To do so, GUARDEN will make use of a suite of methods and tools using deep learning, earth observation, and hybrid modeling to augment the amount of standardized and geo-localized biodiversity data, build-up a new generation of predictive models of biodiversity and ecosystem status indicators under multiple pressures (human and climate), and propose a set of complementary ecological indicators likely to be incorporated into local management and policy. The GUARDEN approach will be applied at sectoral case studies involving end users and stakeholders through Multi-Stakeholder Partnerships, and addressing critical cross-sectoral challenges (at the nexus of biodiversity and deployment of energy/transport infrastructure, agriculture, and coastal urban development). Thus, the GUARDEN DSAs shall help stakeholders engaged in the challenge to improve their holistic understanding of ecosystem functioning, biodiversity loss and its drivers and explore the potential ecological and societal impacts of alternative decisions. Upon the acquisition of this new knowledge and evidence, the DSAs will help end-users not only navigate but also (re-)shape the policy landscape to make informed all-encompassing decisions through cross-sectoral integration.

MAMBO MAMBO project on cordis.europa.eu

Title: Modern Approaches to the Monitoring of BiOdiversity

Duration: From September 1, 2022 to August 31, 2026

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- AARHUS UNIVERSITET (AU), Denmark
- STICHTING NATURALIS BIODIVERSITY CENTER (NATURALIS), Netherlands

- THE UNIVERSITY OF READING, United Kingdom
- HELMHOLTZ-ZENTRUM FUR UMWELTFORSCHUNG GMBH - UFZ, Germany
- ECOSTACK INNOVATIONS LIMITED, Malta
- UK CENTRE FOR ECOLOGY & HYDROLOGY, United Kingdom
- CENTRE DE COOPERATION INTERNATIONALE EN RECHERCHE AGRONOMIQUE POUR LEDEVELOPPEMENT - C.I.R.A.D. EPIC (CIRAD), France
- PENSOFT PUBLISHERS (PENSOFT), Bulgaria
- UNIVERSITEIT VAN AMSTERDAM (UvA), Netherlands

Inria contact: Alexis Joly

Coordinator: AARHUS UNIVERSITET

Summary: EU policies, such as the EU biodiversity strategy 2030 and the Birds and Habitats Directives, demand unbiased, integrated and regularly updated biodiversity and ecosystem service data. However, efforts to monitor wildlife and other species groups are spatially and temporally fragmented, taxonomically biased, and lack integration in Europe. To bridge this gap, the MAMBO project will develop, test and implement enabling tools for monitoring conservation status and ecological requirements of species and habitats for which knowledge gaps still exist. MAMBO brings together the technical expertise of computer science, remote sensing, social science expertise on human-technology interactions, environmental economy, and citizen science, with the biological expertise on species, ecology, and conservation biology. MAMBO is built around stakeholder engagement and knowledge exchange (WP1) and the integration of new technology with existing research infrastructures (WP2). MAMBO will develop, test, and demonstrate new tools for monitoring species (WP3) and habitats (WP4) in a co-design process to create novel standards for species and habitat monitoring across the EU and beyond. MAMBO will work with stakeholders to identify user and policy needs for biodiversity monitoring and investigate the requirements for setting up a virtual lab to automate workflow deployment and efficient computing of the vast data streams (from on the ground sensors, and remote sensing) required to improve monitoring activities across Europe (WP4). Together with stakeholders, MAMBO will assess these new tools at demonstration sites distributed across Europe (WP5) to identify bottlenecks, analyze the cost-effectiveness of different tools, integrate data streams and upscale results (WP6). This will feed into the co-design of future, improved and more cost-effective monitoring schemes for species and habitats using novel technologies (WP7), and thus lead to a better management of protected sites and species.

9.2.2 H2020 projects

COS4CLOUD COS4CLOUD project on cordis.europa.eu

Title: Co-designed Citizen Observatories Services for the EOS-Cloud

Duration: From November 1, 2019 to February 28, 2023

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- TREBOLA ORGANIZACION ECOLOGICA (TREBOLA ORGANIZACION ECOLOGICA), Colombia
- ETHNIKO KAI KAPODISTRIAKO PANEPISTIMIO ATHINON (UOA), Greece
- BINEO CONSULTING S.L., Spain
- SCHMIDT NORBERT CARL (DDQ), Netherlands
- CONSERVATION EDUCATION AND RESEARCH TRUST (EARTHWATCH), United Kingdom
- AGENCIA ESTATAL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS (CSIC), Spain

- SCIENCE FOR CHANGE, SL (SCIENCE FOR CHANGE), Spain
- SVERIGES LANTBRUKSUNIVERSITET (SWEDISH UNIVERSITY OF AGRICULTURAL SCIENCES), Sweden
- 52 NORTH SPATIAL INFORMATION RESEARCH GMBH (52°North GmbH), Germany
- DYNAIKON LTD, United Kingdom
- VEREIN DER EUROPAEISCHEN BURGERWISSENSCHAFTEN - ECSA E.V. (EUROPEAN CITIZEN SCIENCE ASSOCIATION), Germany
- CENTRO DE INVESTIGACION ECOLOGICA Y APLICACIONES FORESTALES (CREAF-CERCA), Spain
- SECURE DIMENSIONS GMBH (SECURE DIMENSIONS), Germany
- THE OPEN UNIVERSITY (OU), United Kingdom

Inria contact: Alexis Joly

Coordinator: CREAF-CERCA

Summary: COS4CLOUD (Co-designed citizen observatories for the EOS-Cloud) aims to design, prototyped and implemented services that address the Open Science challenges shared by Citizen observatories of biodiversity, based on the experience of platforms like: Artportalen, Natusfera, iSpot, as well as other environmental quality monitoring platforms like: FreshWater Watch, KdUINO, OdourCollect, iSpex and CanAir.io. The innovative services will be designed, prototyped and implemented for improving the data and information quality using deep machine learning, automatic video recognition, advanced mobile app interfaces, and other cutting-edge technologies, based on data models and data protocols validated by traditional science. The new services will provide mechanisms to ensure the visibility and recognition of data contributors and the tools to improve networking between various stakeholders. Novel innovative digital services will be developed through the integration of CS products, generated by different providers, following open standards to ensure their interoperability, and offered in agile, fit-for-purpose and sustainable site available through EOSC hub, including a discovery service, to both traditional and citizen scientists. The design of new services will be user oriented, engaging a wide range of stakeholders in society, government, industry, academia, agencies, and research to co-design service requirements. As a result, COS4CLOUD will integrate citizen science in the European Open Science Cloud, bringing Citizen Science (CS) projects as a service for the scientific community and society at large.

9.2.3 Other european programs/initiatives

Creation of a Collaborative Doctoral Partnership between the EU Joint Research Centered of Ispra and the university of Montpellier - first PhD student to be recruited in 2024

9.3 National initiatives

Pl@ntAgroEco (PEPR Agroécologie & Numérique), (2023-2027), 1.6 Meuro.

Participants: Antoine Affouard, Christophe Botella, Hervé Goëau, Hugo Gresse, Alexis Joly, Thomas Paillot.

Agroecology necessarily involves crop diversification, but also the early detection of diseases, deficiencies and stresses (hydric, etc.), as well as better management of biodiversity. The main stumbling block is that this paradigm shift in agricultural practices requires expert skills in botany, plant pathology and ecology that are not generally available to those working in the field, such as farmers or agri-food technicians. Digital technologies, and artificial intelligence in particular, can play a crucial role in removing this barrier to access to knowledge.

The aim of the Pl@ntAgroEco project will be to design, experiment with and develop new high-impact agro-ecology services within the Pl@ntNet platform. This includes : AI and plant science research ; agile development of new components within the platform; organizing participatory science programs and animating the Pl@ntNet user community.

Institut de Convergence Agriculture numérique #DigitAg, (2017-2023), 275 Keuro.

Participants: Alexis Joly, Florent Masseglia, Esther Pacitti, Christophe Pradal, Patrick Valduriez.

#DigitAg brings together in a partnership of seventeen actors (public research and teaching organizations, transfer actors and companies) with the objective of accelerating and supporting the development of agriculture companies in France and in southern countries based on new tools, services and uses. Based in Montpellier with an office in Toulouse and Rennes and led by Irstea, #DigitAg's ambition is to become a world reference for digital agriculture. In this project, Zenith is involved in the analysis of big data from agronomy, in particular, plant phenotyping and biodiversity data sharing.

ANR PerfAnalytics (2021-2024), 100 Keuro.

Participants: Reza Akbarinia, Florent Masseglia.

The objective of the PerfAnalytics project is to analyze sport videos in order to quantify the sport performance indicators and provide feedback to coaches and athletes, particularly to French sport federations in the perspective of the Paris 2024 Olympic games. A key aspect of the project is to couple the existing technical results on human pose estimation from video with scientific methodologies from biomechanics for advanced gesture objectivation. The motion analysis from video represents a great potential for any monitoring of physical activity. In that sense, it is expected that exploitation of results will be able to address not only sport, but also the medical field for orthopedics and rehabilitation.

PPR Antibiorésistance: structuring tool "PROMISE" (2021-2024), 240 Keuro.

Participants: Reza Akbarinia, Florent Masseglia.

The objective of the PROMISE (PROfessional coMmunIty network on antimicrobial reSistancE) project is to build a large data warehouse for managing and anlyzing antimicrobial resitance (AMR) data. It gathers 21 existing professional networks and 42 academic partners from three sectors, human, animal, and environment. The project is based on the following transdisciplinary and cross-sectoral pillars: i) fostering synergies to improve the one-health surveillance of antibiotic consumption and AMR, ii) data sharing for improving the knowledge of professionals, iii) improving clinical research by analyzing the shared data.

PNR "Beerisk" (2022-2025). 200K Keuro.

Participants: Reza Akbarinia, Florent Masseglia.

The objective of this project is to analyze honeybee daily mortality rates, represented as time series, in order to detect anomalies and study the lethal effects of bees exposure to pesticides.

Plan national Ecoantibio "INTERSECTION" (2024-2028), 175 Keuros

Participants: Reza Akbarinia, Florent Masseglia.

The objective of the INTERSECTION project is to produce intersectoral and territorial indicators for monitoring resistance and use of antibiotics in France, and to facilitate the use and analysis of these indicators, in a One health approach.

CASDAR CARPESO (2020-2022), 87 Keuro.

Participants: Julien Champ, Hervé Goëau, Alexis Joly.

In order to facilitate the agro-ecological transition of livestock systems, the main objective of the project is to enable the practical use of meslin (grains and forages) by demonstrating their interests and remove sticking points on the nutritional value of the meslin. Therefore, it develops AI-based tools allowing to automatically assess the nutritional value of meslin from images. The consortium includes 10 chambers of agriculture, 1 Technical Institute (IDELE) and 2 research organizations (Inria, CIRAD).

9.3.1 Others

Pl@ntNet consortium membership (2019-20XX), 80 Keuro / year

Participants: Alexis Joly, Jean-Christophe Lombardo, Hervé Goëau, Hugo Gresse, Mathias Chouet, Antoine Affouard, David Margery.

This contract between four research organisms (Inria, INRAE, IRD and CIRAD) aims at sustaining the Pl@ntNet platform in the long term. It has been signed in November 2019 in the context of the InriaSOFT national program of Inria. Each partner subscribes a subscription of 20K euros per year to cover engineering costs for maintenance and technological developments. In return, each partner has one vote in the steering committee and the technical committee. He can also use the platform in his own projects and benefit from a certain number of service days within the platform. The consortium is not fixed and is intended to be extended to other members in the coming years.

9.4 Regional initiatives

Regional project "DACLIM" (2023-2026), 70 Keuros

Participants: Reza Akbarinia, Florent Masseglia.

The objective of this project is to develop scalable techniques based on massive data distribution to enable the efficient detection of anomalies in large climate databases. The detection of anomalies in climate data can provide climatologists with insights into the behavior of various climatological variables, understanding of extreme events such as heatwaves and cold snaps, as well as the prediction of these types of events.

10 Dissemination

Participants: Reza Akbarinia, Alexis Joly, Antoine Liutkus, Florent Masseglia, Esther Pacitti, Christophe Pradal, Patrick Valduriez.

10.1 Promoting scientific activities

General chair, scientific chair

- R. Akbarinia: Organization chair of the French Database Conference (BDA), October 23-26, 2023, Montpellier
- P. Valduriez: Inria-Brasil Workshop, April 10-14, 2023, Sao Paulo and Rio de Janeiro, Brazil
- P. Valduriez: Inria-Brasil Workshop on Digital Sciences and Energy, PUC-Rio, Rio de Janeiro, Brazil
- A. Joly: LifeCLEF 2023 Workshop chair, part of Conference and Labs of the Evaluation Forum (CLEF) 2023, Thessaloniki, Greece

Member of the conference program committees

- R. Akbarinia: ECML-PKDD 2023, EDBT 2023, IEEE BigData 2023, AIMLSystems 2023.
- E. Pacitti: BDA 2023.
- C. Pradal: FSPM 2023.
- F. Masseglia: ICDM, DSAA, DS, PAKDD, AIKE, SimBIG, SAC
- A. Joly: CLEF 2023

10.1.1 Journal

Editor, Associate editor

- R. Akbarinia: associate editor of IEEE Transactions on Knowledge and Data Engineering (TKDE) journal.

Member of the editorial boards

- R. Akbarinia: Transactions on Large Scale Data and Knowledge Centered Systems (TLDKS).
- P. Valduriez: Distributed and Parallel Databases.

Reviewer - reviewing activities

- R. Akbarinia: IEEE Transactions on Fuzzy Systems, Journal of Supercomputing.
- A. Joly: Nature Communications, Nature reports, PLOS One, Ecology (Wiley), Journal of Supercomputing
- F. Masseglia: JMLR, DMKD

10.1.2 Invited talks

- P. Valduriez
  - "Data Science and Innovation", COPPE/UFRJ, Rio de Janeiro, May 5, 2023 and CEFET, Rio de Janeiro, 3 May 2023.
  - "Life Science Workflow Services (LifeSWS): motivations and architecture", The Data Systems Seminar Series, University of Waterloo, September 5, 2023.
  - "Big Data Technologies", Inria Paris, November 23, 2023.
- A. Joly
  - "Cooperative learning for biodiversity monitoring", STATLEARN 2023, Montpellier, April 6, 2023.

  - "Pl@ntNet under the hood", ML-MTP, March 23, 2023.
  - "Pl@ntNet Citizen observatory & research platform", journées de l'école doctorale de Saint-Malo, Février 6-8, 2023.

- C. Pradal
  - "Towards digital twins for plants : Connecting 3d plant models and phenotyping", Seminar on Digital Twins for agri-environmental applications, Wageningen, August 30, 2023
  - "Data-intensive scientific workflows for model-assisted high-throughput phenotyping", FSPM2023, Berlin, March 27-31, 2023

10.1.3 Leadership within the scientific community

- A. Joly: contributed to an OECD book on "Artificial Intelligence in science", aimed at a broad readership, including policy makers, the public, and stakeholders in all areas of science. We co-authored the chapter entitled "Advancing the productivity of science with citizen science and artificial intelligence" [43].
- E. Pacitti: Member of the Steering Committee of the BDA conference.

10.1.4 Scientific expertise

- R. Akbarinia: member of the evaluation comitee (section 27) of University of Montpellier.
- R. Akbarinia: member of the selection committee for Associate Professor (Maître de Conférence), University Toulouse 3 Paul Sabatier
- C. Pradal: member of the INRAE evaluation comitee CSS (Scientific Specialist Commission) in Plant Integrated Biology
- A. Joly: GENCI expert committee (AI thematic)
- A. Joly: expert for LABEX CEMEB (call for postdoctoral research projects 2023)
- P. Valduriez: consultant on big data for the Software Heritage project

10.1.5 Research administration

- R. Akbarinia: Scientific referent for research data at Inria branch of University of Montpellier; Member of Inria national commission for research data.
- F. Masseglia: deputy scientific director of Inria for the domain "Perception, Cognition And Interaction".
- E. Pacitti: manager of Polytech' Montpellier's International Relationships for the computer science department (100 students).
- P. Valduriez: scientific manager for the Latin America zone at Inria's Direction of Foreign Relationships (DRI) and scientific director of the Inria-Brasil strategic partnership.
- C. Pradal: Team leader with C. Granier of the PhenoMEn team of the AGAP Institute.
- A. Joly: co-manager of a Collaborative Doctoral Partnership between the EU Joint Research Centered of Ispra and the university of Montpellier

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

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

Esther Pacitti:

- IG3: Database design, physical organization, 54h, level, L3, 50 students.
- IG4: Distributed Databases and NoSQL, 80h , level M1, 50 students.
- Large Scale Information Management (Iot, Recommendation Systems, Graph Databases), 27h, level M2, 20 students.
- Supervision of industrial projects
- Supervision of master internships.
- Supervision of computer science discovery projects.

Patrick Valduriez:

- Professional: Big Data Architectures, 24h, level M2, Capgemini Institut.

Reza Akbarinia:

- Big Data Management, 6h, level L3, University of Montpellier 3.

Alexis Joly:

- lecturer at the IDESSAI summer school 2023 .
- lecturer for an Inria academy master class at LVMH.

10.2.2 Supervision

PhD & HDR:

- Defended PhD: Daniel Rosendo, Enabling HPC-Big Data Convergence for Intelligent Extreme-Scale Analytics, Univ. Rennes. Advisors: Gabriel Antoniu, Alexandru Costan, Patrick Valduriez.
- Defended PhD: Camille Garcin, Multi-class classification with high label ambiguity and a long-tailed distribution. Advisors: Joseph Salmon, Maximilien Servajean, Alexis Joly.
- Defended PhD: Joaquim Estopinan, Species Conservation Status Prediction. Advisors: Alexis Joly, François Munoz, Maximilien Servajean, Pierre Bonnet.
- PhD in progress: Cesar Leblanc, Predicting biodiversity future trajectories through deep learning. Advisors: Alexis Joly, Maximilien Servajean, Pierre Bonnet.
- PhD in progress: Tanguy Lefort, Ambiguity of classification labels and expert feedback. Advisors: Joseph Salmon, Benjamin Charlier, Alexis Joly.
- PhD in progress: Kawtar Zaher, Novel class retrieval through interactive learning. Advisors: Olivier Buisson, Alexis Joly.
- PhD in progress: Matteo Contini, Multi-scale monitoring of coastal marine biodiversity. Advisors: Sylvain Bonhommeau, Alexis Joly.
- PhD in progress: Guillaume Coulaud, Anomaly Detection in Big Climate Data. Advisors: Reza Akbarinia, Audrey Brouillet, Florent Masseglia.

10.2.3 Juries

Members of the team participated to the following PhD or HDR committees:

- R. Akbarinia (reviewer): Qtong Wang, Paris Cité University.
- P. Valduriez (reviewer): Saalik Hatia, Sorbonne Université.
- A. Joly (jury president): Samy BENSLIMANE, Université de Montpellier.

A. Joly was a member of the jury of the INRAE participatory research award 2023

10.3 Popularization

10.3.1 Articles and contents

- A. Joly:
  - Author of the article "Tous botanistes !" in the magazine Le numérique est-il un progrès durable?, *Pour la science*, 2023
  - Author of two book chapters in Le lien à la nature à l'ère du numérique, *ISTE publisher*, Émilie Kohlmann, 2023

10.3.2 Interventions

- A. Joly:
  - Main guest of a 1 hour TV program of Sorcier TV - youtube
  - Main guest of the Radio program "Carnets de campagne" (France Inter) - podcast

11 Scientific production

11.1 Major publications

[1] C. Botella, A. Joly, P. Bonnet, F. Munoz and P. Monestiez. 'Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data'. In: *Methods in Ecology and Evolution* 12.5 (1st Feb. 2021), pp. 933–945. DOI: 10.1111/2041-210X.13565. URL: https://hal.umontpellier.fr/hal-03150701.

[2] B. Deneu, M. Servajean, P. Bonnet, C. Botella, F. Munoz and A. Joly. 'Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment'. In: *PLoS Computational Biology* 17.4 (19th Apr. 2021), e1008856. DOI: 10.1371/journal.pcbi.1008856. URL: https://hal.inrae.fr/hal-03220977.

[3] M. Fontaine, R. Badeau and A. Liutkus. 'Separation of Alpha-Stable Random Vectors'. In: *Signal Processing* (Jan. 2020), p. 107465. DOI: 10.1016/j.sigpro.2020.107465. URL: https://hal.inria.fr/hal-02433213.

[4] C. Garcin, M. Servajean, A. Joly and J. Salmon. 'Stochastic smoothing of the top-K calibrated hinge loss for deep imbalanced classification'. In: ICML 2022 - 39th International Conference on Machine Learning. Vol. 162. Baltimore, United States: PMLR, 2022, pp. 7208–7222. URL: https://hal.inria.fr/hal-03828747.

[5] G. Heidsieck, D. de Oliveira, E. Pacitti, C. Pradal, F. Tardieu and P. Valduriez. 'Cache-aware scheduling of scientific workflows in a multisite cloud'. In: *Future Generation Computer Systems* 122 (2021), pp. 172–186. DOI: 10.1016/j.future.2021.03.012. URL: https://hal.archives-ouvertes.fr/hal-03189130.

[6] J. Liu, N. Moreno Lemus, E. Pacitti, F. Porto and P. Valduriez. 'Parallel Computation of PDFs on Big Spatial Data Using Spark'. In: *Distributed and Parallel Databases* 38 (2020), pp. 63–100. DOI: 10.1007/s10619-019-07260-3. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02045144.

[7] A. Liutkus, O. Cífka, S.-L. Wu, U. Şimşekli, Y.-H. Yang and G. Richard. 'Relative Positional Encoding for Transformers with Linear Complexity'. In: ICML 2021 - 38th International Conference on Machine Learning. Proceedings of the 38th International Conference on Machine Learning. Virtual Only, United States, 18th July 2021. URL: https://hal.telecom-paris.fr/hal-03256451.

[8] A. Liutkus, U. Ş. Imşekli, S. Majewski, A. Durmus and F.-R. Stöter. 'Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions'. In: *36th International Conference on Machine Learning (ICML)*. Long Beach, United States, June 2019. URL: https://hal.inria.fr/hal-02191302.

[9] T. Mondal, R. Akbarinia and F. Masseglia. 'kNN matrix profile for knowledge discovery from time series'. In: *Data Mining and Knowledge Discovery* 37.3 (May 2023), pp. 1055–1089. DOI: 10.1007/s10618-022-00883-8. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04225369.

[10] D. Oliveira, J. Liu and E. Pacitti. *Data-Intensive Workflow Management: For Clouds and Data-Intensive and Scalable Computing Environments*. Vol. 14. Synthesis Lectures on Data Management 4. Morgan&Claypool Publishers, May 2019, pp. 1–179. DOI: 10.2200/S00915ED1V01Y201904DTM060. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02128444.

[11] T. Özsu and P. Valduriez. *Principles of Distributed Database Systems - Fourth Edition*. Télécharger la 3ieme et 4ieme édition : lien dans " voir aussi ". Springer, 2020, pp. 1–674. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02265930.

[12] D.-E. Yagoubi, R. Akbarinia, F. Masseglia and T. Palpanas. 'Massively Distributed Time Series Indexing and Querying'. In: *IEEE Transactions on Knowledge and Data Engineering* 32.1 (2020), pp. 108–120. DOI: 10.1109/TKDE.2018.2880215. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02197618.

[13] C. Zhang, R. Akbarinia and F. Toumani. 'Efficient Incremental Computation of Aggregations over Sliding Windows'. In: 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2021). Singapore, Singapore, 2021, pp. 2136–2144. DOI: 10.1145/3447548.3467360. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03359490.

11.2 Publications of the year

International journals

[14] R. Akbarinia, C. Botella, A. Joly, F. Masseglia, M. Mattoso, E. Ogasawara, D. de Oliveira, E. Pacitti, F. Porto, C. Pradal, D. Shasha and P. Valduriez. 'Life Science Workflow Services (LifeSWS): motivations and architecture'. In: *Transactions on Large-Scale Data- and Knowledge-Centered Systems*. Lecture Notes in Computer Science 14280 (2023), pp. 1–24. DOI: 10.1007/978-3-662-68100-8_1. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04173545.

[15] F. Bauget, V. Protto, C. Pradal, Y. Boursiac and C. Maurel. 'A root functional-structural model allows to assess effects of water deficit on water and solute transport parameters'. In: *Journal of Experimental Botany* 74.5 (13th Mar. 2023), pp. 1594–1608. DOI: 10.1093/jxb/erac471. URL: https://inria.hal.science/hal-03915413.

[16] M. van Der Velde, H. Goëau, P. Bonnet, R. d'Andrimont, M. Yordanov, A. Affouard, M. Claverie, B. Czúcz, N. Elvekjær, L. Martinez-Sanchez, X. Rotllan-Puig, A. Sima, A. Verhegghen and A. Joly. 'Pl@ntNet Crops: merging citizen science observations and structured survey data to improve crop recognition for agri-food-environment applications'. In: *Environmental Research Letters* 18.2 (24th Jan. 2023), p. 025005. DOI: 10.1088/1748-9326/acadf3. URL: https://hal.science/hal-03965684.

[17] L. Djebour, R. Akbarinia and F. Masseglia. 'Variable-Size Segmentation for Time Series Representation'. In: *Transactions on Large-Scale Data- and Knowledge-Centered Systems* 53 (2023), pp. 34–65. DOI: 10.1007/978-3-662-66863-4_2. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03882927.

[18] Q. Groom, M. Dillen, W. Addink, A. Ariño, C. Bölling, P. Bonnet, L. Cecchi, E. Ellwood, R. Figueira, P.-Y. Gagnier, O. Grace, A. Güntsch, H. Hardy, P. Huybrechts, R. Hyam, A. Joly, V. K. Kommineni, I. Larridon, L. Livermore, R. J. Lopes, S. Meeus, J. Miller, K. Milleville, R. Panda, M. Pignal, J. Poelen, B. Ristevski, T. Robertson, C. Rufino, J. Santos, M. Schermer, B. Scott, K. Seltmann, H. Teixeira, M. Trekels and J. Gaikwad. 'Envisaging a global infrastructure to exploit the potential of digitised collections'. In: *BDJ Open* 11 (30th Nov. 2023). DOI: 10.22541/au.166678848.82362633/v2. URL: https://inria.hal.science/hal-03871553.

[19] M. Labadie, K. Guy, M.-N. Demené, Y. Caraglio, G. Heidsieck, A. Gaston, C. Rothan, Y. Guédon, C. Pradal and B. Denoyes. 'Spatio-temporal analysis of strawberry architecture: insights into the control of branching and inflorescence complexity'. In: *Journal of Experimental Botany* 74.12 (27th June 2023), pp. 3595–3612. DOI: 10.1093/jxb/erad097. URL: https://hal.inrae.fr/hal-04084880.

[20] S. Levionnois, C. Pradal, C. Fournier, J. Sanner and C. Robert. 'Modeling the Impact of Proportion, Sowing Date, and Architectural Traits of a Companion Crop on Foliar Fungal Pathogens of Wheat in Crop Mixtures'. In: *Phytopathology* 113.10 (8th Nov. 2023), pp. 1876–1889. DOI: 10.1094/PHYTO-06-22-0197-R. URL: https://inria.hal.science/hal-04302997.

[21] J. Liu, D. Dong, X. Wang, A. Qin, X. Li, P. Valduriez, D. Dou and D. Yu. 'Large-scale Knowledge Distillation with Elastic Heterogeneous Computing Resources'. In: *Concurrency and Computation: Practice and Experience* 35.26 (2023), e7272. DOI: 10.1002/cpe.7272. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03740277.

[22] S. Mehreen, H. Goëau, P. Bonnet, S. Chau, J. Champ and A. Joly. 'Estimating Compositions and Nutritional Values of Seed Mixes Based on Vision Transformers'. In: *Plant Phenomics* 5 (10th Nov. 2023). DOI: 10.34133/plantphenomics.0112. URL: https://inria.hal.science/hal-04322179.

[23] C. A. Midingoyi, C. Pradal, A. Enders, D. Fumagalli, P. Lecharpentier, H. Raynal, M. Donatelli, D. Fanchini, I. Athanasiadis, C. Porter, G. Hoogenboom, F. Oliveira, D. Holzworth and P. Martre. 'Crop modeling frameworks interoperability through bidirectional source code transformation'. In: *Environmental Modelling and Software* 168 (Oct. 2023), p. 105790. DOI: 10.1016/j.envsoft.2023.105790. URL: https://hal.inrae.fr/hal-04256535.

[24] T. Mondal, R. Akbarinia and F. Masseglia. 'kNN matrix profile for knowledge discovery from time series'. In: *Data Mining and Knowledge Discovery* 37.3 (May 2023), pp. 1055–1089. DOI: 10.1007/s10618-022-00883-8. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04225369.

[25] J. Talpaert Daudon, M. Contini, I. Urbina-Barreto, B. Elliott, F. Guilhaumon, S. Bonhommeau, A. Joly and J. Barde. 'GeoAI for Marine Ecosystem Monitoring: a Complete Workflow to Generate Maps from AI Model Predictions'. In: *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-4/W7-2023 (22nd June 2023), pp. 223–230. DOI: 10.5194/isprs-archives-XLVIII-4-W7-2023-223-2023. URL: https://hal.science/hal-04204101.

International peer-reviewed conferences

[26] H. Borges, A. Castro, C. Souza, J. Rodrigues, F. A. Machado Porto, R. Coutinho, E. Pacitti and E. Ogasawara. 'STMotif Explorer: A Tool for Spatiotemporal Motif Analysis'. In: SBBD 2023 – Simpósio Brasileiro de Banco de Dados. Belo Honrizonte, Brazil, 25th Sept. 2023. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04283828.

[27] C. Botella, B. Deneu, D. Marcos, M. Servajean, T. Larcher, C. Leblanc, J. Estopinan, P. Bonnet and A. Joly. 'Overview of GeoLifeCLEF 2023: Species Composition Prediction with High Spatial Resolution at Continental Scale Using Remote Sensing'. In: CLEF 2023 - Working Notes of the Conference and Labs of the Evaluation Forum. Vol. 3497. CEUR Workshop Proceedings. Thessalokini, Greece, 18th Dec. 2023, pp. 1954–1971. URL: https://hal.science/hal-04322255.

[28] A. Castro, H. Borges, C. Souza, J. Rodrigues, F. A. Machado Porto, E. Pacitti, R. Coutinho and E. Ogasawara. 'GSTSM Package: Finding Frequent Sequences in Constrained Space and Time'. In: BDA 2023 – 39e Conférence sur la Gestion de Données – Principes, Technologies et Applications. Montpellier, France, 23rd Oct. 2023, pp. 1–4. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04283772.

[29] H. Goëau, P. Bonnet and A. Joly. 'Overview of PlantCLEF 2023: Image-based Plant Identification at Global Scale'. In: *CEUR Workshop Proceedings*. CLEF-WN 2023 - 24th Working Notes of the Conference and Labs of the Evaluation Forum. Vol. 3497. Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023). Thessalonique, Greece, 18th Sept. 2023, pp. 1972–1981. URL: https://hal.science/hal-04345310.

[30] A. Y. Hao Chai, S. Han Lee, F. S. Tay, Y. Lung Then, H. Goëau, P. Bonnet and A. Joly. 'Pairwise Feature Learning for Unseen Plant Disease Recognition'. In: *2023 IEEE International Conference on Image Processing (ICIP)*. ICIP 2023 - 30th IEEE International Conference on Image Processing. Kuala Lumpur, Malaysia: IEEE, 8th Oct. 2023, pp. 306–310. DOI: 10.1109/ICIP49359.2023.10222401. URL: https://hal.inrae.fr/hal-04214623.

[31] A. Joly, C. Botella, L. Picek, S. Kahl, H. Goëau, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc and T. Larcher. 'Overview of LifeCLEF 2023: evaluation of AI models for the identification and prediction of birds, plants, snakes and fungi'. In: *Lecture Notes in Computer Science*. CLEF 2023 - 14th International Conference of the CLEF Association. Vol. LNCS-14163. Experimental IR Meets Multilinguality, Multimodality, and Interaction 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, Proceedings. Thessalokini, Greece: Springer Nature Switzerland, 11th Sept. 2023, pp. 416–439. DOI: 10.1007/978-3-031-42448-9_27. URL: https://hal.science/hal-04322219.

[32] A. Joly, S. Kahl, L. Picek, C. Botella, D. Marcos, M. Šulc, M. Hrúz, S. S. Moussi, M. Servajean, E. Cole, A. Durso, H. Glotin, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet and H. Müller. 'LifeCLEF 2023 teaser: Species Identification and Prediction Challenges'. In: *LNCS. Lecture Notes in Computer Science*. ECIR 2023 - 45th European Conference on Information Retrieval. Vol. 13982. Advances in Information Retrieval : 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III. Dublin, Ireland: Springer, 16th Mar. 2023, pp. 568–576. DOI: 10.1007/978-3-031-28241-6_65. URL: https://inria.hal.science/hal-04204378.

[33] S. Kahl, T. Denton, H. Klinck, H. Reers, F. Cherutich, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué and A. Joly. 'Overview of BirdCLEF 2023: Automated Bird Species Identification in Eastern Africa'. In: *CEUR Workshop Proceedings*. CLEF 2023 - Working Notes of the Conference and Labs of the Evaluation Forum. Vol. 3497. Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023). Thessalonique, Greece, 18th Sept. 2023, pp. 1934–1942. URL: https://hal.inrae.fr/hal-04345437.

[34] R. van der Klis, S. Alaniz, M. Mancini, C. F. Dantas, D. Ienco, Z. Akata and D. Marcos. 'PDiscoNet: Semantically consistent part discovery for fine-grained recognition'. In: ICCV 2023 - International Conference on Computer Vision. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France, Oct. 2023. URL: https://hal.inrae.fr/hal-04183747.

[35] L. Perthame, C. Pradal, A. Jullien, F. Rees, C. Richard-Molard and G. Arman. 'Identification of shoot architectural traits to promote winter oilseed rape vigour during the vegetative growth: a simulation approach'. In: IRC 2023 - 16th International Rapeseed Congress. Sydney, Australia, 24th Sept. 2023. URL: https://hal.inrae.fr/hal-04263001.

[36] L. Perthame, F. Rees, X. Cornilleau, C. Richard-Molard, C. Pradal and A. Jullien. 'SIMBAL: A structural-functional plant model to simulate C and N dynamics andshoot-root architecture of winter oilseed rape associated with legumes'. In: FSPM2023 - 10th International Conference on Functional-Structural Plant Model. Book of Abstracts of the 10th International Conference on Functional-Structural Plant Models. Berlin, Germany, 2023. URL: https://hal.inrae.fr/hal-04262992.

[37] F. Rees, M. Gauthier, R. Barillot, C. Richard-Molard, A. Jullien, C. Chenu, C. Pradal and B. Andrieu. ‘Quantitative importance of various rhizodeposition processes: lessons from a mechanistic functional-structural root model’. In: FSPM 2023 - 10th International Conference on Functional-Structural Plant Models. Berlin, Germany, 2023. URL: https://hal.inrae.fr/hal-04098521.

[38] V. Ribeiro, E. Pena, R. Saldanha, R. Akbarinia, P. Valduriez, F. Arif, J. Stoyanovich and F. Porto. ‘Subset Modelling: A Domain Partitioning Strategy for Data-efficient Machine-Learning’. In: *SBBD 2023 Companion Proceedings series*. SBBD 2023 - Simpósio Brasileiro de Banco de Dados (Proceedings of the 38th Brazilian Symposium on Databases). Belo Horizonte, Brazil, 25th Sept. 2023, pp. 318–323. DOI: 10.5753/sbbd.2023.232829. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04264125.

[39] D. Rosendo, K. Keahey, A. Costan, M. Simonin, P. Valduriez and G. Antoniu. ‘KheOps: Cost-effective Repeatability, Reproducibility, and Replicability of Edge-to-Cloud Experiments’. In: *ACM REP ’23: Proceedings of the 2023 ACM Conference on Reproducibility and Replicability*. REP 2023 - ACM Conference on Reproducibility and Replicability. Santa Cruz, CA, United States: ACM, 28th June 2023, pp. 62–73. DOI: 10.1145/3589806.3600032. URL: https://hal.science/hal-04157720.

[40] D. Rosendo, M. Mattoso, A. Costan, R. Souza, D. Pina, P. Valduriez and G. Antoniu. ‘ProvLight: Efficient Workflow Provenance Capture on the Edge-to-Cloud Continuum’. In: IEEE Cluster 2023 - IEEE International Conference on Cluster Computing. Santa Fe, New Mexico, United States: IEEE, 2023, pp. 1–13. URL: https://hal.science/hal-04161546.

Conferences without proceedings

[41] A. Aniraj, C. F. Dantas, D. Ienco and D. Marcos. ‘Masking Strategies for Background Bias Removal in Computer Vision Models’. In: Workshop "Out of Distribution Generalization in Computer Vision" in conjunction with ICCV 2023. Paris, France, 2nd Oct. 2023. URL: https://hal.science/hal-04184449.

[42] O. Cífka and A. Liutkus. ‘Black-box language model explanation by context length probing’. In: *ACL Anthology*. ACL 2023 - 61st Annual Meeting of the Association for Computational Linguistics. Vol. Volume 2: Short Papers. Annual Meeting of the Association for Computational Linguistics (2023). Toronto, Canada, 2023, pp. 1067–1079. DOI: 10.18653/v1/2023.acl-short.92. URL: https://hal.umontpellier.fr/hal-03917930.

Scientific book chapters

[43] L. Ceccaroni, J. L. Oliver, E. Roger, J. Bibby, P. Flemons, K. Michael and A. Joly. ‘Advancing the productivity of science with citizen science and artificial intelligence’. In: *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*. 23rd June 2023. DOI: 10.1787/69563b12-en. URL: https://inria.hal.science/hal-04204315.

[44] R. Salles, E. Pacitti, E. Bezerra, C. Marques, C. Pacheco, C. Oliveira, F. A. Machado Porto and E. Ogasawara. ‘TSPredIT: Integrated Tuning of Data Preprocessing and Time Series Prediction Models’. In: *Transactions on Large-Scale Data- and Knowledge-Centered Systems LIV : Special Issue on Data Management - Principles, Technologies, and Applications*. Vol. LNCS14160. Lecture Notes in Computer Science. 22nd Sept. 2023, pp. 41–55. DOI: 10.1007/978-3-662-68014-8_2. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04283842.

Doctoral dissertations and habilitation theses

[45] J. Estopinan. ‘A predictive approach to determining the joint conservation status of species’. Université de Montpellier (UM), FRA, 28th Nov. 2023. URL: https://theses.hal.science/tel-04366847.

[46] C. Garcin. ‘Loss functions for set-valued classification’. Université de Montpellier, 29th Sept. 2023. URL: https://inria.hal.science/tel-04379493.

[47] D. Rosendo. ‘Methodologies for Reproducible Analysis of Workflows on the Edge-to-Cloud Continuum’. INSA RENNES, 1st June 2023. URL: https://hal.science/tel-04167278.

Reports & preprints

[48] C. Botella, B. Deneu, D. Marcos, M. Servajean, J. Estopinan, T. Larcher, C. Leblanc, P. Bonnet and A. Joly. *The GeoLifeCLEF 2023 Dataset to evaluate plant species distribution models at high spatial resolution across Europe*. 4th Aug. 2023. URL: https://hal.science/hal-04152362.

[49] O. Cífka, S. Chamaillé-Jammes and A. Liutkus. *MoveFormer: a Transformer-based model for step-selection animal movement modelling*. 6th Mar. 2023. DOI: 10.1101/2023.03.05.531080. URL: https://hal.umontpellier.fr/hal-04023957.

[50] C. Garcin, M. Servajean, J. Salmon and A. Joly. *A two-head loss function for deep Average-K classification*. 2023. DOI: 10.48550/arXiv.2303.18118. URL: https://inria.hal.science/hal-04204318.

[51] T. Lefort, B. Charlier, A. Joly and J. Salmon. *Peerannot: classification for crowdsourced image datasets with Python*. 2023. URL: https://hal.science/hal-04202889.

[52] Q. Leroy, O. Buisson and A. Joly. *How does the degree of novelty impacts semi-supervised representation learning for novel class retrieval?* 17th Oct. 2023. URL: https://inria.hal.science/hal-03871584.

[53] G. Morand, A. Joly, T. Rouyer, T. Lorieul and J. Barde. *Predicting species distributions in the open oceans with convolutional neural networks*. 12th Aug. 2023. DOI: 10.1101/2023.08.11.551418. URL: https://inria.hal.science/hal-04204332.

[54] R. Salles, J. Lima, R. Coutinho, E. Pacitti, F. Masseglia, R. Akbarinia, C. Chen, J. M. Garibaldi, F. A. Machado Porto and E. S. Ogasawara. *SoftED: Metrics for Soft Evaluation of Time Series Event Detection*. 2023. DOI: 10.48550/arXiv.2304.00439. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04280618.

Other scientific publications

[55] P. Leroy, B. Fruchard, T. Goyallon, G. Duprat, C. Avezou, P. Dewilde and L. Reveret. ‘Design of a Software Suite to Support Indexing, Annotating, and Analyzing Climbing Videos’. In: ECSS 2023 - 28th Congress of the European College of Sport Science. Paris, France, 4th July 2023. URL: https://hal.science/hal-04330587.