RESEARCH CENTRE

Inria Centre at Université Côte d'Azur

IN PARTNERSHIP WITH:

CNRS, Université de Montpellier

2024 ACTIVITY REPORT

Team ZENITH

Scientific Data Management

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions)

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

Inria

Contents

Team ZENITH

Creation of the Team: 2024 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]
- Reza Akbarinia [INRIA, Researcher]
- Christophe Botella [INRIA, ISFP]
- Benjamin Bourel [INRIA, Researcher]
- Alexis Joly [INRIA, Senior Researcher]
- Fabio Andre Machado Porto [INRIA, Senior Researcher, from Sep 2024 until Oct 2024]
- Maxime Ryckewaert [INRIA, Starting Research Position]
- Joseph Salmon [UNIV MONTPELLIER, Professor Detachement, from Oct 2024]
- Patrick Valduriez [INRIA, Emeritus]

Faculty Members

- Esther Pacitti [UNIV MONTPELLIER, Professor Delegation, from Sep 2024]
- Esther Pacitti [UNIV MONTPELLIER, Professor, until Aug 2024]
- Joseph Salmon [UNIV MONTPELLIER, Professor, until Sep 2024]

Post-Doctoral Fellows

- Aimie Berger Dauxere [INRAE, until Sep 2024]
- Raphael De Freitas Saldanha [INRIA, Post-Doctoral Fellow]
- Camille Garcin [INRIA, Post-Doctoral Fellow, until Jun 2024]
- Lukas Picek [INRIA, Post-Doctoral Fellow]
- Rebecca Pontes Salles [INRIA, Post-Doctoral Fellow]
- Jules Vandeputte [INRIA, Post-Doctoral Fellow, until Nov 2024]

PhD Students

- Matteo Contini [IFREMER, until Sep 2024]
- Guillaume Coulaud [UNIV MONTPELLIER, until Sep 2024]
- Lo'Aï Gandeel [INRIA, from Oct 2024]
- Cesar Leblanc [INRIA]
- Tanguy Lefort [UNIV MONTPELLIER, until Sep 2024]
- Kawtar Zaher [INA, CIFRE, until Sep 2024]

Technical Staff

- Antoine Affouard [INRIA, Engineer, until Sep 2024]
- Mathias Chouet [CIRAD, Engineer, until Sep 2024]
- Maxime Fromholtz [INRIA, Engineer]
- Hugo Gresse [INRIA, Engineer, until Sep 2024]
- Benoit Lange [INRIA, Engineer]
- Théo Larcher [INRIA, Engineer, until Oct 2024]
- Pierre Leroy [INRIA, Engineer]
- Thomas Paillot [INRIA, Engineer]
- Remi Palard [CIRAD, Engineer]
- Julien Thomazo [LIRMM, Engineer, until Sep 2024]

Interns and Apprentices

- Daniel Akbarinia [TELECOM PARIS, Intern, from Jul 2024 until Jul 2024]
- Louis Aury [INRIA, Intern, from Mar 2024 until Jun 2024]

Administrative Assistant

- Cathy Desseaux [INRIA]

External Collaborators

- Hervé Goëau [CIRAD]
- François Munoz [UGA]
- Christophe Pradal [CIRAD]

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 unended data. 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, this year, we have 15 journal publications versus 24 conference publications.

6 Highlights of the year

6.1 Awards

- Pl@ntNet, with The Brazilian Team led by ESALQ (University of Sao Paulo) won the 3rd place at Xprize Rainforest, a $10 million competition (among more than 300 teams) related to novel technologies for the monitoring of tropical biodiversity.
- Patrick Valduriez was elected AAIA Fellow (Fellow of the Asia-Pacific Artificial Intelligence Association) for contributions in data science, at the heart of data-centric AI.
- César Leblanc, Maxime Ryckewaert and Theo Larcher won the B-cubed Hackaton, Brussels April 2-5, 2024.

6.2 Other

- The Pl@ntNet database has exceeded one billion plant observations.
- Zenith participated to the CESE consultation on the impact of AI on the environment (CESE is one of the 3 assemblies of the French constitution, made up of elected representatives of civil society)
- Patrick Valduriez has been appointed Inria Alumni ambassador for Brazil, with the mission of developing the Inria Alumni network in Brazil.

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

7.2 Open data

Pl@ntNet-CrowdSWE

Contributors: Tanguy Lefort,Antoine AFFOUARD, Benjamin Charlier, Jean-Christophe Lombardo, Mathias Chouet, Hervé Goëau, Joseph Salmon, Pierre BONNET, Alexis Joly

Description: This open dataset is a subset of the full Pl@ntNet database from the extraction of the Pl@ntNet South Western Europe (SWE) crowdsourced annotations. It contains all species identification and user votes for observations made between 2017 and 2023 in the SWE flora. In total, more than 6 699 593 plant observations are labeled by 823 251 users between January 2017 and October 2023. In addition, 98 experts were selected to obtain ground truth values for 26 811 observations. The goal of this dataset is to provide a new large scale benchmark dataset for label aggregation in a setting of crowdsourcing for image classification. It has already been used to demonstrate the current performance of the Pl@ntNet label aggregation strategy in [58].

Dataset PID (DOI,...): `https://doi.org/10.5281/zenodo.10782465`

Project link: `https://zenodo.org/records/10782465`

Publications: `https://hal.science/hal-04603038`

Contact: tanguy.lefort@inria.fr

Release contributions: New release, updated for metadata modifications related to anonymity (required during the reviewing session for the related publications) and founding

Seatizen Atlas

Contributors: Matteo Contini, Julien Barde, Sylvain Bonhommeau, Victor Illien, Alexis Joly

Description: The **Seatizen Atlas** dataset is a collection of geospatial and metadata-enriched images, combining participatory science data with data collected from advanced scientific platforms. This dataset integrates imagery captured using different platforms, including kitesurfs, paddleboards, snorkeling masks, Autonomous Surface Vehicles (ASVs), and Unmanned Aerial Vehicles (UAVs).

Covering vast regions in the Southwest Indian Ocean, including Réunion Island, Seychelles, and Mauritius, the dataset comprises approximately 1,626,830 images. The images are enriched with detailed metadata, including GPS coordinates, camera settings (EXIF/XMP standards), and AI-based predictions. Additionally, 14,492 images have been manually annotated following Global Coral Reef Monitoring Network (GCRMN) and Global Biodiversity Information Facility (GBIF) standards.

This dataset supports diverse scientific and ecological applications, such as:

- Generating historical records of marine ecosystems with precise geolocation for monitoring ecosystem changes.
- Producing high-resolution species distribution maps using AI-based inference models.
- Training and evaluating computer vision algorithms on annotated subsets.
- Creating aerial and underwater photogrammetric reconstructions using accurately georeferenced imagery.
- Validating remote sensing models with ground-truth geolocated data.
- Promoting FAIR (Findable, Accessible, Interoperable, Reusable) data practices through adherence to established standards (e.g., Darwin Core, GCRMN, EXIF/XMP, and COCO).

The data workflow, including image processing, AI-based predictions, and publishing methodologies, is openly shared and reproducible through the GitHub repository: https://github.com/SeatizenDOI.

Dataset PID (DOI,...): 10.5281/zenodo.13951435

Project link: https://seatizenmonitoring.ifremer.re

Publications: Seatizen atlas: a collaborative dataset of underwater and aerial marine imagery. Soon available at https://doi.org/10.1038/s41597-024-04267-z.

Contact: seatizen.ifremer@gmail.com, sylvain.bonhommeau@ifremer.fr, alexis.joly@inria.fr

Release contributions:

- **seatizen_atlas_db.gpkg**: Geopackage for spatial data management.
- **session_doi.csv**: List of published sessions.
- **metadata_images.csv**: Metadata for published images.
- **metadata_multilabel_predictions.csv**: Predictions from the AI model.
- **metadata_multilabel_annotation.csv**: Annotated subset of images.
- **seatizen_atlas.qgz**: QGIS project file.
- **darwincore_multilabel_annotation.zip**: Darwin Core Archive with image annotations.

8 New results

8.1 Distributed Data and Model Management

8.1.1 Data-Driven Model Selection for Spatio-Temporal Prediction

Participants: Fabio Porto, Patrick Valduriez.

Spatio-temporal predictive queries encompass a spatio temporal constraint, which defines a region and a target variable, and an evaluation metric. They produce the future values for the target variable computed by predictive models at each point of the spatio-temporal region. However, training temporal models at each spatial domain point can be prohibitive. In [55], we propose a data-driven approach for selecting pre-trained temporal models to be applied at each query point, which avoids training a different model for each domain point, thus saving model training time. We propose a technique to decide on the best-trained model to be applied to a point for prediction. The experimental evaluation on a case study in temperature forecasting using historical data and auto-regressive models shows that, compared to the baseline, the approach achieves equivalent predictive performance at a fraction of the total computational cost.

8.1.2 Executing Scientific Workflows with Privacy Constraints

Participants: Esther Pacitti.

Containerized and cloud environments are ideal for running scientific workflows because they offer a flexible and easily instantiated setting. Data dispersion and encryption can be adopted in this context, but not independently of workflow scaling, as this can increase the total execution time or the associated financial cost. In [49], we introduce Okinawa, a heuristic for executing workflows in containerized environments with confidentiality constraints. In [50], we introduce CYCLOPS, an approach that aims to execute workflows in cloud computing environments efficiently while considering the confidentiality constraints of the produced data and the workflow structure.

8.1.3 Federated Learning

Participants: Patrick Valduriez.

Federated Learning (FL) enables collaborative model training on distributed data over edge devices. We contributed to FL in two major ways. The first one [42] is the Asynchronous Efficient Decentralized FL framework (AEDFL) for heterogeneous devices. Extensive experimentation on four public datasets and four models demonstrates the strength of AEDFL in terms of accuracy (up to 16.3% higher), efficiency (up to 92.9% faster), and computation costs (up to 42.3% lower). The second contribution [43] is a Fisher information-based efficient curriculum federated learning framework (FibecFed) to fine-tune Large Language Models (LLMs), which comes with two novel methods, i.e., adaptive federated curriculum learning and efficient sparse parameter update. Extensive experimental results based on 10 datasets demonstrate that FibecFed yields excellent performance (up to 45.35% in terms of accuracy) and superb fine-tuning speed (up to 98.61% faster) compared with 17 baseline approaches.

8.1.4 Optimal Checkpointing for Heterogeneous Chains: How to Train Deep Neural Networks with Limited Memory

Participants: Alena Shilova, Alexis Joly.

This work was done in collaboration with the HiePacs Inria team.

Training in Feed Forward Deep Neural Networks is a memory-intensive operation which is usually performed on GPUs with limited memory capacities. This may force data scientists to limit the depth of the models or the resolution of the input data if data does not fit in the GPU memory. The re-materialization technique, whose idea comes from the checkpointing strategies developed in the Automatic Differentiation literature, allows data scientists to limit the memory requirements related to the storage of intermediate data (activations), at the cost of an increase in the computational cost. In [15], we introduce a new strategy of re-materialization of activations that significantly reduces memory usage. It consists in selecting which activations are saved and which activations are deleted during the forward phase, and then recomputing the deleted activations when they are needed during the backward phase. We propose an original computation model that combines two types of activation savings: either only storing the layer inputs, or recording the complete history of operations that produced the outputs. We prove that finding the optimal solution is NP-hard and that classical techniques from Automatic Differentiation literature do not apply. Thus, we propose a weak memory persistence property and provide a Dynamic Program to compute the optimal sequence of computations. Through extensive experiments, we show that our implementation consistently outperforms existing re-materialization approaches for a large class of networks, image sizes and batch sizes.

8.2 Data Analytics

8.2.1 Event Detection in Time Series

Participants: Esther Pacitti, Fabio Porto, Rebecca Salles.

Event detection in times series is a basic function in surveillance and monitoring systems and has been extensively explored over the years. The book [54] provides a general taxonomy for event detection according to the specific event types: anomaly detection, change-point, and motif discovery. It discusses state-of-the-art metric evaluations for event detection methods and on online event detection, including the challenges of incremental and adaptive learning.

We also proposed a time series anomaly detection method [44], based on the Fast Fourier Transform (FFT). The method removes low-frequency components, such as trends and seasonality, which represent the normal behavior of the series, while preserving high-frequency components associated with anomalies. The experimental results show the effectiveness of the method in anomaly detection using high-pass FFT filters that have a cutoff frequency adjusted by change points.

Finally, we addressed the problem of matching detections to events in time series [48], by structuring the association problem using graph theory, as a bipartite graph matching problem and using the Hungarian algorithm as solution. The results demonstrate the effectiveness of the proposed approach, highlighting the impact of improvements in the associations between detections and events.

8.2.2 Metrics for Evaluating Event Detection in Time Series

Participants: Reza Akbarinia, Florent Masseglia, Esther Pacitti, Fabio Porto, Rebecca Salles.

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 [29], 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.

8.2.3 Subset Models for Multivariate Time Series Forecast

Participants: Reza Akbarinia, Raphael De Freitas Saldanha, Patrick Valduriez.

Multivariate time series find extensive applications in conjunction with machine learning methodologies for scenario forecasting across various domains. Nevertheless, certain domains exhibit inherent complexities and diversities, which detrimentally impact the predictive efficacy of global models. In [34], we introduce a modeling framework that accommodates shared feature characteristics and regional variations across diverse units, offering cost-effective training and robust prediction capabilities. The approach involves the following steps: (1) identifying subsets within the dataset that exhibit similar covariate patterns and training models specific to each subset; (2) mapping incoming samples to the appropriate subset based on the similarity between data distributions; and (3) using the model trained on the subset data for prediction.

8.2.4 One Health Data Analytics

Participants: Benoit Lange, Reza Akbarinia, Florent Masseglia, Christophe Pradal.

Many scientific applications require a "One Health" approach to data analytics that integrates data, models and tools from different application domains. We used this approach in two applications: antimicrobial resistance and reduction of pesticide usage.

Antimicrobial resistance (AMR) presents critical health risks for humans, animals, and the environment. In [39], we introduce the PROMISE platform, for managing and analyzing One Health data. It supports 25 academic networks and 42 partners, resolving data heterogeneity through built-in interoperability, allowing users to retain their preferred taxonomies. The platform facilitates cross-disciplinary insights into AMR, showcasing its capability to streamline and unify diverse datasets for effective analysis.

Reducing pesticide use in agricultural areas requires joint consideration of the ecological, economic and social components that contribute to this goal. This requires a "One Health" approach, integrating data, models and simulation tools at various scales, from the plant to the farm to the territory. In [27], we introduce the TRAVERSEES project, for designing a socio-ecosystem model used to simulate agricultural practice trajectories across territories, which subsequently serve as a tool for prospective analysis with stakeholders in the Barrois region (Grand-Est). The project employs a range of methodologies and engages in a transdisciplinary partnership. This approach leads to the emergence of multiple, innovative proposals for territorial transition levers and highlighted the diverse factors considered by farmers, along with varying degrees of sensitivity to these influences.

8.3 Machine Learning for Biodiversity and Agroecology

8.3.1 AI-Based Species Distribution Modeling and Mapping

Participants: Christophe Botella, Benjamin Deneu, Joaquim Estopinan, Alexis Joly, Théo Larcher, César Leblanc, François Munoz, Lukáš Picek.

Although increasing threats on biodiversity are now widely recognized, there are no accurate global maps showing whether and where species assemblages are at risk. In [21], we introduce a new Deep Species Distribution Model trained on 1M occurrences of 14K orchid species to predict their assemblages at global scale and at kilometre resolution. We propose two main indicators of the conservation status of the assemblages: (i) the proportion of threatened species, and (ii) the status of the most threatened species in the assemblage. We show and analyze the variation of these indicators at World scale and in relation to currently protected areas in Sumatra island. Global and interactive maps available online show the indicators of conservation status of orchid assemblages, with sharp spatial variations at all scales. In [46] (Neurips 2024 Dataset and Benchmarks), we designed and developed a new European-scale

dataset and benchmark for Multimodal SDMs at high spatial resolution (10–50m), including more than 10k species (i.e., most of the European flora). It comprises 5M heterogeneous Presence-Only records and 90k exhaustive Presence-Absence survey records, all accompanied by diverse environmental rasters (e.g., elevation, human footprint, and soil), Sentinel-2 RGB and NIR satellite images with 10 m resolution, 20-year time series of climatic variables, and satellite time series from the Landsat program. In addition to the data, we provide an openly accessible benchmark (hosted on Kaggle), which has already attracted an active community and a set of strong baselines for single predictor/modality and multimodal approaches. All resources, e.g., the dataset, pre-trained models, and baseline methods (in the form of notebooks), are available on Kaggle, allowing one to start with the dataset literally with two mouse clicks. In [26], we propose a new method that leverages convolutional neural networks (CNNs) to capture spatial features of environmental variables in the open ocean. We considered 38 taxa comprising pelagic fishes, elasmobranchs, marine mammals, marine turtles and birds. We trained a model to predict probabilities from the environmental conditions at any specific point in space and time, using species occurrence data from the Global Biodiversity Information Facility (GBIF) and environmental data from various sources. These variables included sea surface temperature, chlorophyll concentration, salinity and fifteen others. The classifier accurately predicted the observed taxon as the most likely in 69% of cases and included the observed taxon among the top three most likely predictions in 89% of cases. Additionally, this purely correlative model was then analyzed with explicability tools to understand which variables had an influence on the model's predictions.

8.3.2 Species-to-Habitats Classification

Participants: César Leblanc, Alexis Joly.

In [22], we worked on a deep-learning framework for enhancing habitat identification based on species composition. We leveraged deep-learning techniques, such as transformers and tested different network architectures, feature encodings, hyperparameter tuning and noise addition strategies to identify the optimal model. Our best algorithm, applied to European habitat types, significantly improved habitat classification accuracy, achieving a more than twofold improvement compared to the previous state-of-the-art (SOTA) method on an external data set, clearly outperforming expert systems. The accuracy score on a database containing hundreds of thousands of standardized presence/absence European surveys reaches 88.74%, as assessed by expert judgment. Finally, our results showcase that species dominance is a strong marker of ecosystems and that the exact cover abundance of the flora is not required to train neural networks with predictive performances. The framework we developed can be used by researchers and practitioners to accurately classify habitats. This work was co-authored with 25 other research labs who provided data (in the context of the European Vegetation Archive), developed the initial expert system and contributed ideas and writing improvements.

8.3.3 Unseen Plant Disease Recognition

Participants: Alexis Joly.

This work was done in the context of a collaboration with Swinburne University of Technology.

Automatic plant disease recognition has a great potential for improving agricultural productivity and sustainability by providing timely, accurate, and scalable solutions for managing plant health. One of the key challenge, however, is the impracticality of collecting samples for all diseases and every plant species. In [18] and [40], we worked on the recognition of unseen plant diseases. We propose new models capable of disantangling the pathogen's characteristic information from that of the host plant, so that it can be recombined for any plant/disease pair. In [18], we introduce a Cross Learning Vision Transformer (CL-ViT) model, incorporating self-supervised learning, in contrast to the previous state-of-the-art, FF-ViT, which emphasizes conceptual feature disentanglement with a synthetic feature generation framework. In

[40], we introduce a novel approach that incorporates textual data to guide the vision model in learning the representation of multiple plants and diseases.

8.3.4 Evaluation of Species Identification and Prediction Algorithms

Participants: Alexis Joly, Lukáš Picek, Hervé Goëau, Christophe Botella, Benjamin Deneu, Diego Marcos, Joaquim Estopinan, César Leblanc, Théo Larcher.

We ran a new edition of the LifeCLEF evaluation campaign [36, 37] with the involvement of hundreds of data scientists and research teams worldwide. It delivers a unique view of state-of-the-art performance on species identification and prediction problems, thanks to realistic datasets and controlled evaluation methodologies. One of the main outcome was that domain shift problems remain a major problem for the emergence of new techniques, such as passive acoustic sensors, HD images of plant cover, or remote sensing monitoring. The lack of annotated data for these new domains considerably hinders the progress of supervised methods, and alternative cross-domain methods are struggling to emerge. A great hope may lie in the use of unlabeled data, which will become increasingly available and whose use for domain adaptation or self-supervised learning is beginning to emerge as an effective solution (notably in BirdCLEF [38] and GeoLifeCLEF [45]). Another very promising prospect is multi-modal model learning, which was the key to the success of the best methods for the GeoLifeCLEF challenge and has enabled improvements in other tasks, including FungiCLEF [47] and PlantCLEF [35]. As far as model architectures are concerned, there is a wide disparity between the use of large-scale foundation models such as DinoV2 in PlantCLEF, SnakeCLEF, and FungiCLEF and a certain trend towards frugal architectures in GeoLifeCLEF, FungiCLEF, SnakeCLEF, and BirdCLEF. Finally, it is important to note the strength of collaborative work in the progression of the challenges. The sharing of knowledge, models, or codes, whether by the organizers or the participants themselves, has a direct impact on their subsequent developments and promotes co-construction rather than sole competition.

8.3.5 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 2024.
First, the web front-end of the application has been deeply refactorized with strong UI and UX improvements (thanks to the recruitment of an engineer expert in this domain). Most views of the application were revisited, necessitating substantial development work, including on the back-end, as numerous new data access functionalities had to be developed, impacting all layers (database, data access layer, API). Pl@ntNet web's front-end is used by about 10M users per year.
We also continued the development of new tools for managing and analyzing vegetation plot images (e.g. drone images, quadrat images, road-side views, etc.). We first optimized our tiling approach for the analysis of such HD pictures (in pytorch). This allowed to divide the computation cost by a factor 5. We also enriched the API to make this service more easily usable by external partners and we finalized the development of the web front-end thanks to the feedbacks of beta testers along the year. Those tools are already used by several projects/organizations including: (i) a Biodiversa+ pilot project on the monitoring of invasive alien species (coordinated by the Danish Ministry of the Environment through Aarhus University), (ii) two international consortia who participated to the X-prize rain forest challenge (who obtained the second and third place over 300 participants), (iii) the French project Pl@ntAgroEco.
A third key feature developed in 2024 is the possibility to manage and identifiy other plant attributes than just the species through a mechanism called "tags" (key-valued). In particular, it enables the identification

and management of two key concepts for agro-ecology: (i) plant damages (pathogens, diseases, fungies, eyc.) and (ii), plant varieties (sub-specific levels in the taxonomy such as crop varities, horticicultural varieties). This required in-depth work on the database and the data access layer. The integration of these new features into the web and mobile front-ends is still under development.
A last new feature developed as part of the GUARDEN EU project is to enable the import of observations by batch within the database. Indeed, many institutions and private individuals have large batches of plant observations that they would like to share on the platform. A new interface has therefore been developed in the Pl@ntNet web application.

9 Partnerships and cooperations

9.1 International initiatives

HPDaSc

Title: High Performance Data Science

Duration: 2020 ->

Coordinator: Fabio Porto (fporto@lncc.br)

Partners:

- Laboratório Nacional de Computação Científica Petrópolis (Brésil)

Inria contact: Patrick Valduriez

Summary: Data-intensive science refers to modern science, such as astronomy, geoscience or life science, where researchers need to manipulate and explore massive datasets produced by observation or simulation. It requires the integration of two fairly different paradigms: high-performance computing (HPC) and data science. We address the following requirements for high-performance data science (HPDaSc): Support realtime analytics and visualization (in either in situ or in transit architectures) to help make high-impact online decisions; Combine ML with analytics and simulation, which implies dealing with uncertain training data, autonomously built ML models and combine ML models and simulation models; Support scientific workflows that combine analytics, modeling and simulation, and exploit provenance in realtime and HIL (Human in the Loop) for efficient workflow execution.

To address these requirements, we will exploit new distributed and parallel architectures and design new techniques for ML, realtime analytics and scientific workflow management. The architectures will be in the context of multisite cloud, with heterogeneous data centers with data nodes, compute nodes and GPUs. We will validate our techniques with major software systems on real applications with real data. The main systems will be OpenAlea and Pl@ntnet from Zenith and DfAnalyzer and SAVIME from the Brazilian side. The main applications will be in agronomy and plant phenotyping (with plant biologists from CIRAD and INRA), biodiversity informatics (with biodiversity scientists from LNCC and botanists from CIRAD), and oil & gas (with geoscientists from UFRJ and Petrobras).

9.2 International research visitors

9.2.1 Visits of international scientists

Inria International Chair

Participants: Reza Akbarinia, Alexis Joly, Patrick Valduriez.

Fabio Porto, Laboratório Nacional de Computação Científica (LNCC, Brasil), holds an Inria International Chair for 12 months (from January 1, 2024 to December 31, 2028).

Work on Gypscie, a framework that supports the entire ML lifecycle and manages all ML artifacts (metadata, datasets, models), enabling reuse and import from ML frameworks.

Other international visits to the team

Dennis Shasha

Status Researcher

Institution of origin: University of New-York

Country: USA

Dates: April 17 - June 7

Context of the visit: Projects MAMBO and PROMISE

Mobility program/type of mobility: research stay, lecture

Eduardo Ogasawara

Status Researcher

Institution of origin: Department of Computer Science at the Federal Center for Technological Education of Rio de Janeiro (CEFET/RJ)

Country: Brasil

Dates: June 1-9, and September 30 - October 4

Context of the visit: Own resource of CEFET

Mobility program/type of mobility: research stay, lecture

9.3 European initiatives

9.3.1 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
- YPOURGEIO GEORGIAS, AGROTIKIS ANAPTYXIS KAI PERIVALLONTOS (MINISTRY OF AGRICULTURE, RURAL DEVELOPMENT AND ENVIRONMENT OF CYPRUS), Cyprus
- DREVEN SRL, Belgium
- PLYMOUTH MARINE LABORATORY LIMITED (PML), United Kingdom
- UNIVERSITY OF ANTANANARIVO, Madagascar
- CHAROKOPEIO PANEPISTIMIO (HAROKOPIO UNIVERSITY OF ATHENS (HUA)), Greece

- INSTITUT METROPOLI (BARCELONA INSTITUTE OF REGIONAL AND METROPOLITAN STUDIES), Spain
- 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 AND 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.

9.4 National initiatives

Pl@ntAgroEco (PEPR Agroécologie et 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. The project is leaded by ZENITH (Alexis Joly).

FishPredict (ANR), (2022-2025), 500 Keuro.

Participants: Benjamin Bourel, Alexis Joly, Maximilien Servajean, Julien Thomazo.

Fish-predict ANR project funded in the context of the IA-Biodiv challenge. The projects aims at predicting the biodiversity of reef fishes using AI technologies. Alexis Joly is co-leading of the whole project jointly with David Mouillot, marine ecologist at the MARBEC lab.

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 analyzing 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.

PEPR agroécologie et numérique "RootSystemTracker" (2024-2027), 144 Keuros

Participants: Reza Akbarinia, Christophe Pradal.

Roots play a crucial role in nutrient and water uptake, atmospheric carbon fixation, and soil interactions, significantly influencing resource use efficiency and crop resilience to environmental stresses. The objective of the RootSystemTracker project is to develop efficient methods for the spatio-temporal phenotyping of plant root architectures using heterogeneous data. This involves automatically capturing their topology and geometry over time, despite challenges such as root occlusions and variability in observation conditions.

9.4.1 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.

9.5 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.

9.6 Public policy support

CESE consultation on the impact of AI on the environment : The CESE (Conseil Economique, Social et Envinonemental) is one of the 3 assemblies of the French constitution, made up of elected representatives of civil society (unions, associations, companies, students, etc.). Its role is to provide advice on economic, social and environmental policies to guide public decision-making (governmental in particular). Alexis Joly took part in the consultation entitled “Impacts of artificial intelligence: risks and opportunities for the environment”. He was consulted and interviewed on several occasions and was one of the 3 experts invited to the final plenary session that voted on the recommendations.

OCDE report the advancement of the productivity of science with citizen science and artificial OCDE report on the advancement of the productivity of science with citizen science and artificial)

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- P. Valduriez:
  - Inria-Brasil (hybrid) workshop on Artificial Intelligence and Applications, LNCC, Petropolis, Rio de Janeiro (April 11, 2024)
  - Inria-Brasil (hybrid) workshop on Digital Science and Agronomy, Inria Montpellier (September 10-11, 2024)
- A. Joly:
  - Pl@ntAgroEco summer school 2024, Montpellier (July 10-11, 2024)
  - LifeCLEF 2024 workshop, Grenoble (September 9, 2024)

Member of the conference program committees

- R. Akbarinia: ECML-PKDD 2024, IEEE BigData 2024, AIMLSystems 2024.
- E. Pacitti: BDA 2024.
- P. Valduriez: Ph.D. thesis award committee, BDA 2024.
- F. Masseglia: ECML PKDD 2024, Discovery Science 2024, AIxDKE 2024, ACM SAC 2024, PAKDD 2024, EGC 2024.

10.1.2 Journal

Editor, Associate editor

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

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

- A. Joly: Computers and Electronics in Agriculture journal, EURASIP Journal on Audio Speech and Music Processing, Ecological Informatics
- F. Masseglia: Data mining and Knowledge Discovery; Journal of Machine Learning Research.

10.1.3 Invited talks

- P. Valduriez:
  - "Data Science and Innovation"
  - IMPA, Rio de Janeiro, May 5, 2024
  - IRIT, Toulouse, July 12, 2024
  - LNCC, Petropolis, Rio de Janeiro, July 22, 2024
  - Keynote, SSDBM Conference, Rennes, July 12, 2024

  - Keynote, SBBD Conference, Florianopolis, Brazil, October 14, 2024

- A. Joly

  - Invited keynote speaker at EGC 2024, Dijon, January 22-26, 2024
  - Invited speaker at a symposium at Collège de France entitled "Solutions to Monitor Plants, Pollinators and Their Interactions in a Changing World", Paris, May 23, 2024.
  - Invited panelist at a plenary session of the CESE (Conseil Economique, Social et Environnemental), Paris, September 23, 2024 (replay).

10.1.4 Leadership within the scientific community

- E. Pacitti: Member of the Steering Committee of the BDA conference.
- R. Akbarinia: Member of the Steering Committee of the BDA conference.
- : A. Joly

  - Scientific and Technical director of Pl@ntNet platform
  - Coordinator of the LifeCLEF international virtual lab

10.1.5 Scientific expertise

- C. Pradal: member of the INRAE evaluation committee CSS (Scientific Specialist Commission) in Plant Integrated Biology
- R. Akbarinia: member of the evaluation committee (section 27) of University of Montpellier.
- A. Joly:

  - GENCI expert committee (AI thematic)
  - AI2050 Early Career Fellowship

- P. Valduriez: consultant on big data for the Software Heritage project
- F. Masseglia: Inria International Chair selection committee

10.1.6 Research administration

- F. Masseglia: deputy scientific director of Inria for the domain "Perception, Cognition And Interaction", 50% of his time.
- R. Akbarinia: Scientific referent for research data at Inria branch of Montpellier; Member of Inria national commission for research data.
- 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

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.

10.2.2 Supervision

PhD & HDR:

- 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.
- PhD in progress: Loaï Gandeel, Automatic methods for spatio-temporal reconstruction of root architecture. Advisors: Reza Akbarinia, Romain Fernandez, Christophe Pradal.
- PhD in progress: Raphaël Benerradi, species trends estimation from citizen science data. Advisors: Christophe Botella, Alexis Joly, Maximilien Servajean
- PhD in progress: Théo Larcher, multi-scale species prediction. Advisors: Alexis Joly, Joseph Salmon, Pierre Bonnet, Marijn Van der Velde
- PhD in progress: Sébastien Gigot-Leandri, decision-oriented site occupancy models. Advisors: Alexis Joly, Maximilien Servajean, David Mouillot

10.2.3 Juries

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

- R. Akbarinia (reviewer): Qi Fan, École Polytechnique.
- F. Masseglia: Anton Dolhopolov, Université de Montpellier.
- A. Joly: (i) as reviewer: Salim Khazem (Centrale supelec, 2024), Odile Peyron (University of Montpellier, 2024), (ii) as president of the jury: Letizia Lamperti (University of Montpellier, 2024), Ilyass Moummad (IMT Atlantique, 2024)

10.3 Popularization

10.3.1 Specific official responsibilities in science outreach structures

- F. Masseglia: member of the strategic committee of Fondation Blaise Pascal
- A. Joly: member of the steering committee of Pl@ntNet citizen science platform

10.3.2 Productions (articles, videos, podcasts, serious games, ...)

- P. Valduriez
  - Interview Le Monde Informatique: "L'IA va-t-elle remplacer SQL"
  - Chiche (4 actions): Lycée La Mercy, Lycée Mermoz, Montpellier
- Hugo Gresse, Thomas Paillot: fête de la nature, Zoo de Lunaret, May 25, 2024
- A. Joly
  - Citizen science program La flore des cultures sous l'objectif de Pl@ntNet
  - Shooting of a 50-minute documentary for France 5 on gardening today (directed by Jean-Christophe Chatton, to be broadcasted in 2025)
  - Pl@ntNet's donation campaign 2024
- J. Salmon
  - fête de la science at university of Montpellier (carasciences)
  - bar des sciences, Faut il avoir peur de l'IA ?

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] M. Metz, M. Lesnoff, F. Abdelghafour, R. Akbarinia, F. Masseglia and J.-M. Roger. 'A "big-data" algorithm for KNN-PLS'. In: *Chemometrics and Intelligent Laboratory Systems* 203 (Aug. 2020), p. 104076. DOI: 10.1016/j.chemolab.2020.104076. URL: https://hal.inrae.fr/hal-02899789.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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

[15] O. Beaumont, L. Eyraud-Dubois, J. Herrmann, A. Joly and A. Shilova. 'Optimal checkpointing for heterogeneous chains: how to train deep neural networks with limited memory'. In: *ACM Transactions on Mathematical Software* (2024). URL: https://inria.hal.science/hal-02352969. In press (cit. on p. 14).

[16] C. Botella, P. Gaüzère, L. O'Connor, J. Renaud, Y. Dou, C. Graham, P. Verburg, L. Maiorano and W. Thuiller. 'Don't bite the hand that feeds you: Meta food webs help in the face of the Eltonian shortfall'. In: *Global Change Biology* 30.6 (7th June 2024), e17359. DOI: 10.1111/gcb.17359. URL: https://hal.science/hal-04757080.

[17] S. Bouzaouia, M. Ryckewaert, D. Héran, A. Ducanchez and R. Bendoula. 'Using Dynamic Laser Speckle Imaging for Plant Breeding: A Case Study of Water Stress in Sunflowers'. In: *Sensors* 24 (14th Aug. 2024). DOI: 10.3390/s24165260. URL: https://hal.inrae.fr/hal-04675578.

[18] A. Y. H. Chai, S. H. Lee, F. S. Tay, P. Bonnet and A. Joly. 'Beyond supervision: Harnessing self-supervised learning in unseen plant disease recognition'. In: *Neurocomputing* 610 (Dec. 2024), p. 128608. DOI: 10.1016/j.neucom.2024.128608. URL: https://inria.hal.science/hal-04815172 (cit. on p. 16).

[19] J. Diaz-Olivares, R. Bendoula, W. Saeys, M. Ryckewaert, I. Adriaens, X. Fu, M. Pastell, J.-M. J. .-. Roger and B. Aernouts. 'PROSAC as a selection tool for SO-PLS regression: A strategy for multi-block data fusion'. In: *Analytica Chimica Acta* 1319 (Aug. 2024), p. 342965. DOI: 10.1016/j.aca.2024.342965. URL: https://hal.inrae.fr/hal-04872776.

[20] N. Elvekjaer, L. Martínez-Sanchez, P. Bonnet, A. Joly, M. L. Paracchini and M. van Der Velde. 'Detecting flowers on imagery with computer vision to improve continental scale grassland biodiversity surveying'. In: *Ecological Solutions and Evidence* 5 (23rd May 2024). DOI: 10.1002/2688-8319.12324. URL: https://hal.inrae.fr/hal-04593804.

[21] J. Estopinan, M. Servajean, P. Bonnet, A. Joly and F. Munoz. 'Mapping global orchid assemblages with deep learning provides novel conservation insights'. In: *Ecological Informatics* 81 (1st July 2024), p. 102627. DOI: 10.1016/j.ecoinf.2024.102627. URL: https://hal.inrae.fr/hal-04581266 (cit. on p. 15).

[22] C. Leblanc, P. Bonnet, M. Servajean, M. Chytrý, S. Aćić, O. Argagnon, A. Bergamini, I. Biurrun, G. Bonari, J. A. Campos, R. Ćušterevska, A. Čarni, M. de Sanctis, J. Dengler, E. Garbolino, V. Golub, U. Jandt, F. Jansen, M. Lebedeva, J. R. M. H. Lenoir, J. E. Moeslund, A. Pérez-Haase, R. Pielech, J. Šibík, Z. Stančić, A. Stanisci, G. Swacha, D. Uogintas, K. Vassilev, T. Wohlgemuth and A. Joly. 'A deep-learning framework for enhancing habitat identification based on species composition'. In: *Applied Vegetation Science* 27.3 (28th Aug. 2024), e12802. DOI: 10.1111/avsc.12802. URL: https://hal.science/hal-04700157 (cit. on p. 16).

[23] S. H. Lee, Z. R. Liaw, Y. H. Chai, S. L. Ng, P. Bonnet, H. Goëau and A. Joly. 'Revolutionizing Plant Pathogen Conservation: The Past, Present, and Future of AI in Preserving Natural Ecosystems'. In: *Biodiversity Information Science and Standards* 8 (25th July 2024), e133055. DOI: 10.3897/biss.8.133055. URL: https://hal.inrae.fr/hal-04701227.

[24] T. Lefort, B. Charlier, A. Joly and J. Salmon. 'Identify ambiguous tasks combining crowdsourced labels by weighting Areas Under the Margin'. In: *Transactions on Machine Learning Research Journal* (2024). URL: https://hal.science/hal-03812716.

[25] T. Lefort, B. Charlier, A. Joly and J. Salmon. 'Peerannot: classification for crowdsourced image datasets with Python'. In: *Computo* (2024). DOI: 10.57750/qmaz-gr91. URL: https://hal.science/hal-04202889.

[26] G. Morand, A. Joly, T. Rouyer, T. Lorieul and J. Barde. 'Predicting species distributions in the open ocean with convolutional neural networks'. In: *Peer Community Journal* 4 (2024), e93. DOI: 10.24072/pcjournal.471. URL: https://hal.umontpellier.fr/hal-04754491 (cit. on p. 16).

[27] C. Robert, F. Accatino, A. Barbe, C. Bedos, P. Benoit, C. Bertrand, A. Bourceret, T. Da Costa, M. Dahirel, C. Fournier, T. Griessinger, L. Gauthier, L. Grohens, F. Honore, M. Jacob, J. Lecomte, L. Martin, E. Meunier, P.-A. Précigout, C. Pradal, J.-E. Rougier and P. Smith. 'TRAVERSéES: Territorial levers and transition pathways for reducing pesticide use'. In: *Innovations Agronomiques* 96 (2024), pp. 25–37. DOI: 10.17180/ciag-2024-Vol96-art03-GB. URL: https://hal.inrae.fr/hal-04810884 (cit. on p. 15).

[28] R. Saldanha, R. Akbarinia, M. Pedroso, V. Ribeiro, C. Cardoso, E. Peña, P. Valduriez and F. Porto. 'Zonal statistics datasets of climate indicators for Brazilian municipalities'. In: *Environmental Data Science* 3 (8th Feb. 2024), e2. DOI: 10.1017/eds.2024.3. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04447150.

[29] 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'. In: *Computers & Industrial Engineering* (2025). DOI: 10.1016/j.cie.2024.110728. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04280618. In press (cit. on p. 14).

International peer-reviewed conferences

[30] V. Berry, A. Castelltort, B. Lange, J. Teriihoania, C. Tibermacine and C. Trubiani. 'Is it Worth Migrating a Monolith to Microservices? An Experience Report on Performance, Availability and Energy Usage'. In: *IEEE Xplore*. ICWS 2024 - IEEE International Conference on Web Services. 2024 IEEE International Conference on Web Services (ICWS). Shenzhen, China: IEEE, 7th July 2024, pp. 944–954. DOI: 10.1109/ICWS62655.2024.00112. URL: https://inria.hal.science/hal-04781943.

[31] E. Coindre, L. Chir, M. Ryckewaert, R. Boulord, M. Falcon, T. Laisné, G. Rolland, M. Lis, L. Cabrera-Bosquet, A. Doligez, T. Simonneau, B. Pallas, A. Coupel-Ledru and V. Segura. 'Diversity of leaf functioning under water deficit in a large grapevine panel: high throughput phenotyping and genetic analyses'. In: *IVES Conference Series*. Open GPB 2024 - Open Conference on Grapevine Physiology and Biotechnology. Logrono La Riora, Spain, 7th July 2024. DOI: 10.58233/FEmcETUr. URL: https://hal.inrae.fr/hal-04677545.

[32] E. Coindre, M. Ryckewaert, L. Chir, R. Boulord, M. Falcon, T. Laisné, G. Rolland, V. Bouckenooghe, M. Lis, L. Cabrera-Bosquet, A. Doligez, T. Simonneau, V. Freitas, M. Thomas, B. Pallas, A. Coupel-Ledru and V. Segura. 'NIRS as a high-throughput phenotyping tool for assessing the diversity of leaf functioning under water deficit in a large grapevine panel'. In: *Hélio SPIR*. 25èmes rencontres HélioSPIR. Montpellier, France, 2024, p. 23. URL: https://hal.inrae.fr/hal-04702511.

[33] L. M. Estupinan Suarez, L. Abraham, T. Adriaens, L. Breugelmans, D. A. Clarke, P. Desmet, S. Dove, K. T. Faulkner, M. Fernandez, L. A. Hendrickx, C. Hui, A. Joly, S. Kumschick, W. Langeraert, M. Martini, J. Miller, D. Oldoni, H. Pereira, C. Preda and Q. Groom. 'Biodiversity Data Cubes for Cross-Cutting Science and Policy'. In: *EGU 2024 abstract*. EGU 2024 - General Assembly of European Geosciences Union. Vienna, Austria, 27th Nov. 2024. DOI: 10.5194/egusphere-egu24-6353. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04834859.

[34] R. de Freitas Saldanha, V. Ribeiro, E. Peña, M. Pedroso, R. Akbarinia, P. Valduriez and F. Porto. 'Subset Models for Multivariate Time Series Forecast'. In: *IEEE Xplore*. ICDEW 2024 - IEEE 40th International Conference on Data Engineering Workshops. International Conference on Data Engineering Workshops. Utrecht, Netherlands, 13th May 2024, pp. 86–90. DOI: 10.1109/ICDEW61823.2024.00016. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04711300 (cit. on p. 15).

[35] H. Goëau, V. Espitalier, P. Bonnet and A. Joly. 'Overview of PlantCLEF 2024: multi-species plant identification in vegetation plot images'. In: CLEF 2024 Working Notes - 25th Conference and Labs of the Evaluation Forum. Vol. 3740. CEUR worshop proceedings 187. Grenoble, France, 9th Sept. 2024, pp. 1978–1988. URL: https://hal.inrae.fr/hal-04806900 (cit. on p. 17).

[36] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, M. Šulc, M. Hrúz, M. Servajean, J. Matas, H. Glotin, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, A. Durso, I. Eggel, P. Bonnet and H. Müller. 'LifeCLEF 2024 Teaser: Challenges on Species Distribution Prediction and Identification'. In: *Lecture notes in computer science*. ECIR 2024 - 46th European Conference on Information Retrieval. Vol. LNCS-14613. Advances in Information Retrieval. ECIR 2024 - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part VI. Glasgow, United Kingdom: Springer Nature Switzerland, 20th Mar. 2024, pp. 19–27. DOI: 10.1007/978-3-031-56072-9_3. URL: https://hal.inrae.fr/hal-04667635 (cit. on p. 17).

[37] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, M. Šulc, M. Hrúz, M. Servajean, H. Glotin, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet and H. Müller. 'Overview of LifeCLEF 2024: Challenges on Species Distribution Prediction and Identification'. In: *Lecture notes in computer science*. CLEF 2024 - 15th International Conference of the Cross-Language Evaluation Forum for European Languages. Vol. LNCS-14959. Experimental IR Meets Multilinguality, Multimodality, and Interaction : 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II. Grenoble, France: Springer Nature Switzerland, 19th Sept. 2024, pp. 183–207. DOI: 10.1007/978-3-031-71908-0_9. URL: https://inria.hal.science/hal-04830385 (cit. on p. 17).

[38] S. Kahl, T. Denton, H. Klinck, V. Ramesh, V. Joshi, M. Srivathsa, A. Anand, C. Arvind, H. Cp, S. Sawant, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué and A. Joly. 'Overview of BirdCLEF 2024: Acoustic Identification of Under-studied Bird Species in the Western Ghats'. In: CLEF 2024 Working Notes - 25th Conference and Labs of the Evaluation Forum. Vol. 3740. CEUR workshop proceedings. Grenoble, France, 9th Sept. 2024, pp. 1948–1957. URL: https://hal.inrae.fr/hal-04719578 (cit. on p. 17).

[39] B. Lange, R. Akbarinia and F. Masseglia. 'A One-Health Platform for Antimicrobial Resistance Data Analytics'. In: CIKM 2024 - 33rd ACM International Conference on Information and Knowledge Management. CIKM '24: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise, United States, 21st Oct. 2024, pp. 5230–5233. DOI: 10.1145/3627673.3679237. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04774438 (cit. on p. 15).

[40] J. Z. Liaw, A. Y. H. Chai, S. H. Lee, P. Bonnet and A. Joly. 'Can Language Improve Visual Features For Distinguishing Unseen Plant Diseases?' In: *Lecture Notes in Computer Science*. ICPR 2024 - 27th International Conference on Pattern Recognition. Vol. 15330. Lecture Notes in Computer Science. Kolkata, India: Springer Nature Switzerland, 4th Dec. 2024, pp. 296–311. DOI: 10.1007/978-3-031-78113-1_20. URL: https://inria.hal.science/hal-04851648 (cit. on pp. 16, 17).

[41] J. Lima, L. Tavares, E. Pacitti, J. E. Ferreira, I. Santos, I. Guimaraes Siqueira, D. Carvalho, F. Porto, R. Coutinho and E. Ogasawara. 'Online Event Detection in Streaming Time Series: Novel Metrics and Practical Insights'. In: IJCNN 2024 - International Joint Conference on Neural Networks. Yokoama, Japan, 30th June 2024, pp. 1–8. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04674128.

[42] J. Liu, T. Che, Y. Zhou, R. Jin, H. Dai and P. Valduriez. 'AEDFL: Efficient Asynchronous Decentralized Federated Learning with Heterogeneous Devices'. In: SDM 2024 - SIAM International Conference on Data Mining. Houston, TX, United States: Society for Industrial and Applied Mathematics, 11th Jan. 2024, pp. 833–841. DOI: 10.1137/1.9781611978032.95. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04597263 (cit. on p. 13).

[43] J. Liu, J. Ren, R. Jin, Z. Zhang, Y. Zhou, P. Valduriez and D. Dou. 'Fisher Information-based Efficient Curriculum Federated Learning with Large Language Models'. In: EMNLP 2024 - Conference on Empirical Methods in Natural Language Processing. Miami, Fl, United States, 1st Oct. 2024, pp. 1–27. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04734309 (cit. on p. 13).

[44] E. Paixão Silva, H. Balbi, E. Pacitti, F. Porto, J. A. F. dos Santos and E. S. Ogasawara. 'Cutoff Frequency Adjustment for FFT-Based Anomaly Detectors'. In: SBBD 2024 - Simpósio Brasileiro de Banco de Dados. Florianapolis, Brazil, 14th Oct. 2024, pp. 1–5. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04683135 (cit. on p. 14).

[45] L. Picek, C. Botella, M. Servajean, C. Leblanc, R. Palard, T. Larcher, B. Deneu, D. Marcos, J. Estopinan, P. Bonnet and A. Joly. 'Overview of GeoLifeCLEF 2024: Species Composition Prediction with High Spatial Resolution at Continental Scale using Remote Sensing'. In: CLEF 2024 Working Notes - 25th Conference and Labs of the Evaluation Forum. Vol. 3740. CEUR workshop proceedings 186. Grenoble, France: CEUR, 9th Sept. 2024, pp. 1966–1977. URL: https://hal.inrae.fr/hal-04720817 (cit. on p. 17).

[46] L. Picek, C. Botella, M. Servajean, C. Leblanc, R. Palard, T. Larcher, B. Deneu, D. Marcos, P. Bonnet and A. Joly. 'GeoPlant: Spatial Plant Species Prediction Dataset'. In: NeurIPS 2024 - 38th Conference on Neural Information Processing Systems. Vol. Track on Datasets and Benchmarks. Vancouver, Canada, 10th Dec. 2024. URL: https://inria.hal.science/hal-04852083 (cit. on p. 15).

[47] L. Picek, M. Šulc and J. Matas. 'Overview of FungiCLEF 2024: Revisiting Fungi Species Recognition Beyond 0-1 Cost'. In: CLEF 2024 Working Notes - 25th Conference and Labs of the Evaluation Forum. Vol. 3740. CEUR workshop proceedings 185. Grenoble, France, 9th Sept. 2024, pp. 1958–1965. URL: https://inria.hal.science/hal-04852127 (cit. on p. 17).

[48] M. Reis, R. Salles, G. Xexeo, R. Coutinho and E. Ogasawara. 'Matching Detections to Events in Time Series'. In: SBBD 2024 - Simpósio Brasileiro de Banco de Dados. Florianapolis, Brazil, 14th Oct. 2024, pp. 1–6. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04683212 (cit. on p. 14).

[49] R. A. P. Silva, W. Ferreira, E. Pacitti, Y. Frota and D. de Oliveira. 'A Heuristic for Executing Confidentiality-Constrained Workflows in Containerized Environments'. In: SBBD 2024 - Simpósio Brasileiro de Banco de Dados. Florianapolis, Brazil, 14th Oct. 2024, pp. 1–13. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04683224 (cit. on p. 13).

[50] R. A. P. Silva, Y. Y. Frota, D. de Oliveira and E. Pacitti. 'Assegurando a Confidencialidade de Dados de Workflows Executados em Nuvens de Computadores: Abordagens Heurísticas e Exatas'. In: SBPO 2024 - Simpósio Brasileiro de Pesquisa Operacional. Fortaleza, Brazil, 4th Nov. 2024, pp. 1–12. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04687976 (cit. on p. 13).

National peer-reviewed Conferences

[51] A. Dubar, T. Lefort and J. Salmon. 'peerannot: A framework for label aggregation in crowdsourced datasets'. In: JDS 2024 - 55es Journées de Statistique. Recueil de l'ensemble des résumés longs est téléchargeable. Bordeaux, France, 27th May 2024. URL: https://hal.science/hal-04562830.

[52] T. Lefort, A. Affouard, P. Bonnet, B. Charlier, A. Joly and J. Salmon. 'Cooperative learning of Pl@ntNet's Artificial Intelligence algorithm using label aggregation'. In: *Recueil des résumés longs : 55èmes Journées de Statistique de la SFdS*. JDS 2024 - 55es Journées de Statistique. Recueil des résumés longs : 55èmes Journées de Statistique de la SFdS. Bordeaux, France, 27th May 2024, pp. 35–44. URL: https://hal.science/hal-04574223.

Conferences without proceedings

[53] T. Lefort, B. Charlier, A. Joly and J. Salmon. 'Weighted majority vote using Shapley values in crowdsourcing'. In: CAp 2024 - Conférence sur l'Apprentissage Automatique. Lille, France, 13th May 2024. URL: https://hal.science/hal-04573727.

Scientific books

[54] E. S. Ogasawara, R. Salles, F. Porto and E. Pacitti. *Event Detection in Time Series*. Synthesis Lectures on Data Management (SLDM). Springer, Feb. 2025, pp. 1–178. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04776385 (cit. on p. 14).

Scientific book chapters

[55] R. Zorrilla, E. Ogasawara, P. Valduriez and F. Porto. 'A Data-Driven Model Selection Approach to Spatio-Temporal Prediction'. In: *Transactions on Large-Scale Data- and Knowledge-Centered Systems LVI : Special Issue on Data Management - Principles, Technologies, and Applications*. Vol. LNCS-14790. Lecture Notes in Computer Science. Transactions on Large-Scale Data- and Knowledge-Centered Systems. 21st July 2024, pp. 98–118. DOI: 10.1007/978-3-662-69603-3_4. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04672000 (cit. on p. 13).

Edition (books, proceedings, special issue of a journal)

[56] R. Akbarinia, T. Allard and A. Bonifati, eds. *Actes de la conférence BDA 2023, Montpellier*. BDA 2023 - 39ème Conférence sur la Gestion de Données Principes Technologies et Applications. 25th June 2024. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04627853.

[57] A. Hameurlain, A. M. Tjoa, R. Akbarinia and A. Bonifati, eds. *Transactions on Large-Scale Data- and Knowledge-Centered Systems; Vol. LVI, LNCS 14790. Special Issue on Data Management - Principles, Technologies, and Applications*. Lecture Notes in Computer Science 14790 (2024). DOI: 10.1007/978-3-662-69603-3. URL: https://hal.science/hal-04744662.

Reports & preprints

[58] T. Lefort, A. Affouard, B. Charlier, J.-C. Lombardo, M. Chouet, H. Goëau, J. Salmon, P. Bonnet and A. Joly. *Cooperative learning of Pl@ntNet's Artificial Intelligence algorithm: how does it work and how can we improve it?* 5th June 2024. DOI: 10.5281/zenodo.10782465. URL: https://hal.science/hal-04603038 (cit. on p. 11).

Other scientific publications

[59] E. Porcher, P. Bonnet, C. Damgaard, P. de Frenne, N. Deguines, B. Ehlers, J. Frei, M. García, C. Gros, U. Jandt, A. Joly, G. Martin, D. Michez, O. Pescott, T. Roth and D. Waller. 'Can we harmonize the monitoring of plants and pollinators?' In: *New Phytologist* 244.1 (13th Sept. 2024), pp. 39–42. DOI: 10.1111/nph.20038. URL: https://hal.science/hal-04680061.