V 3.0 – April 2025
Gaëlle Jaouen

EcoFoG policy for scientific data management and openness

1. Introduction

The scientific data produced by EcoFoG cover a wide range of thematic areas (biodiversity inventories, environmental descriptors, ecophysiological traits, materials science, chemistry, genomics, gas flows, etc.), all connected to ecological research. These data serve as fuel for the various one-off or long-term scientific projects of the joint research unit and its partners. To optimize this role, these data must be managed as rigorously as possible, so that they are preserved, available and unambiguously intelligible. These aspects are summarized in the acronym FAIR (Findable, Accessible, Interoperable and Reusable). The FAIRization of data is at the heart of this policy, and enables us to prepare and optimize, when appropriate, their publication. This policy applies to data heritage acquired in the past (before the implementation of a unit policy on scientific data), as well as to the production of new data, anticipating all aspects and needs (practical, IT, infrastructural, legal...).

Whatever their origin and nature, data must be centralized, controlled, standardized, documented, perpetuated and opened up as much as possible, in accordance with the application of the appropriate legal framework (French, European, other) and the institutional policies of the unit's various member institutes, or their recommendations (AgroParisTech, Cirad, CNRS, INRAe).

The gradual and structured aggregation of all scientific data produced by the unit is in line with the overall philosophy of this policy, which is to interconnect as many data as possible in a cross-thematic information system, the main inter-thematic link being georeferencing.

This policy has been drawn up by the unit's management and the person in charge of the unit's scientific data management (hereinafter referred to as the “SDM” for Scientific Data Manager), and must be followed and applied by all unit staff. They are accompanied and trained by the SDM.

2. Management of the data heritage

2.1 Centralization of old and newly produced data

Theme by theme, the data produced over many years, as well as that currently being produced or in the course of future projects, must be aggregated and banked in databases in order to secure them. The aim is also to avoid any loss or uncontrolled duplication, leading to a perilous multiplication of versions. For some scientific themes, this work was initiated as soon as the first data was created; for others, the data is scattered over several media, some permanent, some not. This means:

- Continuously evolve existing systems towards virtuous practices in terms of security, reliability and durability
- Structuring themes for which this is not yet the case, in order to bring together all data sources, using or developing suitable procedures and the most appropriate receptacles.

In practice: The unit's data producers must therefore make every effort to report and transmit any old, new or future data sources to the SDM, with a view to integrating them into the cross-functional information system and ensuring their compatibility with existing tools, or developing new tools for processing, storing and opening them.

2.2 Data curation and enrichment

As data is aggregated, it is explored, checked and, if necessary, corrected to ensure:

- Homogeneity: to ensure compliance with common standards (units, language, etc.) and reference systems (taxonomic, geographical, ontological), which are regularly updated.
- Completeness: is everything filled in? Are gaps justified? Are null values authorized?
- Quality: precision, accuracy, consistency (simple statistical and graphical checks, compliance with limits/ranges).

As far as possible, all data is enriched with as much relevant information as possible (georeferencing, newly acquired and potentially retroactive information, environmental descriptions).

In practice: These control and enrichment procedures are carried out to a greater or lesser extent, or even entirely, by the SDM, in conjunction with the scientific managers for each theme and the data producers (who are often the same person), who are asked to pre-format their data according to models established in advance with the SDM. However, the data must be inspected and checked as thoroughly as possible before being transmitted to the SDM.

2.3 Data documentation

For each dataset, the richest possible metadata must be filled in (origin, genesis, protocol...), data dictionaries describing each variable (meaning, type, unit, origin...) created and Data Management Plans (DMPs) drawn up. The DMPs define the content of the data sets, their origin, their genesis (author, context, funding), the precautions to be taken (legal, ethical, technical, etc.) and the management strategy applied to them (file organization, file and data naming conventions, storage, backup, sharing, publication and the person(s) responsible for each part of the strategy). DMPs must also be produced for as many scientific projects as possible, including PhDs, and may be based on DMPs for cross-disciplinary datasets.

A global DMP for the unit is drafted by the SDM and regularly updated to accompany the present policy. It can serve as a resource for the drafting of various project or thematic DMPs.

In practice: This documentation is supervised by the SDM, and is drafted in collaboration with the scientific managers of the various themes or projects. It is very important that the organization of the data produced, the naming conventions and the methods of data sharing during the course of the project are defined in advance of each project, and that these rules are applied by all those involved. The SDM is there to advise and support project managers in these discussions and the definition of these rules.

2.4 Application of the legal framework

Every data producer has a duty to ensure compliance with the legal framework when handling or producing data, particularly when collecting biological material (ABS), handling personal data (RGPD) or sensitive data.

In practice: Agents confronted with this type of data should contact the SDM for support and advice on the necessary steps to take.

3. Technical operation, data infrastructures

3.1 Creation and management of databases

Once the data has been curated by the SDM, and validated by the scientific managers of each theme and project, it is integrated into databases designed specifically for each field. The design of each database is done according to the data content and its purpose. The databases remain scalable and can be perfected at any time according to the progress of projects, changing needs and the arrival of new data to aggregate, as well as new variables.

These databases are created with Microsoft SQL Server, the software historically used in the unit, on the unit's dedicated server. A migration to PostGreSQL, a much more widespread, open, free and equally powerful software, integrating fully-fledged management of geographic data, is planned in the medium term, with full transparency for users.

In practice: The SDM manages the design, supply and evolution of the databases, keeping in mind the need to be able to interconnect them. The database server is maintained jointly by the unit's IT manager and the SDM.

3.2 Storage

As most ecological data is fairly light, our storage capacity for this type of data is not currently limited and does not require major investment. On the other hand, for heavy data, such as remote sensing, genomics and any project involving the manipulation or creation of large quantities of data (in excess of 50GB), dedicated servers exist within the unit and may be backed up from time to time by appropriate devices.

Any new project or any new acquisition of large amounts of data must be notified as far in advance as possible, so that an appropriate storage strategy can be planned, or even outsourced if necessary.

In addition, everyone must take care not to accumulate copies of data unnecessarily, in order to avoid their uncontrolled circulation on the one hand, and the abusive and ecologically damaging clogging up of our storage devices on the other.

In practice: Storage needs must be anticipated, and the SDM and the unit's IT manager must be informed and involved in the discussions as far in advance as possible of the arrival of the data.

3.3 Backups

3.3.1 Local backups

The virtual machine used as a database server is automatically backed up twice a week. These backups are managed by the unit's IT manager.

In addition, if major modifications are made to a database, a manual one-off backup is made before and after modifications by the SDM. The pre-modification backup is stored on another campus instance with an explicit date and name.

3.3.2 External backups

Automatic database backups are made on a server at CIRAD's IT Department in Montpellier:

- Every week for “active” databases (i.e. those in which data is regularly added or modified)

- Every 6 months, for “inactive” databases
- Every year, a copy of all the databases is made and stored on this server, year after year.

Back-up copies of the databases are also made regularly and stored securely in French Guiana, but outside the campus.

4. Data access for EcoFoG members

4.1 Roles and rights management

The only authentic versions of data are those contained in the databases. To ensure that this data is never compromised, it is essential to manage access and action rights on these databases. User groups are created by theme, and these groups are assigned more or less restricted rights to the data, according to their roles. With a few exceptions, no one other than the SDM can modify the data.

In practice: The SDM manages the constitution of groups and the allocation of rights in agreement with the scientific managers of each dataset.

4.2 Visualization interfaces and simplified data access

For several of the databases, Access interfaces have been developed to meet the expectations of those responsible for each theme and the needs of users. These interfaces, which are scalable according to demand, essentially enable data to be viewed, combined by query and extracted quickly and easily. Several distinct interfaces can be created for the same data source, depending on requirements.

In practice: The SDM develops and upgrades these interfaces as required. Their operation is based on the rights granted to the user (4.1).

4.3 Data access scripts for advanced analysis

In order to be able to manipulate and analyse data directly with the software most commonly used in the unit, mainly R, extraction routines are developed. These can be scaled so that the output corresponds to targeted analysis packages. These developments concern each theme, but can also combine several cross-disciplinary scientific themes and include GIS procedures.

In practice: The SDM develops these scripts and the necessary connection bridges between the analysis software and the database management system. All scripts developed are centralized and documented, and many are available on the EcoFoG's GitHub. All perennial scripts *developed by the unit's agents must be deposited in software forges and documented. Institutional forges are currently being developed (INRAe, CIRAD). The SDM is responsible for communicating progress on these points to unit staff.*

4.4 Individual requests

For anyone who is not fluent in R, unfamiliar with the use of Access, is not on the Kourou campus and therefore does not have authorized access to the database server, or for all data queries more complex and/or cross-disciplinary than the various access modes listed above allow, requests should be addressed to the SDM. Authorization to access the data is requested from the project's scientific manager if the requester is not already part of an authorized group, then a customized extraction is made and sent to the requester with all the information needed to understand the data (2.3) as well as the terms of use defined by the scientific manager.

5. Data Openness

5.1 Paracou data platform

For forest census data from the Paracou site (currently the most frequently requested data), an online platform has been integrated into the Paracou website, enabling them to be made available on scientifically motivated request. This site also includes a module for disseminating Paracou GIS data.

Since 2023, use of the data platform has been reduced thanks to the development of the EcoFoG data repository (5.3), but it remains online as it enables finer, more targeted queries.

From 2024-25, use of the GIS platform should be abandoned in favour of an alternative solution that is much more robust and durable: CIRAD's IDG (geographic data infrastructure), on which our GIS data will gradually be published (with metadata, durable identifiers and distribution licenses). This IDG should be fully operational by 2025.

5.2 Data publication in thematic repositories

In line with the policies of the unit's member institutes, some of the unit's data is published in thematic data repositories, to give it a targeted display. This type of publication should be encouraged. Such existing publications are regularly updated and enriched.

In practice: When data is ready to be published, those responsible for the dataset should contact the SDM to study the various publication options and receive support in the process.

5.3 EcoFoG data repository

An EcoFoG data repository, managed by the SDM, has been created to publish data for which there is no suitable thematic repository. CIRAD's institutional data repository was chosen to host this space because, on the one hand, functionalities and security are more or less the same from one institutional space to another and, on the other hand, most of our data, in terms of volume, consists of data from the Paracou scientific station, which is a CIRAD device. In addition, since July 2022, all French institutional repositories have been brought together in Recherche Data Gouv, where all published data will gradually appear as harvesting progresses. The EcoFoG data repository has been designed as a tree structure, with thematic branches (forest census, arthropods, wood science, ecophysiological traits, etc.) each subdivided according to its particular features. Spaces dedicated to research projects (for example) can be created on request, enabling data to be temporarily shared between partners, accessible for scientific articles reviewers and then published.

Each branch contains the data, metadata, data dictionaries, DMPs and all the documents and references required to make the data intelligible and reusable (analysis scripts, protocols, projects, articles, etc.).

Data access conditions are set on a file-by-file basis, so data can be published with varying degrees of visibility. They may be accessible only on scientifically motivated request, remain under embargo or be freely open, but whatever the choice, they are always accompanied by a distribution license defining their conditions of reuse (traceability, citation of sources, authorized or unauthorized commercial use).

In practice: The SDM manages the unit data repository and assigns administration rights, branch by branch, to the appropriate scientific managers. Any development of a new thematic area can be carried out by contacting the SDM, who will design the organization of this area, its feeding and the management of access rights with the thematic manager.

5.4 DataPapers publication

Whenever possible, published data should be accompanied by DataPapers. Several have already been published by the unit.

In practice: Dataset managers can ask the SDM to help them choose the Data Journal and write the DataPaper. The SDM must, at the very least, proofread the article to ensure that it is in line with the unit's data management procedures.

6. Agent training and dissemination of information and resources

To ensure that this unit policy is applied by everyone, as easily as possible throughout the data life cycle, it is necessary that i) everyone is trained to the extent of their needs, ii) they have access to the information resources required for the day-to-day management of their data, iii) they have access to the unit's own management documents.

Throughout the year, the SDM distributes information on data management and openness, whatever its origin, as well as training offers from various organizations, whether or not they are part of the unit's supervisory body.

Twice a year, the SDM offers a day of webinar training (one session in French in autumn, one in English in June) for the unit's staff and doctoral students. The dates of these training sessions are circulated as soon as they are known, so that everyone can make arrangements to attend. Training courses can also be organized on request.

Finally, a page on the unit's website is dedicated to the theme of scientific data. A summary of the unit's internally available and published data is presented, along with links to the unit's data management documents (DMP, unit policy) and training resources.

In addition, the SDM is available at any time for personalized, individual or group support (when setting up or launching a project, for example, or when starting an internship or PhD).