Selecting a trustworthy subject-specific repository for self-depositing data: methodology and analysis of existing services

Frédéric de Lamotte, Véronique Stoll, Cécile Arènes, Marie-Emilia Herbet, Stéphane Debard, Françoise Genova, Christine Hadrossek, Emilie Lerigoleur, Gaëlle Leroux, Gilles Ohanessian, et al.

Ouvrir la science !

Selecting a trustworthy subject-specific repository for self-depositing data: methodology and analysis of existing services

Research Data College Working Group

July 2024

Selecting a trustworthy subject-specific repository for self-depositing data: methodology and analysis of existing services

_____

Research Data College Working Group

_____

Frédéric de LAMOTTE – Co-chair of the Research Data College
INRAE

Véronique STOLL – Co-chair of the Research Data College
Observatoire de Paris - PSL

Cécile ARENES – Management of the working group
Sorbonne University

Marie-Emilia HERBET – Management of the working group
Université Jean Moulin Lyon 3

Stéphane DEBARD
IRD - UMR Espace-Dev

Françoise GENOVA
CNRS, Observatoire astronomique de Strasbourg

Christine HADROSSEK
CNRS, DDOR

Émilie LERIGOLEUR
CNRS, UMR Géode Toulouse

Gaëlle LEROUX
CNRS, Centre de recherche en Neurosciences de Lyon

Gilles OHANESSIAN
CNRS

Christelle PIERKOT
Data Terra

Marie STAHL
French School of Athens

The Research Data College of the French Committee for Open Science would like to thank the following for their help and expertise:

Thomas JOUNEAU, Université de Lorraine
Maria-Grazia SANTANGELO, Université Grenoble Alpes

———

July 2024 (translated in November 2024)

———

Graphic design: opixido

Contents

CONTEXT **5**

METHODOLOGY **6**

DEFINITION 6
ID SOURCE FOR THE REPOSITORIES 7
COORDINATION WITH OTHER WORK CARRIED OUT ON DATA REPOSITORIES. 7
Research Data Alliance *7*
Cat OPIDoR *7*
DoRANum *7*

AREAS OF WORK AND DELIVERABLES **8**

LIST OF TRUSTWORTHY CRITERIA FOR A SUBJECT-SPECIFIC REPOSITORY 8
Goals *8*
Key players *8*
Deliverables *8*
Timetable *8*
TO SUGGEST A METHODOLOGY AND A LIST OF TRUSTWORTHY SUBJECT-SPECIFIC REPOSITORIES. 8
Goals *8*
Key players *8*
Deliverables *8*
Timetable *8*
STRATEGY FOR ENSURING THE SUSTAINABILITY OF THE LIST OF REPOSITORIES 9
Goals *9*
Key players *9*
Deliverables *9*
Timetable *9*

CRITERIA FOR EXCLUDING AND DESCRIBING REPOSITORIES **9**

CRITERIA FOR EXCLUDING REPOSITORIES 9
No moderation of data deposits *10*
No guarantee about the sustainability of the infrastructure *10*
No long-term identifier allocated *10*
Repositories that assign rights *10*
Excessive pricing policy *11*
Location of physical data storage outside the European Union for specific types of data. *11*
Data deposits limited by institutional affiliation *11*
CRITERIA FOR DESCRIBING REPOSITORIES 11
Disciplinary field *12*
Data accepted *12*
Long-term identifier provided *12*
Data sustainability *12*
Type of moderation *12*
Possibility of embargo *12*
Volume limit *13*
Observations *13*

RESULTS AND ANALYSIS BY DISCIPLINE **13**

APPLICATION OF EXCLUSION AND DESCRIPTION CRITERIA 14
ANALYSIS BY DISCIPLINE 15

Astronomy *15*
Biology *16*
Chemistry *16*
Physics *17*
Environmental Sciences *17*
Humanities and Social Sciences *17*

INITIAL INFORMATION FOR UPDATING THE LIST OF REPOSITORIES 18

CONCLUSION 20

ANNEX: ENGAGEMENT LETTER 21

Table of figures

List of exclusion criteria ..... **Erreur ! Signet non défini.**
List of description criteria ..... **Erreur ! Signet non défini.**
Subject-specific distribution of selected repositories ..... **Erreur ! Signet non défini.**
Ranking of criteria based on difficulty obtaining information **Erreur ! Signet non défini.**

Suggested workflow for updating the list of repositories

Context

The research data policy of the French Ministry of Higher Education and Research has a specific purpose: to ensure that data is slowly but surely structured in accordance with the FAIR principles (Findability, Accessibility, Interoperability and Reuse of digital assets), and that the data is preserved and shared or opened *via* trustworthy data repositories. The issue of open research data is now official policy in France and at European level. Researchers are encouraged to make available to the community any research material that could contribute to the understanding of a scientific result. This data provision is contingent on complying with certain principles intended to ensure that the relevant data is intelligible and reusable: i.e., it is backed up by documentation and metadata that reflect the specific characteristics of the disciplines. It also assumes that access to this data can be guaranteed over time. Achieving these goals depends in large part on the choice of the repository. The level of requirement demanded when depositing data (the moderation policy or the nature of the metadata), together with the sustainability of the infrastructure, in large part shape the "FAIRisation" level of the data.

Including these institutional requirements in the landscape of scientific communication on a longstanding basis means that researchers can have suitable infrastructures for depositing and presenting their data. Some communities engaged with the issue of data sharing at an early stage (crystallography, astrophysics, genomics, etc.), equipping themselves with dedicated repositories that are today recognised far and wide. Others, however, are still struggling to identify the subject-specific repositories that will host their data. With no specific recommendations from research funders, learned societies or communities, "the selection of a suitable repository is delegated to the researcher". [1]

This lack of guidance can bring about two risks: on the one hand, a proliferation of erratic data deposits in general repositories that do not have a strict data description policy. On the other hand, an upswing in repositories supported by commercial publishers, which researchers might be directed to since there is no alternative service they know about or that has been recommended.

This observation is made in a paradoxical context, where some tools, such as catalogues or data repository directories, do exist. Using these tools to deposit data comes up against numerous hurdles:

- The array of repositories that require a specific institutional affiliation for depositing data.
- Unsuitable allocation by discipline.
- Reporting on unmaintained repositories.
- Complicated navigation paths for users.

Each of these factors wastes the researcher's time.

On March 3, 2023, the French Ministry of Higher Education and Research gave the Research Data College of the French Committee for Open Science the two-fold mission of:

- Proposing a definition of the trustworthy criteria for assessing the quality of a repository against the FAIRisation data objectives.
- Working to identify subject-specific trustworthy repositories suitable for *harvesting by Recherche Data Gouv*.

The Recherche Data Gouv ecosystem provides research teams with a multidisciplinary service for depositing, publishing and reporting their data in synergy with subject-specific and institutional repositories. Accordingly, it is essential to identify trustworthy subject-based French and international

[1] Source: https://www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf

repositories, whether they are CoreTrustSealcertified or not, so that research teams can be guided towards the most suitable repository for sharing and opening data in their discipline.

As a result, this briefing note suggests a method for identifying recommended subject-specific repositories for communities so they can self-deposit data, together with an initial list derived from the selected analysis criteria framework. It is based on the work undertaken since 2022 by the Research Data College of the Committee for Open Science[2]. The French Ministry of Higher Education and Research specifically commissioned the Research Data College to draw up a list of criteria to guide the selection of trustworthy subject-specific repositories for depositing and publishing datasets, prioritising the most active/organised disciplines in terms of data management[3].

In accordance with the Charter of the Colleges and expert groups of the Committee for Open Science (May 2023), the list of trustworthy repositories drawn up by the Research Data College is the subject of regular exchanges with the Ministry of Higher Education and Research. The latter carries out the final validation before sending the list to the appropriate organisations for dissemination.

Methodology

Definition

The Research Data College has defined a subject-specific data repository as a service and storage infrastructure that facilitates the depositing, description, open-access sharing, discovery and reuse, by humans or machines, of datasets specific to a scientific community. These datasets, which are associated with metadata, are retained for the medium or long term.

The services provided can be used to organise data logically and coherently, by (for example) adopting unique identifiers, implementing a moderation policy and guaranteeing a minimum length of time for preserving data.

Data repositories may have specific requirements and/or statutory restrictions regarding:

- The subject matter or area of research.
- Data quality.
- The origin of the data.
- The reuse of and access to data.
- File formats and data structure.
- The types of metadata.

A repository differs from a catalogue in that it has the capacity to host, manage and curate data, and not just the information system (cataloguing and presenting metadata harvested from other bodies).

[2] https://www.ouvrirlascience.fr/research-data-college/

[3] Mission letter from the ministerial administrator of data, algorithms and source code, 3 March 2023; voir annexe.

ID source for the repositories

The work to identify and analyse repositories was based on five key sources of information:

- The dedicated working group of the Research Data College consisting of members of the committee and external experts.
- Scientific and grey literature (articles describing how repositories operate).
- Repository directories (CatOpidor, Re3data, Fairsharing, Opendoar).
- Discipline-specific platforms dedicated to managing research data (such as the German NFDI consortium or Datacc.org) providing an initial inventory of subject-specific repositories.
- Feedback from the scientific community.

The information featured on each repository website was leveraged as a matter of routine. Where documentation was missing or incomplete, contact was made with repository managers to obtain more details (usually about the repository's moderation procedure or long-term sustainability).

Coordination with other work carried out on data repositories.

Research Data Alliance

In 2022 and 2023, the Research Data Alliance's *Data Repository Attributes Working Group* (DRAWG) started drawing up a list of "high-level" criteria to characterise research data repositories. This work was submitted for comments at the beginning of 2024. Since it is not finished, it has a fortiori not been validated or published by the RDA. This involves delivering a list of key attributes without exclusion criteria or recommendation objectives. The list is designed for all research stakeholders, not just data-producing scientists. The approach is, therefore, somewhat different from that described here. However, all the criteria adopted here can be found in the DRAWG list in identical or similar form. The list of criteria used here is more limited to facilitate its use.

Cat OPIDoR

Cat OPIDoR is a service "wiki" dedicated to research data; it provides exhaustive lists of French data repositories. This inventory also includes CNRS research data, a catalogue that lists the repositories and services specialising in CNRS research data and the bodies that support them. The objectives of Cat OPIDoR and this study are not the same: the former is dedicated to the latest developments relating to French repositories, while the latter has defined inclusion and exclusion criteria, giving rise to a list of French and international repositories.

DoRANum

In December 2023, a new resource was published on DoRANum, a self-learning platform for research data that has joined the resource centres of the Recherche Data Gouv ecosystem. The examples of uploaded data repositories focus on the human and social sciences with around 10 listed repositories, which are either rigorously disciplinary in nature or more general. Around 20 criteria were adopted, represented in the form of pictograms. The type of data accepted, which is a criterion for describing repositories in the present study, was not specified in the files uploaded on to DoRANum.

Areas of work and deliverables

The working group defined three main areas of work in response to the assignment it was given:

- To draw up a list of trustworthy criteria for assessing the quality of a subject-specific repository.
- To suggest a methodology note and an initial list of subject-specific trustworthy repositories.
- To put forward a strategy for ensuring the sustainability of the list of trustworthy subject-specific repositories.

List of trustworthy criteria for a subject-specific repository

Goals

To draw up a list of suitable criteria for selecting trustworthy subject-specific repositories that enable research communities to self-deposit data and publish datasets.
These criteria will form a standard analysis framework. They will provide a well-argued ID profile for each repository studied.

Key players

Working Group (WG) members

Deliverables

List of description and exclusion criteria for a selection of subject-specific repositories

Timetable

Second quarter of 2023

Methodology note and a list of subject-specific trustworthy repositories

Goals

- To draw up a methodology note explaining the approach used to compile the list of subject-specific repositories (sources of repository reporting, experts consulted, contact made with repository administrators, etc.).
- To suggest a list of trustworthy repositories, to be fleshed out in stages, based on the defined trustworthy criteria, and to offer guidance to French research communities when publishing their datasets. The working group draws on external experts from different scientific communities to ensure that this selection is pertinent. The list, drawn up as part of the Research Data College's roadmap, will feed into the work of the larger WG of the Recherche Data Gouv's catalogue module.

Key players

WG members and outside experts

Deliverables

- Methodology note.
- List of recommended subject-specific repositories.

Timetable

- Methodology note: third quarter of 2023.

- First version of the list of repositories: Third quarter of 2023.
- Second version of the list of repositories: first quarter of 2024.

Strategy for ensuring the sustainability of the list of repositories

Goals

- To suggest scenarios for ensuring the sustainability of the list of trustworthy repositories so that it will be updated or expanded.
- To take part in discussions about updating the list of trustworthy repositories harvested by Recherche Data Gouv.

The deliverable provided by the Research Data College of the French Committee for Open Science may be updated during the current term of office by the members of the working group responsible for monitoring the repositories. Furthermore, it is vitally important to suggest scenarios for regularly revising not only the list of repositories but also the validity of the trustworthy criteria, whose required level is likely to grow over time.
The list of subject-specific repositories will be revised on an annual basis and/or following changes to the requirements of the research funding agencies.

Key players

WG members

Deliverables

Strategy note and scenarios for updating the list of trustworthy subject-specific repositories and the list of criteria.

Timetable

First quarter of 2024

Criteria for excluding and describing repositories

Criteria for excluding repositories

The Research Data College has adopted a series of seven exclusion criteria for selecting trustworthy subject-specific repositories. These repositories will be able not only to accept the depositing and publication of datasets, but also contribute to their dissemination and subsequent reuse by scientific communities. There are three that relate to criteria about the quality of the service provided, and four to organisational criteria.

This basic set of criteria was drawn up to establish an initial list of repositories that research teams can easily identify and use. At the same time, it aims to be general enough so as not to restrict the available services to certified repositories.

List of exclusion criteria

No moderation of data deposits
No long-term identifier
No guarantee about the sustainability of the infrastructure
Repositories that assign rights
Excessive pricing policy
Data located outside the European Union for certain data types
Data deposits limited by institutional affiliation

No moderation of data deposits

The following should be excluded: repositories that do not practise data moderation (human or automated) designed to ensure a minimum quality of the metadata, which makes it possible to avoid transferring data that is incomplete or poorly described. In this example[4], the dataset has a meaningless title "Supplemental table S1", with no keywords, context or associated documentation. We feel that the moderation stage helps encourage the depositor to clarify and better document their dataset.

No guarantee about the sustainability of the infrastructure

We recommend prioritising maintained repositories with a data-preservation period of at least five years, following the practices put in place by Recherche Data Gouv. As a minimum requirement, if a repository can demonstrate previous durability, and it is still in operation, that can be sufficient evidence of credibility.

No long-term identifier allocated

In accordance with the FAIR principles, the use of a long-term identifier provided by the repository (PID, DOI for example) makes it easier to find and cite datasets (e.g. in a publication).

Repositories that practice the transfer of rights

The intellectual property practices of some publishers do not guarantee unrestricted access and unrestricted reuse of data deposited in the repositories that they develop and recommend. This is the case, for example, of ACS in chemistry, which accepts the depositing of nuclear magnetic resonance data in the form of FID files at the research data centre without detailing the licencing policy.

As a result, repositories that practice the transfer of rights are excluded. This position is in line with the guide "Partager les données liées aux publications scientifiques" (Sharing data relating to scientific publications) published by the Research Data College of the Committee for Open Science (2022). This advises against the practice of "trapping users in environments controlled by major commercial players in scientific publishing".

[4] https://doi.org/10.5281/zenodo.3725604

Excessive pricing policy

The purpose of this criterion is to exclude repositories where every low-volume data deposit automatically incurs a fee. This is the case, for example, with Dryad, which charges $150 or more for every data deposit depending on the volume[5] or the Digital Archaeological Record (tDAR), which charges a retention fee of $10 per 10 MB together with curation fees (basic metadata and file control) at a rate of $90 an hour[6].

On the other hand, repositories that may require a financial contribution in return for depositing significant volumes of data (more than 50 GB) have not been excluded.

Location of physical data storage outside the European Union for specific types of data.

Certain types of data (health data, survey results) that identify people even if pseudonymization and anonymization techniques are used. In this case, open-access communication is excluded and is rigorously regulated via the General Data Protection Regulation (GDPR). It was decided, therefore, to exclude data repositories located outside the European Union for depositing personal data that cannot be anonymised, with the exception of Switzerland, Great Britain, Japan and Argentina, which are deemed to be GDPR-compliant.[7] For other types of data, the reporting of repositories outside the European Union was included. This is all the more important since researchers lean towards repositories with an international dimension (Prost and Schöpfel, 2015)[8].

Data deposits limited by institutional affiliation

Ruled out are subject-specific repositories that limit data deposits to certain scientific communities where only scientists affiliated to the institution hosting the repository are authorised to deposit. The selection proposed at the end of the work is designed, therefore, to report infrastructures that are widely open and accessible to as many researchers as possible, irrespective of their affiliation.

Criteria for describing repositories

A brief ID profile is drawn up for each repository that includes the information research teams need for depositing datasets. In addition to the general descriptive information (name, URL, host institution), we chose to focus on seven areas:

[5] https://datadryad.org/stash/faq

[6] https://core.tdar.org/cart/add

[7] https://www.cnil.fr/fr/la-protection-des-donnees-dans-le-monde

[8] https://hal.univ-lille.fr/hal-01198379v1/document

List of description criteria

Disciplinary field
Data accepted
Long-term identifier provided
Data sustainability
Type of moderation
Possibility of embargo
Volume limit

Disciplinary field

Where possible, the disciplinary field employs the nomenclature used by HAL, which proposes a corollary tailored to the human and social sciences, which is not always the case with other existing nomenclatures. Furthermore, it offers up to three different levels of granularity, which makes it possible to provide better descriptions of the selected repositories.

Data accepted

Here it is a question of describing the type of data accepted by the repository, ensuring that the terminology specific to each discipline is employed in order to make the depositor's choice easier (e.g. NMR spectra, 3D structures of biological molecules, TEI-encoded corpus, etc.).

Long-term identifier provided

This section includes any long-term identifier provided by the repository (DOI, ARK, Handle, etc.) that helps to locate datasets based on FAIR principles.

Data sustainability

This criterion addresses aspects relating to the sustainability of the infrastructure and/or the repository's commitment to preserving deposited data for a period of time that is expressly defined.

Type of moderation

This criterion stipulates the type of moderation: metadata check, scientific control of data, human or automated intervention, etc.

Possibility of embargo

Some research teams may be keen to delay open-access publication of their datasets, specifying a specific embargo period. This criterion specifies whether the repository offers the possibility of combining the deposit with an embargo.

Volume limit

This information can prove important for researchers from disciplines that generate substantial volumes of data. It also helps to anticipate the expected cost if the repository requires a financial contribution above a certain volume of data.

Observations

This section flags up any additional information useful for reporting and characterising the repository (details on deposit procedures, etc.).

Results and analysis by discipline

The work carried out by the Research Data College between October 2022 and December 2023 led to the creation of an initial list of 49 subject-specific repositories covering the hard sciences and the human and social sciences, distributed as follows:

Subject-specific distribution of selected repositories

Field	Number of repositories
Astronomy	2
Biology	13
Chemistry	9
Physics	2
Environmental Sciences	7
Humanities and Social Sciences	16 (including 6 in archaeology and 3 in linguistics)

The inventory of these repositories is largely derived from the information provided by the repository websites, as well as by the operators of the various infrastructures. It follows that the collected information is based on self-reporting, and the information could not always be verified, in particular regarding moderation practices involving manual intervention.

The criterion regarding the duration of data preservation is also one of the items of information to be considered with caution. Given the difficulties involved in identifying the length of a repository's commitment, the frame of reference has on occasion been relaxed, drawing on other evidence of credibility, such as the supervision of the infrastructure or the life span of the repository.

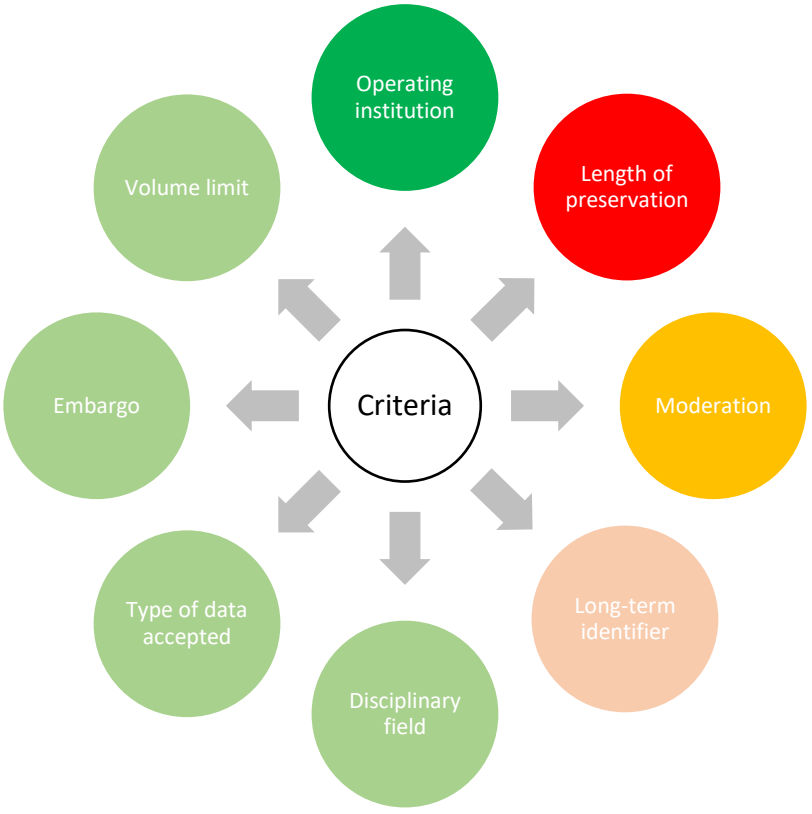
The aim of the work is to ensure that suitable repositories are visible to different communities so that it is easier for them to present their data. All of the suggested repositories comply with the general FAIR principles but do not detail the FAIRisation level of discipline metadata or how this metadata is readable by machines.

This inventory came up against several obstacles that reflect the differences in practices between disciplines, as well as the inherent vulnerability of the economic model backed by data repositories. The work called for direct contact with those responsible for managing 31 of the repositories. The research also led to approximately 40 repositories being ruled out.

Application of exclusion and description criteria

As stated above, the inventory was compiled according to a series of exclusion criteria that were used to identify qualitative and organisational criteria. This was drawn up after analysing other studies carried out in France and at the international level.[9] The criteria framework selected by the Research Data College provides the minimum basic set of essential information required for informing research teams when deciding on their choice of repository. This list of criteria is deliberately concise so that it can meet the needs of users more effectively. It has generated a wide-scale information-collection campaign, with some information proving difficult or even impossible to obtain.

Ranking of criteria based on difficulty obtaining information



Level of difficulty depending on the type of information sought

[9] Among the listed description frameworks, see the repository selection-support tool from Datacc.org (Lyon 1/UGA), the lists of preferred criteria from the NIH and the US federal government open science sub-committee, the COAR list of desirable or essential criteria and the work carried out as part of RDA (RDA Fairsharing WG).

The criteria for the volume limit, embargos, the type of data accepted, and the detailed disciplinary fields were among the easiest to identify. They can be found regularly on repository documentation, even if it was necessary on occasion to contact the operators to obtain precise details about embargoes or volume limits, for instance.

Obtaining information about long-term identifiers proved to be more complex than expected, for at least three reasons:

- Allocation of internal identifiers whose long-term ID value was difficult to confirm.
- Presence of DOIs assigned to publications associated with the datasets and not to the latter.
- Allocation of long-term identifiers only on request.

Furthermore, the issue of moderation required numerous discussions with repository operators to clarify the policy in place (automatic and/or human moderation, systematic nature or not of this moderation, possible assessment of the quality of the data, etc.). Since the notion of moderation is polysemous, it gave rise to numerous discussions about which repositories to select, especially since a repository may also change its moderation policy during the selection phase.

The criterion about the duration of preservation proved impossible to obtain in many instances. Large numbers of repositories are backed by non-sustainable funding resulting from short-term projects. This compromises the ability of infrastructures to commit to preserving data for a longer period than the duration of the project itself. The working group even observed that one repository dedicated to the humanities and social sciences indicated that it no longer accepted new data due to a lack of new funding. Insofar as very few repositories are in a position to commit to sustainable data retention (which makes the ecosystem of subject-specific repositories very vulnerable), the working group had to amend this exclusion criterion to include the lifespan of a repository and its different sources of funding.
The difficulties the project team encountered are reflected in the studies carried out on the sustainability of data repositories. Recent analyses have shown that 6.2% of the repositories listed in re3data had ceased operating after a median period of around 12 years. Among them, we note an over-representation of subject-specific repositories (136 repositories out of 191), mainly in life sciences and natural sciences.[10]

Analysis by discipline

It may appear that the list based on the criteria explained is incomplete or partial. Nevertheless, it reflects the shifting landscape of subject-specific repositories, where some disciplines are very well organised with several recognised repositories available, while others are less well equipped.

Identification also depends on the disciplines represented within the working group, whose members do not currently cover them all exhaustively. This first list will be gradually added to and updated based on the methodology described above and according to a sustainability strategy that is still to be approved.

Astronomy

Astronomy repositories have often been up and running for a long time, due to the fact that the discipline is structured around the IVOA (*International Virtual Observatory Alliance*). They are generally supplied by documentalists who are responsible for depositing and curating data for research teams. This may be space-mission data, whose management is designed as an integral part of the mission and whose data depositing is

[10] This proportion may also be explained by the type of the sample of repositories taken from re3data, which includes an over-representation of subject-specific repositories in these disciplines. “Compared to all repositories indexed in re3data at the time of data collection, repositories with these characteristics are also overrepresented in the sample.” p.10 https://arxiv.org/pdf/2310.06712.pdf

organised collectively. There is also data obtained from a ground-based observatory, which can be collected from the observatory after an embargo of one year as a rule. These astronomy-specific characteristics explain why few repositories in the discipline accept self-depositing data, which is one of the criteria of this study. This explains the relatively low number of repositories reported in this discipline.

While the ecosystem of repositories in astronomy is very international, the two repositories selected because self-depositing is possible are French: the Strasbourg astronomical Data Center and the Paris Astronomical Data Center. The former has been collecting astronomy data since 1972, while the latter has been in business for 20 years. In both cases, it is possible to deposit substantial volumes of data. At the same time, it is worth noting that some astrophysics sub-disciplines do not have dedicated repositories and may, therefore, be split among other repositories.

Biology

The range of biology services is wider than for the other disciplines studied. The 13 repositories reported in this document represent only a fraction of the infrastructures suitable for depositing and sharing data. Nevertheless, this abundance includes several biases. So, we note that repositories based in English-speaking countries are over-represented, and there is also substantial financial support from the NIH (National Institutes of Health), which is identified as the institution that backs three of the listed repositories. This reality is offset by the presence of the European organisation EMBL-EBI (EMBL's European Bioinformatics Institute), which plays a role in the ENA and PRIDE repositories. The DNA and RNA data repository is historically organised around three major international bases: GenBank in the United States, ENA in Europe and DDBJ in Japan. Insofar as these three bases are committed to a cooperative process to facilitate the sharing of genetic[11] sequences, we opted for reporting the European infrastructure, which seems to be the natural entry point for researchers working in France.

This discipline is also characterised by repositories with a long history (over 50 years for PDB and nearly 40 years for ENA). The nature of the data eligible for depositing in the listed repositories focuses mainly on neurobiology and genomics. In neurobiology, we have been careful only to recommend depositing non-human brain imaging data in repositories in the United States. Plant biology does not appear to be so well resourced, with only one repository reported in the list. The use of PIDs in repositories in this discipline is often based on internal identifiers.

Chemistry

Chemistry is a field where data structuring and dissemination is still not mature, except for crystallography, which boasts historical expertise in the area. This is reflected in the long history of the Cambridge Structural Database, whose genesis dates back to the 1970s. The second repository in this specialty, the Crystallography Open Database, has been in business for 20 years. It has been committed to the entirely independent reuse of data since its inception.

We have identified a total of nine suitable repositories, six of which involve German institutional porting or co-porting.

Most of the repositories we describe conform to the defined qualitative criteria, although the recent nature of some initiatives calls for heightened vigilance regarding the preservation policies in place. Nevertheless, their inclusion could be justified by the relevance of these repositories for the community. The Open Reaction Database, which has only been operational since 2021, has an ergonomic interface for depositing data and structured metadata, in a specialty where open access is still in its early days, and work on standardising experiment protocols is relatively new.

[11] See https://www.insdc.org/

In overall terms, chemistry repositories are relatively subject-specific based on a precise type of data (molecular modelling in theoretical chemistry, intermolecular interactions in supramolecular chemistry, chemical reactions in organic chemistry, etc.).

Physics

Physics is a discipline where data production can be on a massive scale. It has a very limited range of repositories for self-depositing data independently of the supporting institution. We identify two repositories, both European, that fulfil the defined qualitative criteria, one in particle physics, hosted by the University of Durham and the European Organization for Nuclear Research (CERN), and the other in materials sciences. The purpose of this last repository, NOMAD, is to house computational data in the material sciences in fields other than physics.

Environmental Sciences

With seven listed repositories, the number for environmental sciences is satisfactory. The list also reflects the presence of French institutional support in this area for over half of the listed repositories: SEANOE, SEXTANT, EasyData and Data Indores, with the last two launched recently (2023 and 2021 respectively).

Repositories hosting environmental sciences and ecology data are characterised by a relatively well-organised moderation policy, the possibility of depositing large volumes and the significant presence of certified repositories. The type of eligible data may vary greatly: while certain repositories, such as Data Indores, have a relatively general approach, other infrastructures - such as SEANOE or the World Data Center of Climate - put the focus on a specific type of data (geolocated marine data for the former, climate simulation for the latter).

Humanities and Social Sciences

With 16 listed repositories, the human and social sciences have a heterogeneous supply of centres: some cater to the requirements of specific communities (such as linguistics or archaeology), while others meet more generic needs, when the type of data likely to be submitted falls within the Humanities and Social Sciences (HSS) in the broad sense, without thematic targeting (Nakala).

The main feature of the HSS repository landscape is the high percentage of French infrastructures (10 repositories out of the 16 listed) in an almost exclusively European offering, except for the Qualitative Data Repository (QDR) for qualitative data and OpenContext in archaeology, both based in the United States.

Compliance with qualitative criteria is particularly high for repositories specialising in survey data (Socio-Political Data Centre - CDSP, Quetelet-Progedo-Diffusion and Qualitative Data Repository - QDR), where the deposit policy is supervised. This is due to the very nature of the survey data, which can be made available but that must guarantee anonymity in compliance with the GDPR. The contextualisation of the survey is also a significant task and specific to these repositories. The moderation policy is also well developed in archaeology and linguistics, where data formatting is widespread and verified by automatic and/or human moderation.

In other cases, although the moderation policy is only partial or non-existent, we nevertheless chose to report these repositories in order to fulfil the needs of the history community or to anticipate the introduction of a future moderation policy.

Unlike astronomy, for instance, HSS communities have tended to structure themselves around a model of independent self-depositing. Metadata moderation is gradually being introduced for some of the repositories studied but it does not form part of the standards of the entire human and social sciences community. However, the attribution of long-term identifiers is relatively widespread. This can be seen in the choice of reported repositories, except for Geovistory, which we nevertheless highlight due to its specific character in the HSS landscape (relational database between historical figures, places and organisations).

The granular quality of the data collected in HSS repositories is highly variable, with targeting especially pronounced in archaeology (modelling, 3D reconstruction of artefacts, bio-archaeological samples, etc.) and to a certain degree in linguistics (sound recordings, TEI or XML-encoded corpora, lexicons etc.). Depending on the type of data targeted, some disciplines - such as geography - must sign up to several repositories: Nakala for survey data, for example, Data Station Archaeology for GIS, Data Indores for observation data collected in the field or Pangaea or EasyData for geoscience data.

The HSS repositories studied are relatively recent, and it was very difficult - even impossible in some instances - to obtain information about the infrastructure's sustainability. For many repositories, the commitment to the length of data retention is weak due to funding for short-term projects or ignorance about the future directions of the supporting establishments.

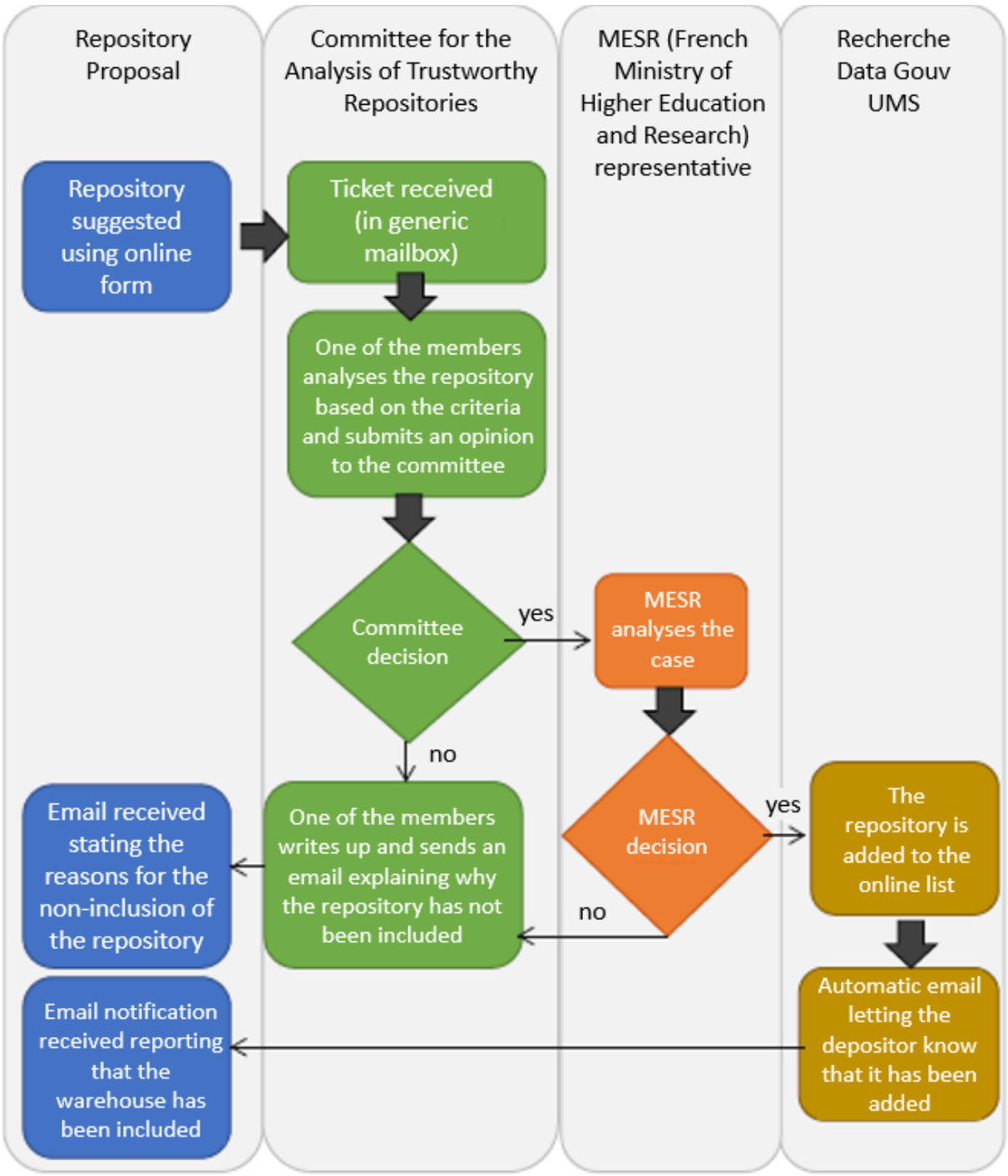
Initial information for updating the list of repositories

The current mandate of the Research Data College will finish at the end of the first half of 2025. Until then, the working group will carry on updating the list and looking at new proposals made by users or the various players assisting the research teams (data workshops, subject-specific resource centres, etc.) via a form, for example.

This time could equally be used to put a long-term analysis committee in place, which would take over from the Research Data College on a sustainable basis, tasked with helping to enhance the list of repositories according to the criteria drawn up in this briefing note. This committee will be given training in how to use the criteria by the working group of the Research Data College and will consist of members from:

- The repository-catalogue team from Recherche Data Gouv.
- Recherche Data Gouv data workshops.
- Recherche Data Gouv subject-specific reference centres.
- Former members of the Collège working group.
- Any disciplinary expert.

Suggested workflow for updating the list of repositories



Conclusion

At the end of this work to draw up the inventory of repositories, several lessons can be learned. Although the working group reached a consensus quickly about the series of qualitative criteria to be adopted, applying these criteria to the analysis of subject-specific repositories raised a number of questions and even difficulties.

In the first place, the piecemeal nature of publicly available information about the repositories (*via* online sites or the scientific literature) was overcome in part by making contact with repository operators. Nevertheless, this approach did not meet every expectation, including the particularly worrying issue about the sustainability of the infrastructures and their commitment to conserve the collected data on a medium or long-term basis. This information features only rarely on repository websites, meaning it is totally inaccessible to depositors, which can create problems regarding guarantees. In discussions with the repository operators, the subject of project-based financing often came up, i.e. limited in time, or the issue of the on-going search for funding. The uncertainty inherent in these responses was a cause of concern among group members regarding the robustness of the discipline-based repository ecosystem. Our observation - and we hope it is only temporary - is that the deposited datasets have a lifespan of between approximately three and six years based on the length of the financing granted to the repositories. This fragility is offset in part by the measures put in place to continue housing data when a repository is required to close. The working group had the opportunity to observe this for an HSS repository whose funding had come to an end. It was left with no choice other than to organise the migration of data to another repository, all the while encountering problems with some data formats. This work could be continued with a study on archiving research data, undertaken in conjunction with the archives departments of the different establishments. This would examine how to create a good synergy between sharing data in subject-specific repositories and its long-term retention, for example on the electronic archiving systems (EAS) that some establishments already have. The provisions of the Research Code on Scientific Integrity, reaffirmed by Decree No. 2021-1572 of December 3, 2021, on compliance with the requirements of scientific integrity, make establishments responsible for retaining the raw results of the research so they can be verified. Accordingly, the challenge is two-fold: to fall in line with the legal framework without delay, and it is essential to ensure the long-term preservation of these new heritage objects that are datasets. This applies equally to the creation of scientific heritage, which has become natively digital, for years to come.

Secondly, the exclusion criteria, which seemed reasonable and rational when the initial research into subject-specific repositories was carried out, has proved to be complex to apply for two of them. Moderation does not mean the same thing in every discipline: some are keen to ensure a very high degree of freedom for the depositor, while others control the dataset's metadata, the intrinsic metadata and the data itself. Against this background, it has not been easy to find the right balance, and some of the listed repositories do not entirely meet the moderation criterion. Nonetheless, they have been selected for their compliance with the other criteria and for their importance within their discipline.

This initial list should not be seen as an end, but as a starting point intended to provide a first level of referral for research teams involved in data sharing. It is designed, therefore, to be brought up to date based on a minimum of four key priorities: updating the information on the listed repositories, adding new repositories, removing repositories from the list, if needs be, and modifying the selection criteria. The landscape of subject-specific repositories is developing fast, and the list provided will quickly become obsolescent. It will only continue to be of interest if it is updated on a regular basis.

Annex: engagement letter

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Directorate-General of Research and Innovation
Directorate-General of Higher Education and Occupational Integration

Paris, March 3, 2023

Isabelle BLANC
National chief Data and Software officer
to
Véronique Stoll and Pierre-Yves Arnould
Joint Managers of the Research Data College
French Committee for Open Science

Re: Subject-specific research data repositories

Dear Sir or Madam,

The purpose of the research data policies introduced by the French Ministry of Higher Education and Research is to ensure that this data is gradually organised to comply with the FAIR principles (Easy to find, Accessible, Interoperable, Reusable), and that it is preserved and shared or opened up by trustworthy data repositories.

One of the key features of these policies is to develop services to work alongside research teams as they manage, share, open up and reuse research data. *Recherche Data Gouv* is an ecosystem that aims to share and open up research data. It has been designed to support research teams in their endeavours to structure data so that it complies with FAIR. Among other things, *Recherche Data Gouv* provides a multidisciplinary service for depositing, publishing and reporting research data. This complementary service of subject-specific repositories gives scientific disciplines that do not have a repository a sovereign solution for sharing and opening up their data. It also aims to report research data shared or opened up by third-party repositories.

It is vital to identify trustworthy French and international thematic repositories, whether they are Core Trust Seal-certified or not, for two reasons: to direct research teams towards the most relevant repository for sharing and opening up their subject-specific data; and to expand the *Recherche Data Gouv* data catalogue that will collate the data available in third-party repositories.

Against this background, we need to have a list of appropriate criteria for selecting trustworthy, subject-specific repositories that can be used not only to deposit and publish datasets but also to contribute to their dissemination and later reuse by the scientific communities.

In response to this challenge, I would like to assign the following missions to the Research Data College of the Committee for Open Science:

1/ In the context of the work that is part of its responsibilities:

- You will draw up the list of suitable criteria for selecting trustworthy, subject-specific repositories for depositing and publishing datasets based on the work that has already been undertaken. As a priority, you will factor in the most active/organised disciplines regarding data management.
- You will propose the methods for monitoring and developing these criteria.
- Drawing on a circle of experts that you will have identified in advance, you will contribute to identifying the functional requirements of a tool to help select subject-specific repositories for research teams.
- You will help to draft educational materials to provide support to research teams for using this tool.
- You will put forward scenarios for ensuring the sustainability of the project, incorporating future developments in the repository landscape.

2/ Working in close coordination with the team tasked with developing *Recherche Data Gouv*:

- You will suggest an initial list of trustworthy repositories that can be "harvested" by the *Recherche Data Gouv* catalogue module derived from the list drawn up as part of the work of the Collège des Données de la Recherche.
- Within the scope of subject-specific repositories, you will take part in discussions about updating the list of trustworthy repositories collated by *Recherche Data Gouv*.
- You will contribute to work meetings on the *Recherche Data Gouv* catalogue.

I would like to thank you for putting a working group in place and appointing managers to head up this assignment (see the composition of the working group in the Annexe). In addition, thank you in advance for clarifying, as promptly as possible, the assignments, deliverables, working methods and different deadlines by means of a brief framework note.

Thanks to the Research Data College for their involvement; I know I can rely on their 100% commitment to supporting the research community.

Isabelle Blanc