

FAIR x FAIR

Feasible, Affordable and Implementable
Requirements for a FAIR research data
repository

Consorti de Serveis Universitaris de Catalunya (CSUC)

MARCH 2019

Acknowledgements

This document has been prepared jointly with representatives of the University of Barcelona, the Universitat Autònoma de Barcelona, the Universitat Politècnica de Catalunya, the Pompeu Fabra University, the University of Girona, the University of Lleida, the Universitat Rovira i Virgili, the Open University of Catalonia, the University of Vic-Central University of Catalonia and the Ramon Llull University.

We would like to express our sincere appreciation to all experts cited in this document for the time they devoted to evaluating and preparing this report.

Written by Mireia Alcalá, CSUC's Information Resources Expert

Coordinated by Lluís Anglada, CSUC's Open Science Director



This document is licensed under the Creative Commons Attribution (<http://creativecommons.org/licenses/by/4.0/>).

Summary

| | |
|---|----|
| Executive summary..... | 4 |
| 1. Increasing importance of open research data..... | 7 |
| 2. The consortial actions and the methodology for writing this report..... | 9 |
| 3. The Research Data Management Support Service today..... | 10 |
| 4. Contextual determinants..... | 12 |
| 5. Minimum feasible functional requirements..... | 14 |
| 5.1 Persistent identifiers..... | 14 |
| 5.2 High storage capacity..... | 15 |
| 5.3 Medium- to long-term preservation..... | 15 |
| 5.4 Interoperability with other systems..... | 17 |
| 5.5 Management of special characteristics..... | 18 |
| 6. Best practices..... | 22 |
| 6.1 Data curation..... | 22 |
| 6.2 Dataset selection..... | 23 |
| 6.3 Encouraging the use of open formats..... | 23 |
| 6.4 Using widely-accepted controlled standards, protocols and vocabularies..... | 23 |
| 7. Final recommendations..... | 24 |
| References..... | 25 |
| Appendices..... | 28 |
| Appendix 1 – FAIR principles..... | 28 |
| Appendix 2 – The experts..... | 29 |
| Appendix 3 – Countries..... | 33 |
| Appendix 4 – Additional documentation..... | 37 |
| Appendix 5 – Assigning DOIs..... | 38 |
| Appendix 6 – Hardware and software for a data repository..... | 39 |
| Appendix 7 – Glossary..... | 40 |

Executive summary

In the last few years, the data collected, generated and used in research projects have received the attention of the scientific community and research management bodies worldwide. At a European level, the Horizon 2020 framework programme and its Open Research Data Pilot have promoted what has been called ‘research data management’. This is an umbrella term that encompasses activities related to the creation, organization, structuring, storage, preservation and sharing of data.

For projects funded by the European Commission (EC), a data management plan must be drawn up, and the data must be deposited in open access in accordance with the FAIR (Findable, Accessible, Interoperable and Reusable) principles in order to increase the efficiency and transparency of research through the rapid dissemination of results, and to facilitate reuse.

The universities of Catalonia initiated the Research Data Management Support Service in September 2016. This service is pivoted on three main areas:

- Facilitating the drafting of data management plans.
- Informing about repositories where the data can be made public, and in some cases modifying institutional repositories to support data.
- Preparing materials that allow institutions to establish their open access to data policy.

The service currently offered has two shortcomings:

- Little demand, probably for several reasons, but mainly the novelty of the subject.
- Lack of the infrastructure required by the European Commission to publish research data in FAIR form.

For this reason, the vice-rectors for research who form the Functional Commission of the Open Science Area (ACO) of the University Services Consortium of Catalonia (CSUC) decided to commission a report that would determine the feasible functional requirements that a data repository must have in order to comply with the FAIR requirements. The decision (CF ACO, 02.11.17) stated that ‘Work should start to prepare a proposal of functional requirements for the creation of consortial research data repositories’. These requirements had to follow the guidelines established in the European Open Science Cloud Declaration (EOSC, 2017).

To draw up this report, a specific working group was created, experts from inside and outside the CSUC were consulted, similar experiences in Flemish Belgium, Finland, the Netherlands, Portugal and Sweden were examined, and the main documents published recently on this subject were studied.

Experts were interviewed with the aim of drawing up a list of functional requirements. In the first meetings and interviews the experts stated that, in addition to the technical requirements, there were two fundamental conditions for setting up a research data repository: to establish

the contextual conditions of the repository and to develop good practices in establishing the criteria of FAIR datasets.

The experts repeatedly stated that the management of research data and its open publication has a long history but is still incipient. This means that any requirements established now will probably need to be extended or modified in the coming years. Furthermore, although we cannot establish stable requirements, **the experts recommended creating a research data infrastructure immediately in order to generate expertise and best practices for its management.**

Given that data repositories have not yet been consolidated or standardized, several options for decision making in different contexts can be observed in Europe. This report accepts that the recommended options for the universities of Catalonia in relation to the research data repository are the following:

- Do it now in order to have a place to publish data and gain experience with data management, even though it will need to be readapted.
- Do it in Catalonia, even though the data could be published in repositories already in operation in Europe, because data are now considered strategic, and a local repository will facilitate compliance with the legal measures.
- Use only permanent or final data, even though researchers also need to manage temporary data files.
- Use only data from disciplines that do not have consolidated data repositories, because for those who have one the best option is to publish there.

The experts also stressed that an infrastructure for publishing research data is insufficient in itself, and that research data management requires the development of a series of best practices, including the following:

- Curate the data, i.e. document them so that they are understandable to users other than those who generated them.
- Create protocols and criteria for selection and expurgation the datasets.
- Encourage the use of open or non-proprietary formats and, to the fullest extent possible, migrate proprietary formats to open formats.
- Use widely-accepted controlled standards, protocols and vocabularies.

The main part of this report focuses on describing the minimum functional requirements for ensuring that the infrastructure created meets the FAIR requirements. Based on prior work of the ACO commission, these requirements were related to the following groups:

- persistent identifiers
- high storage capacity
- medium- to long-term preservation
- interoperability with other systems, and
- management of special characteristics.

Given the above considerations, the document ends with the following final recommendations:

1. To create immediately in Catalonia a repository in which research data can be published in FAIR form in order to develop expertise and best practices in research data management. This repository must meet the following requirements:
 - a. Meet the current FAIR requirements and any that are established by the EC in the immediate future.
 - b. Offer added-value benefits in comparison with the current options, such as the assignation of digital object identifiers, interoperability with the Research Portal of Catalonia and enhanced preservation.
 - c. Use existing open-source software.
2. Promote and facilitate the publication of open research data through actions in which universities publicize the existing research data management service to allow data management plans to be drafted and to offer advice on data publication. Information on the service must be provided to the units conducting research, with the active participation of research offices and services.
1. Provide training on the concepts of open science in general and research data management in particular. In accordance with the guidelines of the EC's FAIR Data Expert Group, this training should include all members of the university community but, to be effective, it should distinguish between advanced users, young researchers and university support staff.

1. Increasing importance of open research data

Science has always been based on the use of data, but until recently data were difficult to collect, preserve, share and reuse. The ability of computers and communications networks to process, preserve and communicate data has changed the scientific process by making it more efficient, transparent and collaborative. As noted in the recent statement on open science of the Conference of Rectors of Spanish Universities (2019), current research activity generates, consumes and processes data intensively, and preservation and reuse are key in 21st-century research.

The bases of this movement were laid in Paris in 2004, when the ministries of science and technology of the OECD member countries, together with China, South Africa, Israel and Russia, approved the *Declaration on Access to Research Data from Public Funding* (OECD, 2004).

The European Commission (EC) gave a global dimension to these changes in 2013, when it launched a public consultation (European Commission, 2013) on the impact of this movement, which now goes under the name ‘Open Science’. In 2016, under the Dutch presidency of the European Union, the ‘Open Science - From Vision to Action’ conference was held in Amsterdam. The final document of the conference stated that research results must be public and reusable, and datasets of research projects must therefore have management plans and follow FAIR principles (see Appendix 1), which state that research data must be ‘Findable, Accessible, Interoperable and Reusable’ (Government of the Netherlands, 2016).

The EC, the major funding agency in Europe, has been preparing the ground for researchers to start managing their data. Under the Horizon 2020 framework programme (2014-2020), it launched the EC Open Research Data Pilot (ORD pilot), which aims to ‘improve and maximise access to and reuse of research data generated by Horizon 2020 projects and takes into account the need to balance openness and protection of scientific information, commercialisation and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions’ (European Commission, 2016). The ORD pilot only affected specific areas of Horizon 2020, for which it required submission of a data management plan (DMP) within the first six months of a project and open availability of the data (unless there were reasons to keep them closed).

In 2017, the pilot extended to all areas and adopted the principle ‘as open as possible, as closed as necessary’. In the next Horizon Europe framework programme (2021-2027), it will be mandatory to make data openly available and create a DMP. The data must be made public under the FAIR principles and must be open by default, although in certain cases they may be closed.

Several European bodies (the EC, research funders, universities and scientific associations) are giving great importance to these changes. This is reflected in the approval of national plans to facilitate the achievement of the open science goals, such as those of Finland (2014), Slovenia

(2015), Portugal (2016), the Netherlands (2017), France (2018) and Serbia (2018). The objective of all these plans is that research data should be openly available.

The objective of making research data openly available is also reflected in the statements and roadmaps for open science that have been made by entities such as the European University Association (2018), the League of European Research Universities (2018), Young European Research Universities (2018), LIBER - Association of European Research Libraries (2018) and the Conference of Rectors of Spanish Universities (2019).

The roadmaps include the Open Science Policy Platform Recommendations (2018), which state that one of the priorities for the European Open Science Agenda is to make research data openly available according to the FAIR principles. The recommendations also state that funders and research-performing organizations should give credit to FAIR data from research work similar to the credit given to publications. DMPs should be mandatory for all research projects and should be machine-readable. Finally, the data resulting from publicly funded research must be made FAIR and citable, and be ‘as open as possible, as closed as necessary’.

In conclusion, the open publication of data in FAIR form is a cornerstone of all these statements, roadmaps and national plans.

2. The consortial actions and the methodology for writing this report

When the Open Science Area (ACO) of the University Services Consortium of Catalonia (CSUC) was set up, a Work Plan for the period 2017-2019 was approved (Doc. CO17/01). This plan set out two objectives in relation to research data: to advise researchers on the preparation of DMPs and to guarantee the existence of repositories for research data.

In the meantime, the CSUC's Research Support Working Group (RSWG) was already starting to set up the Research Data Management Support Service (see below). The RSWG drew up working proposals on open research data (Doc. CO17/11) that were seen and approved by the ACO's Functional Commission. The document addressed the objective of the CSUC's General Action Plan: 'to analyse the possibilities for creating a consortial data repository and to draw up the proposal for it'.

The Commission considered that it was necessary to determine the functional requirements of a data repository that, to achieve economies of scale, should be cooperative but not necessarily centralized. The repository would improve the services rendered and meet needs that were not covered by the current infrastructure. The Commission made the following recommendations:

- Given that research data are strategic, they should be stored in a dedicated infrastructure.
- The repository must offer qualitatively better features than those offered by the current institutional repositories.
- Economies of scale must be considered with regard to costs and specialization of the data to be stored.
- The repository must offer interoperability between institutional repositories, CRIS systems and research portals.

A technical committee was set up to determine the feasible functional requirements for a research data repository to comply with the FAIR requirements. The members of this committee (see Appendix 1) come from the university services that are involved in the processing of research data: libraries, research offices and ICT services. The committee has met twice.

In parallel, the following actions were taken:

- Interviews were held with 32 experts on research data management (see Appendix 2).
- Information was collected on the solutions adopted in places with characteristics similar to those of Catalonia: Flemish Belgium, Finland, the Netherlands, Portugal and Sweden (see Appendix 3).
- Various technical reports related to the subject were compiled and studied (see Appendix 4).

3. The Research Data Management Support Service today

In mid-2015, the RSWG devoted itself almost exclusively to research data management in order to offer support to projects financed by the ORD pilot in three main areas: DMPs, data repositories and the open access to data policy.

As a result of the collaborative tasks, in September 2016 the Catalan universities launched the Research Data Management Support Service. Its services include the following:

- Assistance in the preparation of research data management plans.
- Recommendations on selecting a repository for research data.
- Extension of the services of the institutional repositories to depositing data.
- Preparation of requirements for a consortial data repository.
- Dissemination and training.

With regard to data management plans, the universities offer support through the Research Data Management Plan tool (<https://dmp.csuc.cat>), which allows DMPs to be created for projects funded by the Horizon 2020 programme and the EC's European Research Council. The tool can be used to create, share and export FAIR DMPs and is an adaptation of DMPRoadmap, an open source software that is distributed under the MIT licence and was developed jointly by the Digital Curation Center and the University of California Curation Center. The added value of the tool is information (CSUC, 2016) on what the plan must contain, with real descriptions and examples provided by the RSWG that are maintained jointly but can be customized by each institution.

Several solutions are offered for the data repository. First, researchers are advised on platforms for publishing their data. The recommendations (CSUC, 2017) show a range of available options of both thematic and multidisciplinary data repositories. Some universities have adapted their institutional repositories (CSUC, 2017) and offer them as a possibility for some cases. Though this option is a good start, it does not provide all the benefits of a FAIR repository. Currently, much of the work focuses on establishing minimum feasible functional requirements for a data repository to be FAIR. Finally, the results are monitored using a variety of indicators, the service is publicized through infographics or video capsules, for example, and training is given to university research support staff.

In relation to the open access to data policy, the vice-rectors for research that form the ACO's Functional Commission approved as recommendations a document that set out the subjects that a university should consider when preparing a data management policy. The recommendations made in this document, 'A data management policy for a university', (CSUC, 2018) cover the responsibility of data management, the place of deposit, the retention period, the preparation of a DMP, and the preservation and conservation of data.

To determine their research data management needs to be met by the service, two surveys of researchers were carried out in Catalan universities (CSUC, 2019). It was found that the service has two clear weaknesses. The first is the low level of use: there is a great distance between what is offered and the demand for it. The latest survey reveals that 80% of people are unaware of the Research Data Management Support Service at their university. The main

reasons for this are probably that the subject is new, the current research projects are not yet subject to compulsory data publication, and information on the service has not reached researchers who would potentially use it.

The second weakness is that the service does not have a repository of its own for making research data openly available. Research data on some scientific disciplines can be found in consolidated repositories such as the Protein Data Bank for 3D protein structures and the National Oceanographic Data Centre for oceanographic data, which are the most appropriate places to publish data of this type. In most cases, however, the options are to deposit the data in a general repository outside the researchers' institution, in that of an institution participating in the research project that has a data repository, or in the researchers' own institutional repository. Institutional repositories do not meet FAIR requirements with regard to identifiers, preservation and storage capacity, for example, and this fact inhibits the publication of data.

4. Contextual determinants

The experts consulted agreed that, despite the strategic importance given to open publication of research data (see Section 1), data repositories have not yet been consolidated or standardized. Indeed, today there are many documents and reports on the subject but few functioning repositories. The repositories are very diverse and often in their early stages, so it not possible to select and copy best practices.

From the meetings and interviews with experts it emerged that a series of decisions that would affect the future development of the repository should be taken prior to setting it up.

The options were the following:

To create the repository now or to wait

- The experts agreed that open publication of research data is an emerging subject and that there are still many questions to be determined. However, they also agreed that protocols and best practices must be developed for open publication of research data, and that this can only be done by having a functioning repository.
- It was therefore decided that the repository must be created now.

To consider the repository as under development or definitive

- As there are still questions to be determined, it must be assumed that the repository will have to be readapted to any requirements and practices that are consolidated internationally. Research data repositories are less advanced than other data repositories, but the necessary experience can only be obtained by having our own.
- It was decided that the repository must be considered to be under development.

To use only final data or to also include provisional data

- Research projects use different sets of data at different times. Many of the data are provisional or instrumental and have only a temporary value for the research. Researchers need to manage their research data in addition to publishing them, but the needs of data management differ from those of data publication.
- In order to reduce the requirements and the cost of the repository, it was decided that it should only publish final data for the moment.

To set up the repository in Catalonia or to use external ones

- To meet the needs created by the Horizon 2020 programme in this still incipient stage of open publication of research data, some repositories that have been created allow any researcher to publish data in them. The benefits of these repositories are good, and they can obviously be used to publish data, but it is widely thought that research

data are a strategic resource that should be within the research system itself. Additionally, due to issues related to the regulation of personal data protection, the publication of data is legally facilitated if it is done within the national territory.

- For strategic and legal reasons, despite the fact that data can be published in repositories already functioning in other European countries, it was decided that the repository should be located in Catalonia.

To focus on disciplines without consolidated data repositories or to accept all disciplines

- Some scientific disciplines (genomics is a clear example) have in recent years undergone great advances precisely thanks to data sharing. They have well-established repositories and have developed best practices for publishing their research data. In these cases, it is best to consolidate the disciplinary repositories. However, in many disciplines no repository is as yet available.
- It was therefore decided to limit the data repository to disciplines that do not have a well-established one, because the best option for those that do is to publish their data there.

In accordance with the above comments, and taking into account that these decisions affect the establishment of requirements, **this report recommends creating the repository now, although it must be considered as under development, creating it in Catalonia, using it for publishing final research data, and focusing on disciplines that do not have a well-established thematic data repository.**

5. Minimum feasible functional requirements

In order to define the minimum feasible functional requirements (FR) that a data repository must have, interviews were conducted with 32 experts in various fields. Articles dealing with this subject were also taken into account, and include the following:

- Amorim, R. C. (2017). A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, 16(4), 851-862. doi: 10.1007/s10209-016-0475-y
- Kim, S. (2018). Functional Requirements for Research Data Repositories. *International Journal of Knowledge Content Development & Technology*, 8(1), 2-36. doi:10.5864/IJKCT.2018.8.1.025.

The requirements gathered were classified according to the categories identified in the working proposals on open research data contained in the document ‘Consortial Research Data Repositories (RDRs)’ (CSUC, 2017):

- persistent identifiers,
- high storage capacity for different formats,
- medium- to long-term preservation,
- interoperability with other systems, and
- management of special characteristics.

5.1 Persistent identifiers

Persistent identifiers (PIDs) must be used in order to improve access to research data. PIDs are ‘constructed and implemented such that the identified resource will remain the same independently of the location of its representation and independent of the fact that several copies are available at various locations’ (IASA-TC04, 2009).

FR₁: To assign DOIs as the identifier

There are several types of PID, but digital object identifiers (DOIs) have recently gained accepted as an international standard for research data. A DOI is a unique alphanumeric string that identifies materials published online. It offers extra benefits, according to the ANDS (2018): it guarantees the quality and accuracy of data, increases the number of citations, guarantees long-term accessibility, provides an easy link, and gives datasets an equivalent status to other scholarly publications.

The non-profit organization DataCite is currently the leading body providing DOIs for research data. DOIs are not assigned directly by DataCite but by its members, which act as agents (see Appendix 5).

FR₂: To support ORCID

The universities of Catalonia agreed (CSUC, 2016) to use the ORCID unique identifying system for the authors of the scientific output of their institutions. It is therefore an essential requirement for the data repository to support ORCID as a PID and to associate researchers with the data sets they generate.

5.2 High storage capacity

FR₃: To support files of up to 10GB by default

The recommendations for selecting a repository for the research data repository (CSUC, 2017) compared the five most prominent multidisciplinary data repositories that can be used freely. Two of them allow depositing of files of 10GB or more, but the other three only accept files of up to 5 GB.

Furthermore, on two occasions¹ Catalan researchers have been asked about the size of the datasets they use and it has been found that most are less than 10 GB. Provision must also be made for storing larger files that must be made publicly available to meet the requirements of funding agencies.

FR₄: Elasticity for growth

Although the publication of research data is still incipient, it is considered that it will increase exponentially. The capacity for data storage must therefore be elastic so that it can adapt to demand.

5.3 Medium- to long-term preservation

All digital files that are being created today are liable to become obsolete, and research data are no exception (Johnston, 2017). Digital data are more fragile than those recorded on paper, because according to the type of medium in which they are stored (magnetic, optical, etc.), they may be exposed to damage or decomposition over time. Having a preservation strategy helps mitigate these risks to ensure long-term access to the data.

FR₅: To store files for at least 10 years

The Leaders Activating Research Networks (LEARN) project collected European research data management policies (LEARN, 2017) and found that 10 years was generally established

¹ The surveys of 2016 and 2018 both asked respondents about file size. They can be consulted at RECERCAT.

as the data preservation period. It was therefore decided that the repository must allow the files to be available for at least 10 years.

The ten-year period should count from the last consultation or download rather than the date of publication in order to preserve the most important and widely used data.

FR₆: To have at least two copies at different geographic locations

The National Digital Stewardship Alliance includes the number of copies as one of the elements to be considered within its levels of digital preservation (Phillips, 2013). It states that the first requirement to ensure access to files in the future is to have at least one second copy of the file. In addition, the copies must be geographically distributed in order to avoid local threats such as natural or human disasters.

The repository must have access to one of the copies in real time and have at least one other dark copy in a different geographic location.

FR₇: To check the data integrity periodically

In order to protect the data integrity, it will be mandatory to check periodically that the stored data have not been corrupted or modified accidentally or deliberately. The process consists of adding each of the basic components of a system (usually each byte) and storing the resulting value for later comparison. If the value is the same, it is considered that the file has not been altered.

The repository must perform these checks periodically. In addition, a record with information on the integrity of the files must be kept in order to detect those that are damaged.

FR₈: To follow the OAIS preservation model

The Open Archival Information System (OAIS) model is currently used for long-term preservation of scientific information in digital format (Hirtle, 2001). In addition to capturing the metadata available during the ingestion, data repositories often distribute this information to research search engines or repository indexers, thus improving the visibility of publications.

It was decided that the repository must follow this model, because it is the basis for obtaining the CoreTrustSeal certificate.²

² CoreTrustSeal offers certification to data repositories based on the catalogue and the procedures.

5.4 Interoperability with other systems

FR₉: To communicate with other repositories

Data generally need to be integrated with other data and to interoperate with applications or workflows for analysis, storage or processing (GO FAIR, 2018), and to avoid duplicating tasks.

For this reason, the repository should allow its data to be exchanged with disciplinary repositories to involve the rest of the community, and also with institutional repositories of universities, software repositories such as GitHub, and research management tools of universities such as CRIS.

FR₁₀: To communicate with storage tools in the cloud

During research projects researchers store, manage and share their data. Although these processes are usually carried out physical units (hard drives, CDs, etc.) of universities, more and more, storage tools are used in the cloud (such as Dropbox, Google Drive, Amazon and UNIDISC, among others).

In order to help researchers to upload datasets from cloud storage tools, the repository must communicate with them via APIs. In this communication with the cloud, mechanisms must be established to guarantee quality control of the data.

FR₁₁: To export the metadata to discovery tools

One of the objectives of publishing and sharing research data is that they should be reused. For this reason, exposure of the repository's content to other research platforms improves its visibility and increases its reach (Amorim, 2017). It is therefore essential for the repository to export its metadata to discovery tools.

Initially, the Research Portal of Catalonia must serve as a discovery interface for the research carried out in Catalan universities. The information must also be made visible to the European Open Science Cloud,³ in addition to other organisms or products that researchers may use, such as OpenAIRE and Google DataSearch.

In addition, the repository must be accepted as a discovery tool in R3Data, the world registry of research data repositories.

³ The requirements to participate and make research data findable through the EOSC are currently being established. For this reason, the repository will have to adapt its characteristics as the requirements are developed.

FR₁₂: To use standard communication protocols

The information contained in data repositories is also relevant for other information systems (for instance, those of funding bodies), and it must therefore communicate with them.

The repository must therefore allow metadata to be consumed and distributed through the Open Archive Initiative-Protocol for Metadata Harvesting (OAI-PMH). This is an interoperability protocol for collecting metadata from other repositories and allowing a repository's metadata to be reused (Devarakonda, 2011).

FR₁₃: To use standard data formats

Initially, each specific dataset and its associated applications were coded with their own formats, but this did not foster data interoperability or extensibility. For this reason, standard data formats were created to allow the data of one machine to be stored or processed on other machines.

The repository must support two of the most common standard formats, *eXtensible Markup Language* (XML) and *JavaScript Object Notation* (JSON) (DeYoung, 2015).

5.5 Management of special characteristics

FR₁₄: To allow different versions of the same dataset

Digital datasets are much more flexible and are updated periodically (versioned) as new data are collected. However, this involves reproducibility problems for citation (FREYA, undated).

The European project FREYA⁴ is currently studying the best solution for assigning DOIs to versions. Following its recommendations, to help the user navigate between different versions, the repository must assign a DOI to each dataset and add the final version to it. The DOI will always take the user to a landing page with the latest version and will contain a record of all changes.

FR₁₅: To manage different metadata schemes

Metadata are the backbone of data curation (Higgins, 2007), because they inform an object or resource descriptively or contextually. They include a series of types: descriptive, technical, administrative, management and preservation metadata. Standard metadata schemes include all

⁴ The FREYA project is funded by the European Commission under the Horizon 2020 programme and aims to expand the infrastructure of persistent identifiers (PIDs) as a basic component of open research, both in the European Union and worldwide

or some of these elements and are used to make it easier for the community to use data from other researchers.

The repository must allow different metadata schemes to be used, from a general and basic scheme (such as Dublin Core) to ones for specific disciplines. Because metadata requirements may vary according to disciplines, the repository must be flexible and adapt to each content.

Regardless of the scheme that is chosen, the repository must include metadata of funding bodies, projects and other associated information. In addition, it must link the dataset with related results of articles or other datasets.

Also, bearing in mind that some of the metadata related to a dataset (the funding entity, researchers, etc.) are often repeated, the repository should allow copying or replication of these data in order to facilitate and streamline tasks.

FR₁₆: Type of access

Following the premises of the EC (European Commission, 2016), the data must be as open as possible and as closed as necessary. Therefore, the repository must allow open publication of data by default or, if necessary, allow a reasonable embargo period.

In addition, in cases in which legal or ethical reasons prevent sensitive or confidential data from being shared openly, the repository must allow them to be deposited in closed access but inform that they are deposited there. The access controls must always be proportional to the type of data and the level of confidentiality. The repository must allow access to some of these data through access controls by IP address, invitation, etc.

The repository will allow a single sign-on, so that users only have to sign on once to access various computer applications or websites. The CSUC universities use UNIFICAT, in addition to other systems.

FR₁₇: To accept any type of format

The format and the software with which the research data are created tend to depend on how the researchers collect and analyse the data, which is in turn often determined by rules and customs specific to each discipline (UK Data Service, 2014). For this reason, the repository must accept any type of format that emerges in any discipline.

Some proprietary formats, such as Microsoft Excel and SPSS, are widely used and are likely to be accessible for some time. Therefore, the safest option to guarantee long-term access to data is to use open standard formats.⁵

⁵ The UK Data Service has established the recommended and accepted formats for each type of data. For example, though CSV files are recommended for tabulated data, XLS files are also accepted.

FR₁₈: To allow different types of ingestion

The repository must allow the transfer of data through different types of ingestion, including centralized ingestion, delegated ingestion (through university research support staff), self-archiving (by researchers themselves) and batch ingestion.

FR₁₉: To offer the recommended citation

Citation is one of the basic pillars of scientific and academic publication. Citation of data, as well as citation of other sources and trials, is a good practice and is part of the academic ecosystem that promotes the reuse of data (Data Citation Synthesis Group, 2014).

The repository must offer the recommended citation following the standards that are currently being established in work communities such as the Data Citation WG⁶ of the Research Data Alliance (RDA). In addition, citing of data will allow us to measure their use through data-level metrics, as is done with article-level metrics.

FR₂₀: To allow dissemination of datasets through social networks

More and more researchers want to disseminate their research data through social networks to promote reproducibility and compare research results (Weller & Kinder-Kurlanda, 2016). In addition, social networks allow researchers to communicate in real time and know what other researchers think.

For this reason, the repository must allow the use of the APIs of the most widely used social networks, such as Twitter, Facebook and LinkedIn.

FR₂₁: To manage different types of licences

When research data are published, it is mandatory to grant a licence, which may range from the most open to the most restrictive. The EC states that ‘As far as possible, projects must [...] take measures to enable third parties to access, mine, exploit, reproduce and disseminate [...] research data. One straightforward and effective way of doing this is to attach Creative Commons Licences (CC BY or CC0) to the data deposited.’

However, there are many ways to do this (Ball, 2014), so the repository must allow different types of licences (such as the GNU GPL) to be managed and granted. It must also explain to its users the implications of granting one licence or another.

⁶ The RDA working group on data citation aims to bring together experts to discuss the problems, requirements, advantages and shortcomings of existing approaches to dataset citation.

FR₂₂: To offer analytical data on the use of the platform

On an online platform, the statistics module is essential to monitor usage. The repository must offer statistics of use at the institutional level, by department, by researcher and by dataset.

It must also show the number of times that datasets are downloaded and cited.

FR₂₃: To provide metadata for reuse

The data repository must be able to provide metadata via APIs for reuse, grant a CC0 licence for reuse of metadata, and state the access and reuse policies on the website.

FR₂₄: User-friendliness

The user-friendliness of a data repository is important in order to ensure that users can access, upload, download and cite the data easily. Therefore, the repository must have a simple and intuitive interaction with the web interface, following the parameters of responsive design to adapt to any type of screen.

User-friendliness must also be applied to stewardship, so that university research support staff can easily manage their collections from the institutional level to the level of researchers.

The repository must also preview the datasets on the platform so that it is not necessary to download the file to view them. The preview must be for all open source file formats, and as far as possible for the other formats. The repository must also have mechanisms for viewing data contained in dehydrated or compressed files.

Finally, the repository must use a consensual terminology that is understandable to the research community.

FR₂₅: To comply with current legislation

One of the reasons why it is necessary to store data is to ensure compliance with current legislation..

6. Best practices

The interviews with experts held to prepare this report were intended to collect a list of functional requirements for the repository. However, in the first meetings and interviews, the experts stated that, in addition to the technical requirements, it was essential to develop best practices for research data management.

The access and reuse of research data does not depend only on the characteristics of the repository where they are published. In many cases the potential utility of data is associated with their availability and that of their metadata. Whereas scientific articles have a long history of publication, very few data have yet been published and shared. In addition, the type, size and format of data vary greatly, so the lack of standardized or consolidated practices is one of the main obstacles for data management.

Because of the lack of experience in data management, best practices have yet to be established, and can only be acquired through practical experience with a data repository. Although the first function of the repository is to publish data, the second, no less important, is to generate expertise and best practices from having it operational.

The best practices that the experts consider should be acquired and established can be divided into the following categories:

- Data curation
- Dataset selection
- Encouraging the use of open formats
- Using widely-accepted controlled standards, protocols and vocabularies

6.1 Data curation

Research data curation is the process of managing data throughout their life cycle so that they can be available and reusable in the long term. Curation involves discovering, identifying, selecting, obtaining, verifying, analysing, managing, archiving, publishing and citing.

All these activities require the data to be correctly documented in order to guarantee the transparency and reproducibility of the studies in the future for the researchers and others (Radboud Univeristy, 2018).

For this reason, the data must be accompanied by the correct documents, including files that explain the context (how the research was done, including records of versions, laboratory notes, standardized methodologies or protocols, software, etc.), the structure (often readme.txt files that contain an overview of the folders and files) and the content (the concepts, their meanings, the numerical values, and the classification codes and schemes used).

6.2 Dataset selection

Once the data have been collected or generated, the resulting files can be updated several times until they reach the final version for publication. It is therefore necessary to establish protocols and criteria for selecting the final data that must be preserved in the long term, as well as the reasons to rule out some other.

The criteria include uniqueness or repeatability, value, quality, cost of reproduction, risk of loss, and indications for reuse in publications or by other users (Tjalsma & Rombouts, 2011).

Catalan universities will have to provide information and guidelines in order to help researchers decide what data they should preserve, delete and publish in the repository.

6.3 Encouraging the use of open formats

In order to ensure that data are usable and retrievable in the long term, selecting the file format is essential and will always be associated with a software and a hardware. Open or non-proprietary formats are those in which the software code is available free of charge for anyone to use in their own software without any limitation on reuse (Biblioteca de la CEPAL, 2019).

Therefore, researchers should be encouraged to use non-proprietary formats and open and documented standards that are commonly used within their community and transmitted through standard representation formats (ASCII and Unicode), and that are not encrypted or compressed. To facilitate the work of researchers, it would be advisable to develop information materials on best practices regarding formats.

Whenever possible, proprietary formats should be migrated to open formats.

6.4 Using widely-accepted controlled standards, protocols and vocabularies

Each discipline has specialized metadata schemes, protocols and vocabularies. To ensure the understanding and reuse of the data, they must be described with appropriate metadata standards for the discipline.

Researchers should also be helped to use controlled vocabularies in the metadata to promote interoperability between different systems.

7. Final recommendations

According with the information gathered and the comments laid out above, the following recommendations are made:

1. To immediately create a repository in Catalonia for publication of research data according to FAIR requirements in order to develop expertise and best practices in research data management. This repository must meet the following requirements:
 - a. Meet the current FAIR requirements and any that are established by the EC in the immediate future.
 - b. Offer added-value benefits in comparison with the current options, such as the assignation of DOIs, interoperability with the Research Portal of Catalonia and enhanced preservation.
2. The repository that is created must
 - a. comply with the requirements mentioned in Section 5 of this document, which are considered both minimal and reasonable, and
 - b. use available open-source software.
3. In parallel to the start-up of the repository, the following must be done:
 - a. The service must be publicized and researchers must be informed of the importance of publishing research data in open access and making DMPs.
 - b. The service must be publicized in a coordinated way, all units conducting research in universities and centres must participate, and the research offices and services must be actively involved.
 - c. Training must be provided on the concepts of open science in general and research data management in particular. In accordance with the guidelines of the EC's FAIR Data Expert Group, this training should include all the members of the university community but, to be effective, it should distinguish between advanced users, young researchers and university support staff.

References

- Amorim, R. C. (2017). A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, 16(4), 851-862.
- Ayris, P., Bernal, I., Cavalli, V., & al, e. (2018). *LIBER Open Science Roadmap*. doi:10.5281/zenodo.1303002
- Ball, A. (2014). *How to licence Research Data*. Digital Curation Center. Recollit de http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_To_License_Research_Data.pdf
- Biblioteca de la CEPAL. (2019). *Formatos abiertos y cerrados*. Recollit de Gestión de datos de investigación: <https://biblioguias.cepal.org/gestion-de-datos-de-investigacion/formatos>
- Couto Corrêa, F. (2016). *Gestión de datos de investigación*. Barcelona: UOC.
- CRUE. (2019). *Compromiso de las universidades ante la Open Science*. Recollit de http://www.crue.org/Documentos%20compartidos/Informes%20y%20Posicionamientos/2019.02.20-Compromisos%20CRUE_OPENSCIENCE%20VF.pdf
- CRUE. (2019). *Compromisos de las universidades ante la Open Science*. Recollit de http://www.crue.org/Documentos%20compartidos/Informes%20y%20Posicionamientos/2019.02.20-Compromisos%20CRUE_OPENSCIENCE%20VF.pdf
- CSUC. (2016). *Data management plans: Version 2, December 2016*. RECERCAT. Recollit de <http://hdl.handle.net/2072/270395>
- CSUC. (2016). *Plans de gestió de dades: Versió 2, Desembre 2016*. RECERCAT. Recollit de <http://hdl.handle.net/2072/273194>.
- CSUC. (2016). *Proposta per establir una política d'accés obert a les dades de recerca a les Universitats de Catalunya (Doc.16/30)*.
- CSUC. (2016). *Universitats catalanes acorden l'ús de l'identificador ORCID als seus investigadors*. Recollit de <https://www.csuc.cat/ca/novetat/universitats-catalanes-acorden-l-us-de-l-identificador-orcid-per-als-seus-investigadors>
- CSUC. (2017). *Ampliació de prestacions dels repositoris institucionals per dipositar dades (Doc.CO17/03)*.
- CSUC. (2017). *Propostes de treball referent a les dades de recerca obertes (Doc.CO17/11)*.
- CSUC. (2017). *Recomanacions per seleccionar un repositori per al dipòsit de dades de recerca: Versió 3, Maig 2017*. RECERCAT. Recollit de <http://hdl.handle.net/2072/284974>
- CSUC. (2018). *Model de política de gestió de dades de recerca per a una universitat (Doc.CO18/03)*.
- CSUC. (2019). *Gestió de dades de recerca: resultats de l'enquesta de 2018 (Doc.CO18/19)*.
- Data Citation Synthesis Group. (2014). *Joint Declaration of Data Citation Principles*. San Diego CA: FORCE11. doi:10.25490/a97f-egykh

- Devarakonda, R. P. (2011). Data sharing and retrieval using OAI-PMH. *Earth Sci Inform*, 4(1). doi:10.1007/s12145-010-0073-0
- DeYoung, L. (2015). *An Analysis of XML and JSON*. Recollit de COMP 150-IDS: 2015 Spring Term Final Papers: https://www.cs.tufts.edu/comp/150IDS/final_papers/lizzied.3/FinalReport.html
- Dutch Ministry of Education, Culture and Science. (2017). *National Plan Open Science*. doi:10.4233/uuid:9e9fa82e-06c1-4d0d-9e20-5620259a6c65
- EOSC (2017). *European Open Science Cloud Declaration*. https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf
- European Commission. (2013). *Digital science in Horizon 2020*. Recollit de <https://ec.europa.eu/digital-single-market/en/news/digital-science-horizon-2020>
- European Commission. (2016). *Guidelines on FAIR Data Management 2020*. Recollit de http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- European Commission. (2016). *Open innovation, open science, open to the world: A vision for Europe*. doi:10.2777/061652
- European Commission. (2016). *Open Research Data in Horizon 2020*. Recollit de http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf
- European Commission. (2018). *Open Science Policy Platform Recommendations*. doi:10.2777/958647
- European University Association. (2018). *EUA Roadmap on Research Assessment in the Transition to Open Science*. Recollit de <https://eua.eu/downloads/publications/eua-roadmap-on-research-assessment-in-the-transition-to-open-science.pdf>
- FREYA. (sense data). *Case study: Versioning with identifiers*. Recollit de <https://project-freya.readme.io/docs/creating-dois-and-doi-metadata>
- GO FAIR. (2018). *FAIR Principles*. Recollit de <https://www.go-fair.org/fair-principles/>
- Government of the Netherlands. (2016). *Amsterdam Call for Action on Open Science*. Recollit de <https://www.government.nl/binaries/government/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science/amsterdam-call-for-action-on-open-science.pdf>
- Government of the Republic of Slovenia. (2015). *National strategy of Open Access to scientific publications and research data in Slovenia (2015-2020)*. Recollit de http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/Znanost/doc/Zakonodaja/Strategije/National_strategy_for_open_access.pdf
- Higgins, S. (2007). *What are Metadata Standards*. Recollit de Digital Curation Center: <http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards>
- Hirtle, P. (2001). OAI and OAIS: what's in a name? *D-lib magazine*, 7(4).
- Johnston, L. (2017). *Curating research data: A handbook of current practice*. Chicago: Association of College and Research Libraries.

- Kim, S. (2018). Functional Requirements for Research Data Repositories. *International Journal of Knowledge Content Development & Technology*, 8(1), 2-36. doi:10.5864/IJKCT.2018.8.1.025
- LEARN. (2017). *LEARN Toolkit of Best Practice for Research Data Management*. doi:10.14324/000.learn.00
- Lee DJ, S. B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLoS ONE*, 12(3). doi:10.1371/journal.pone.0173987
- LERU. (2018). *Open Science and its role in universities: a roadmap for cultural change*. Recollit de <https://www.leru.org/files/LERU-AP24-Open-Science-full-paper.pdf>
- Ministère de l'enseignement supérieur, de la recherche et de l'innovation. (2018). *National plan for Open Science*. Recollit de https://libereurope.eu/wp-content/uploads/2018/07/SO_A4_2018_05-EN_print.pdf
- Ministry of Education, Science and Technological Development. (2018). *Open Science Platform*. Recollit de <https://www.openaire.eu/blogs/serbia-has-adopted-a-national-science-policy>
- OCDE. (2004). *Declaration on Access to Research Data from Public Funding*. Paris. Recollit de <http://goo.gl/Iovbt7>
- Phillips, M. B. (2013). The NDSA levels of digital preservation: Explanation and uses. *Archiving Conference Society for Imaging Science and Technology*, 1, 216-222.
- Presidência do Conselho de Ministros. (2016). *Resolução do Conselho de Ministros n.º21/2016 para a implementação de uma Política Nacional de Ciência Aberta*. Recollit de <https://dre.pt/pesquisa/-/search/74094659/details/maximized>
- Radboud Univeristy. (2018). *Documenting data*. Recollit de Research data management: <https://www.ru.nl/rdm/processing-data/documenting-data/>
- The Ministry of Education and Culture's Open Science and Research Initiative. (2014). *The Open Science and Research Roadmap*. Recollit de <http://www.avointiede.fi/>
- Tjalsma, H., & Rombouts, J. (2011). *Selection of Research Data: Guidelines for appraising and selecting research data*. Data Archiving and Networked Services (DANS). Recollit de <https://dans.knaw.nl/en/about/organisation-and-policy/legal-information/DANSselectionofresearchdata.pdf>
- UK Data Service. (2014). *Formatting and organising research data*. Recollit de <https://www.ukdataservice.ac.uk/media/440281/formattingorganising.pdf>
- Weller, K., & Kinder-Kurlanda, K. (2016). A manifesto for data sharing in social media research. *Proceedings of the 8th ACM Conference on Web Science* (p. 16-172). Hannover: ACM. doi:10.1145/2908131.2908172
- YERUN. (2018). *YERUN Statement on Open Science*. Recollit de https://www.yerun.eu/wp-content/uploads/2018/05/YERUN_OpenScience_Statement-3.pdf

Appendix 1 – FAIR principles

In 2016 the ‘FAIR Guiding Principles for scientific data management and stewardship’ were published in Scientific Data. The authors sought to provide guidelines for improving research, accessibility, interoperability and reuse of digital resources. These principles take into account the potential of machines, since humans increasingly rely on support to deal with data as a result of the increasing volume, complexity and speed of data creation.

The principles:

- To be Findable:
 - F1. (meta)data are assigned a globally unique and eternally persistent identifier.
 - F2. data are described with rich metadata.
 - F3. (meta)data are registered or indexed in a searchable resource.
 - F4. metadata specify the data identifier.
- To be accessible:
 - A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
 - A1.1 the protocol is open, free, and universally implementable.
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
 - A2 metadata are accessible, even when the data are no longer available.
- To be interoperable:
 - I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
 - I2. (meta)data use vocabularies that follow FAIR principles.
 - I3. (meta)data include qualified references to other (meta)data.
- To be re-usable:
 - R1. meta(data) have a plurality of accurate and relevant attributes.
 - R1.1. (meta)data are released with a clear and accessible data usage license.
 - R1.2. (meta)data are associated with their provenance.
 - R1.3. (meta)data meet domain-relevant community standards.

Appendix 2 – The experts

1. Members of the Commission

| | |
|------------------------------------|--|
| Ignasi Labastida i Juan | Universitat de Barcelona. Head of the Office for Knowledge Dissemination and the Research Unit at the CRAI |
| Jordi Hernández Sánchez | Universitat Autònoma de Barcelona. Commissioner of the Rector for Information and Communication Technologies |
| Anna Rovira Fernández | Universitat Politècnica de Catalunya. Head of the Research Support Unit at the Library, Publications and Archives Services |
| Antoni Borràs i Escorihuela | Universitat Pompeu Fabra. Project manager at Computing Service |
| Brigit Nonó Rius | Universitat de Girona. Head of Project development unit at Library |
| Leticia Carro de Diego | Universitat de Lleida. Head of Research Area |
| José Luis González | Universitat Rovira i Virgili. Head of CRAI Scientific Production Management Section |
| Rosa Padrós Cuxart | Universitat Oberta de Catalunya. Expert of Research Library |
| Mireia Salgot | Universitat de Vic-Universitat Central de Catalunya. Library director |
| Anna Caellas Camprubí | Universitat Ramon Llull. Expert of Research and Innovation Office |

2. Spanish experts

| | |
|--|--|
| Ernest Abadal i Falgueras | Universitat de Barcelona. Head of Information, Communication and Culture Research Centre |
| Lluís Alfons Ariño Martín | Universitat Rovira i Virgili. Head at IT Services |
| Mercè Cabo Rigol | Universitat Pompeu Fabra. Vice-Chancellor of the Services, Technology and Information Resources Department |
| Anna Maria Casaldàliga Riera | Universitat Pompeu Fabra. Deputy director at the Library |
| Eva Estupinyà Pinyol | Universitat de Lleida. Head of Services to Users at the Library |
| Jorge García Pérez | Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria. Head of section in the Computing Service |
| Alexandre López-Borrull | Universitat Oberta de Catalunya. Director of the Degree in Information and Documentation |
| Manuel Lozano Nebro | Universitat Pompeu Fabra. Head of Computing Service |
| Teresa Malo de Molina Martín-Montalvo | Universidad Carlos III de Madrid. Library director |
| Eva María Méndez Rodríguez | Universidad Carlos III de Madrid. Deputy Viceminister of Scientific Policy |
| María Fernanda Peset Mancebo | Universitat Politècnica de València. Researcher and librarian |
| Antonio Juan Prieto Jiménez | Universitat Politècnica de Catalunya. IT developer at Library Service |
| Marta Renato | Barcelona Supercomputing Center. RES Officer and project suport coordinator |

| | |
|---------------------------------|--|
| Sandra Reoyo Tudó | Consorci de Serveis Universitaris de Catalunya. Open Science Manager |
| Pilar Rico Castro | Fundación Española para la Ciencia y la Tecnología (FECYT). Head of the Open Access Unit, Repository and Magazines |
| Laia Ros i Blanco | Universitat Ramon Llull. Head of the Research and Innovation Office |
| Antonio Sánchez-Padial | Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria. Head of the Biometrics Service |
| Jordi Sorribas Cervantes | Centre Mediterrani d'Investigacions Marines i Ambientals (CMIMA), CSIC. Director |
| Miquel Térmens Graells | Universitat de Barcelona. Dean of the Faculty of Library Science and Documentation |
| Nadia Tonello | Barcelona Supercomputing Center. Data management manager |
| David Vicente Dorca | Barcelona Supercomputing Center. User suport manager |
| Ricard de la Vega Sivera | Consorci de Serveis Universitaris de Catalunya. Computing and Applications Manager |

3. International experts

| | |
|----------------------------------|---|
| Lucy Amez | Vrije Universiteit Brussel. Policy Advisor Scientific Publications, Bibliometrics & Open Science, Department Research and Data Management |
| Jessica Parlant von-Essen | CSC-IT Center for Science in Finland. Senior coordinator |
| Joakim Phillipson | University of Stockholm. Research Data Analyst |
| Jan van Mansum | DANS. Software developer coordinator |
| Linda Reijnhoudt | DANS. Software developer |
| Eloy Rodrigues | Universidade do Minho. Director Documentation Service |
| Jääarno Saarti | University of Eastern Finland. Library director |

Appendix 3 – Countries

Flemish Belgium

The Vlaamse Interuniversitaire Raad ([VLIR](#)) is an advisory body that represents the five Flemish universities (Katholieke Universiteit Leuven, Universiteit Antwerpen, Universiteit Gent, Universiteit Hasselt and Vrije Universiteit Brussel) to facilitate interuniversity cooperation and interaction with the Flemish government.

Through a working group, the VLIR prepared a survey to determine the needs and requirements of a research data management service. They have launched an online tool for developing DMPs and have published the report [Research Data Management en de Vlaamse Universiteiten: White Paper](#) with recommendations on investments that the government should make in infrastructure, education, legislation and incentives.

The VLIR recommends providing a shared sustainable infrastructure for archiving and preserving data.

Finland

Fairdata Services are a set of interoperable tools based on the data life cycle: storage and preservation (IDA), description for publication (Qvain) and discovery (Etsin). All these services depend on the Ministry of Education and Culture and were developed by the [CSC-IT Center for Science Ltd.](#)

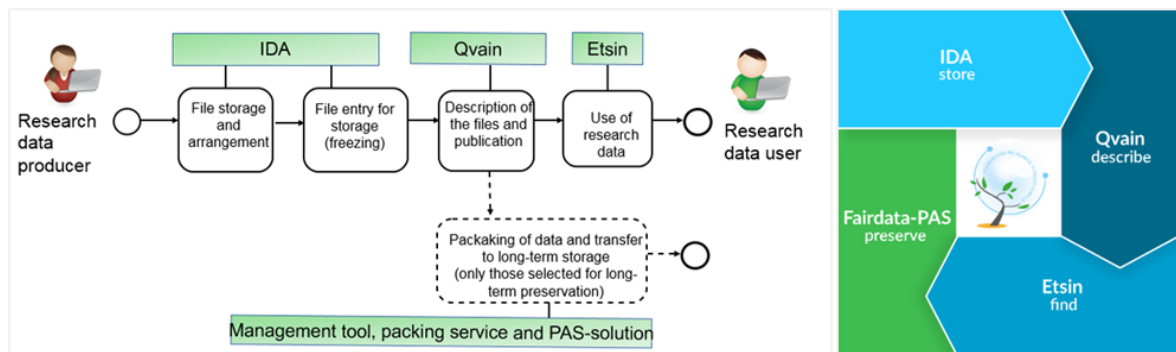


Figura 1. Fairdata Services in Finland

[IDA](#) offers reliable storage for research data during the active phase of research. The selected data are frozen to an immutable state that is valid for publication. Technical metadata are generated on the files and they are transferred to Qvain.

[Qvain](#) is a tool for creating metadata to allow the data to be published. The datasets are given a persistent identifier and a landing page. The finalized metadata records are published in Etsin.

[Etsin](#) is a data discovery tool that harvests metadata from Qvain and from external resources.

The Netherlands

Data Archiving and Networked Services ([DANS](#)) is an institute of the Dutch Academy KNAW funded by the Netherlands Organization for Scientific Research (NWO). DANS has the mission of promoting and providing permanent access to digital research resources and encourages researchers to make their publications and research data searchable, accessible, interoperable and reusable.

DANS coordinates the services in the ‘back office’ and the universities offer support to the users in the ‘front office’. The services are organized according to the data life cycle: DataverseNL while the project is ongoing, EASY when the project is completed to preserve the data, and NARCIS for discovery.

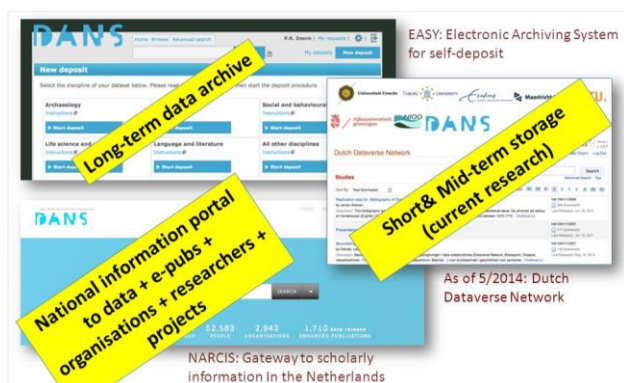


Figura 2. The services offered by DANS

[DataverseNL](#) was set up in 2014 and includes 12 institutions (Eindhoven University of Technology, Leiden University, Delft University of Technology, Maastricht University, NIOO-KNAW, Protestantse Theologische Universiteit, Tilburg University, the University of Groningen, the University of Twente, Utrecht University, the University of Applied Sciences Utrecht and Vrije Universiteit Amsterdam). This repository offers online storage and sharing of research data up to 10 years after its completion and uses the Handle system for persistent identifiers. The institutions pay a fee to participate.

[EASY](#), set up in 2005, is a certified online preservation repository with the Data Seal of Approval, World Data System certification and the Nestor Seal, which ensure compliance with a set of transparent criteria regarding quality, sustainability and accessibility. Unlike DataverseNL, this repository assigns DOIs as persistent identifiers.

Finally, [NARCIS](#) is the national portal for scientific information, including research data.

Portugal

The Fundação para a Ciência e a Tecnologia ([FCT](#)) is the national public agency for supporting research in science, technology and innovation in all areas of knowledge. It depends on the Portuguese Ministry of Science, Technology and Higher Education.

This institution drew up a proposal for a consortial data repository that has been placed on standby because it has not obtained funding. The Universidade do Minho is therefore implementing its own institutional data repository.

Sweden

The SND Consortium is made up of seven universities (Göteborg Universitet, Karolinska Institutet, Lunds Universitet, Sveriges lantbruksuniversitet, Stockholm Universitet, Umea Universitet and Uppsala Universitet) and works under the Swedish Research Council. The universities of the Consortium have developed best practices and knowledge in several disciplines through specialists who have served as a link between researchers and the support offices in each university.

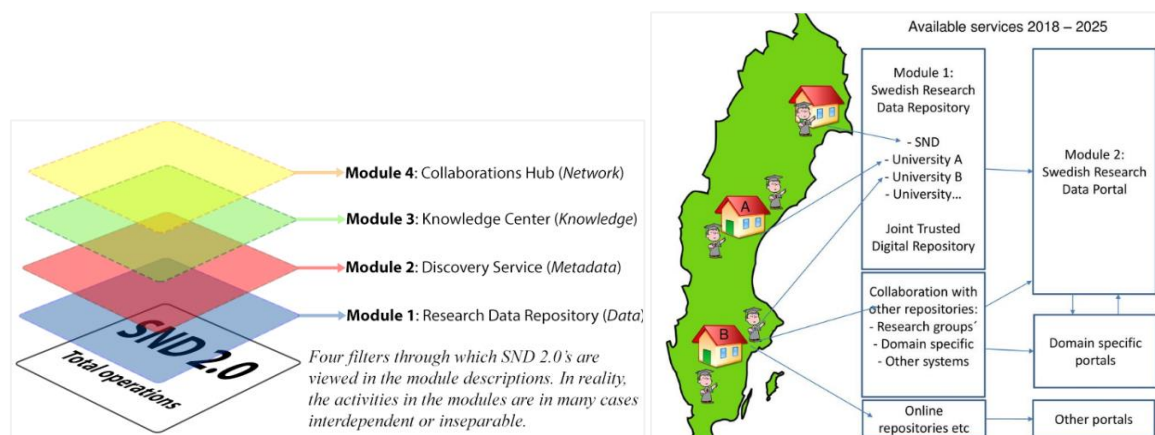


Figura 3. Modules of SND 2.0

El Swedish National Data Service 2.0 ([SND 2.0](#)) is divided into four modules: the Research Data Repository (for data), the Discovery Service (for metadata), the Knowledge Centre (for knowledge) and the Collaborations Hub (to generate a network).

The Swedish Research Data Repository ([SRDR](#)) offers the services of ingestion, curation, access and quality assurance. At this stage, best practices are offered on dataset versions and DOIs are assigned as persistent identifiers, because citation of the data through the DOI is much easier and allows it to be indexed like any other type of publication.

This repository is operated jointly with the Swedish Research Data Discovery Service, which aims to be the sole search portal for Swedish research data. The intention is to collect the datasets of both the national repository and other disciplinary repositories.

At present, EUDAT is used for storage and preservation. It will be sufficient for the next few years and is free of charge. However, a pilot project has been set up with the major data-producing centres in Sweden to develop a common solution.

References

CSC-IT (2019). *CSC-IT Center for Science Ltd.* <https://www.csc.fi/>

DANS (2019). *Data Archiving and Network Services.* <https://dans.knaw.nl/en>

DANS (2019). *DataverseNL.* <https://dataverse.nl/>

DANS (2019). *EASY*. <https://easy.dans.knaw.nl/ui/home>

DANS (2019). *NARCIS*. <https://www.narcis.nl/>

FAIRDATA.FI (2019). *Etsin*. Research dataset finder. <https://www.fairdata.fi/en/etsin/>

FAIRDATA.FI (2019). *IDA*. Research data storage service. <https://www.fairdata.fi/en/ida/>

FAIRDATA.FI (2019). *Qvain*. Research dataset metadata tool. <https://www.fairdata.fi/en/qvain/>

FCT (2019). *Fundação para a Ciência e a Tecnologia*. <https://www.fct.pt/>

SND (2018). *Description of the infrastructure and its activities*. [SND 2.0](#))

SND (2019). *Swedish National Data Service*. <https://snd.gu.se/en>

UNIFI (2018). *Open Science and Data. Action programme for the Finnish scholarly community*. <http://urn.fi/URN:NBN:fi-fe2018111648265>

VLIR Werkgroep Research Data Management & Open Science (2018). *Research Data Management en de Vlaamse Universiteiten: White Paper*. http://www.vlir.be/media/docs/Onderzoeksbeleid/20180525%20White%20Paper_RDM%20en%20de%20Vlaamse%20Universiteiten_addendum.pdf

VLIR (2019). *Vlaamse Interuniversitaire Raad*. <http://www.vlir.be/>

Appendix 4 – Additional documentation

In addition to the documents available in the reference section of these reports, the following documents related to FAIR data were also consulted:

Allen, R. & Hartland, D. (2018). *FAIR in practice*. JISC report on the Findable Accessible Interoperable and Reusable Data Principles. DOI: [10.5281/zenodo.1245568](https://doi.org/10.5281/zenodo.1245568)

Austin et al (2015). Research Data Repositories: review of current features, gap analysis, and recommendations for minimum requirements. *IASSIST Quarterly* 39(4). DOI: [10.29173/iq904](https://doi.org/10.29173/iq904)

COPDESS (2018). *Enabling FAIR data commitment Statement in the Earth, Space, and Environmental Sciences*. <http://www.copdess.org/enabling-fair-data-project/commitment-to-enabling-fair-data-in-the-earth-space-and-environmental-sciences/>

CORE TRUST SEAL (2018). *Core Trustworthy Data Repositories Extended Guidance*. <https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf>

European Commission (2018). *Commission staff working document. Implementation Roadmap for the European Open Science Cloud*. Brussels. SWD(2018) 83 final

Hodson et al. (2018). *FAIR Data Action Plan*. Interim recommendations and actions from the European Commission Expert Group on FAIR data. DOI: [10.5281/zenodo.1285290](https://doi.org/10.5281/zenodo.1285290)

Hodson et al. (2018b). *Turning FAIR data into reality*. Interim report of the European Commission Expert Group on FAIR data. DOI: [10.5281/zenodo.1285272](https://doi.org/10.5281/zenodo.1285272)

Principe & Rodrigues (2018). *Data RepositórioUM: Projeto de implementação do repositório de dados para a Universidade do Minho*. 4ª Fórum de Gestão de Investigação

Rodrigues, E. (2019). *Definición e implementación de estrategias y servicios institucionales para la gestión de datos de investigación*.

SPARC (2018). *FAIR and Open Data: a briefing for policymakers and senior managers*. <https://sparceurope.org/new-briefing-paper-explores-fair-and-open-data/>

Appendix 5 – Assigning DOIs

In order to create new DOIs and assign them to datasets, it is necessary to be a DataCite community member or to collaborate with one of the members.

Organizations of all types can become members, including data centres, publishers and libraries, if they show their support to sharing research data in the following way:

- By demonstrating a high level of commitment to research data and open science.
- By forming part of a global community of data dissemination, learning, collaborating and defending with a leading-edge network of data research experts .
- By supporting and participating in the creation and management of persistent identifiers (DOIs) for research results.
- By playing a critical role in the progress of data sharing.

To join the community, you must present⁷ an application, which will be evaluated by the directors of DataCite, and pay a subscription.

⁷ https://datacite.org/assets/datacite_application.pdf

Appendix 6 – Hardware and software for a data repository

A data repository capable of meeting the functional requirements of Section 5 of this document requires a combination of storage hardware and software.

6.1 Hardware

The repository must have an elastic capacity and at least two geographically remote copies. Current disc cabinets facilitate this elasticity, and if at least two of them are available separately, both requirements can be covered. The storage infrastructure can be local to a data processing centre that has sufficient security measures and availability, or in the cloud.

The storage costs will recur and grow as data are uploaded to the repository for preservation.

6.2 Software

Several options have been analysed in studies (Amorim, 2017; CSUC, 2017; Rodrigues, 2019) that compare requirements similar to those detailed in Section 5 of this document. There are two types of software:

- software of institutional repositories or transparency portals that have been adapted to include research data (such as DSpace, ePrints and CKAN) and
- software made specifically for research data.

Following the recommendations of this report, among the latter we have focused only on those that can store data on local servers. We mention the main features of three software programs, which to a greater or lesser extent, and directly or through plugins, meet all the requirements detailed in Section 5 of this document. These are the following:

- Figshare as a commercial option
- Dataverse as an open source option
- Invenio as an open source option using the adaptation made by EUDAT

In order to increase compliance with the preservation requirements, follow the OAIS model and allow the data to be stored in at least two geographically separate cabinets, the above software should be complemented with other specialized software for these functions.

Appendix 7 – Glossary

| Item | Description |
|------------------------------------|---|
| European Open Science Cloud | The European Open Science Cloud is promoted by the European Commission to provide all researchers, innovators, companies and citizens with seamless access to an open-by-default, efficient and cross-disciplinary environment for storing, accessing and reusing data, tools, publications and any resource. |
| FAIR | <i>Findable, Accessible, Interoperable and Reusable</i> . Not only for research data, but also for all research results. Fair data aims to facilitate the discovery, integration and analysis of relevant data searches and the algorithms and workflows associated with them. |
| Research data management | The development, execution and supervision of the plans, policies, programmes and practices that control, protect, deliver and improve the value of research data. |
| Interoperability | A process that allows users to share data between different organizations. The goal is to create a shared understanding of data. |
| Licence | Describes the terms in which a material can be reused, stored, redistributed, etc. |
| Metadata | Describes the basic characteristics of the data. It usually includes authorship, title, date, summary, keywords and license information. |
| Data management plan | A formal document that describes how data must be managed throughout the entire life cycle. |
| Repository | An infrastructure that allows the persistent, efficient and sustainable storage of digital objects. |



Consorci de
Serveis Universitaris
de Catalunya