DATA MANAGEMENT IN R&D&I CENTRES

XUNTA
DE GALICIA

Institució
CERCA
Centres de Recerca
de Catalunya

GOSCI

GUIDE TO DATA MANAGEMENT IN R&D&I CENTRES

TRAINING PROGRAMME IN STRATEGIC AND OPERATIONAL MANAGEMENT OF RESEARCH CENTRES AND UNITS OF EXCELLENCE

An initiative of the Consellería de Cultura, Educación e Universidade in collaboration with FEUGA

This document constitutes a deliverable of the Training Programme in Strategic and Operational Management of Research Centres and Units of Excellence (GESCI), funded by the Consellería de Cultura, Educación e Universidade and by the CERCA Institution.

It has been drafted by the Fundación Empresa - Universidad Gallega (FEUGA) with the participation of different experts: Ignasi Labastida, Nadia Tonello and Lluís Rovira. Other collaborators include Faustino Infante and Mª Jesús Tallón, from the Secretaría Xeral de Universidades, and the consulting firm Estratega Consulting Global S.L.

This guide does not represent the opinion or the official position of the Xunta de Galicia on this issue.

Legal deposit C2121 – 2021

Printed in November 2021

www.gesci.es

CONTENTS

Chapter 1

INTRODUCTION

1 INTRODUCTION

The **GUIDE TO DATA MANAGEMENT IN R&D&I CENTRES** is the **fifth deliverable of the "Collection of Guides for the Management of R&D&I Centres"**, developed within the framework of the Training Programme in Strategic and Operational Management of Research Centres and Units of Excellence (GESCI).

The basic objective of this fifth guide is to **review, from a practical perspective, the fundamental elements that should be taken into account when managing research data in research institutions,** providing clear guidelines for action.

This publication is **aimed primarily** at the following groups:

- **Members of governing bodies and management teams** of R&D&I centres responsible for drawing up the strategy and defining management policies at all levels.
- **Research data managers,** as well as non-specialised management staff who fully or partially assume functions related to the management of research data.
- **Research staff with responsibilities in the management of research projects**, who are increasingly required, by funding entities, to submit research data management plans.

The guide is internally organised into separate **chapters which can be read independently of each other.** However, a complete read-through of the guide is recommended to ensure a more comprehensive and in-depth understanding of research data management. The document ends with a **glossary** that includes the main concepts related to the subject.

Chapters have a uniform structure and include an explanation of the topic in question, a series of associated recommendations and, in addition, other relevant information.

Based on the concept of research data and its cycle, and using the *"FAIR* principles" for data management as a frame of reference, this document deals with the following **content**:

- The **management of research data in a centre and its basic concepts.**This chapter begins by describing research data, the different classifications involved, the *FAIR* principles and their implications, the research data management cycle, and the role of the data steward in the centre.
- **Planning.** This chapter focuses on the management of research data, using the data management plan of a project as reference, defining its main contents and identifying some examples and tools available for its drafting and monitoring. It goes on to describe the core protocols used to ensure the uptake of the data management principles in an institution, over and above the data management that is carried out in specific projects, and addresses two main dimensions that should be taken into account in all planning processes: the management of personal data and the management of intellectual property.
- **Working with data.** This section offers a practical and disciplinary cross-cutting perspective on aspects such as the sources of research data, the terms of access, use and publication of data, and the citation of sources before delving into the subject of data formats and the implications on the *FAIR* principles, time and cost involved in its use. The chapter ends with a look at the practical aspects related to the organisation and storage of data.
- **Sharing, publishing and preserving data.** This chapter addresses these steps in the management cycle, delving into the duality of open and closed data and the different types of licence involved in data management and existing repositories.
- The final chapter deals with **estimating the cost** of managing data and developing a **data management policy model.**

This guide offers a range of knowledge, tools and resources aimed at establishing a **culture of research data management** as a foundation of excellence in Galician research centres, in line with the philosophy of the European Research Area and the GESCI programme of which it is part.

Chapter 2

RESEARCH DATA MANAGEMENT: BASIC CONCEPTS

2

RESEARCH DATA MANAGEMENT: BASIC CONCEPTS

Research Data Management (RDM) comprises the decisions regarding activities that have to do with the **life cycle of research data. In other words, the collection, organisation, processing, analysis, preservation and publication of the data used in a research project.** The possibility of reusing the data created in a project for other projects is the characteristic that reactivates the management process and makes it cyclical in nature. In this context, the ***FAIR*** (Findable, Accessible, Interoperable, Reusable) principles were established as a response to the need to standardise the "proper management of **research data**". The following diagram shows the Research Data Life Cycle.

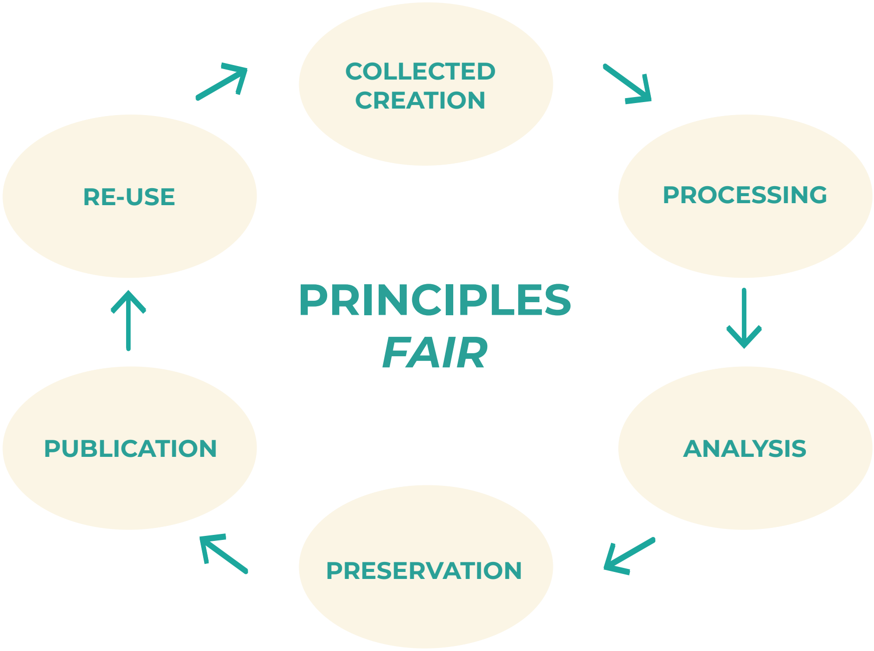


Figure 1: Research data life cycle. Source: compiled by author.

The different types of data and the variety of formats, repositories or disciplines make it difficult to establish a single recommended data management strategy. Each centre must develop its own to ensure adequate management of its data that results in an improvement of its research. Doing so will allow centres to:

- Be more transparent in the validation of research results.
- Ensure that data is findable, accessible, interoperable, and reusable.
- Improve the profile of researchers and the impact and visibility of its projects.
- Promote data protection and minimise the risk of data loss.
- Comply with the legislation on the matter and with the requirements of the funding entities.
- Save time and make efficient use of available resources.

There are **three key figures** involved in an institution's research data management, with the following responsibilities:

- The Data Steward of a centre is the person who **coordinates the institution's data management policy and acts as a link between the research groups,** the centre's policies and those of the funding entities, manages the application of laws or regulations and the infrastructures available to researchers. Their RDM responsibilities range from analysing the data management needs of the group they serve, providing advice, coordinating the interests of stakeholders regarding the data, training and disseminating best practices in management, developing the centre's *RDM* policies, etc.
- **The data analyst or Data Scientist** is the member of the research team **who is in charge of the research data.** They establish the guidelines on how to collect, process and prepare the data for analysis. Their *RDM* responsibilities range from collaborating with other researchers in data collection and analysis, conducting technology watch, creating information analysis and visualisation strategies, identifying problems and providing data analysis solutions, using statistical techniques and data mining models for these purposes, evaluating data collection sources and techniques, staying up to date regarding technologies, techniques and methods, etc.
- **The researcher** is the person who carries out the research activity and who is responsible, within the scope of their own projects, for **data collection** and the **implementation of best practices in** management, **compliance with ethical and legal requirements** and decision-making in accordance with the data management policy. Moreover, **when reusing data, they must consider the conditions** to which they are subject and **respect the creators' right to be cited.**

Finally, if any of the aforementioned roles does not exist, they must manage the research data throughout the entire cycle, assuming the above-mentioned functions.

2.1 Research data

Research data is a **set of information, digitalised in files or on any physical medium, which in the research process contributes to the production of a scientific result.** According to the *CASRAI* consortium, which proposes the establishment of standards to describe research, this type of data is defined as the evidence in a research process that validates its findings and results[1].

Research data can be **figures that have been obtained through an experiment or simulation**, but it can also be **images, recordings, or transcripts of interviews.** Therefore, research data **can have diverse origins and formats,** it can be data that is observed, simulated, collected, processed, etc.

Also, within the framework of a data management process, this data can be grouped into collections or databases. Throughout this guide, the terms data and data sets are used interchangeably, although there are differences when it comes to the management of intellectual property, as will be explained later in the guide.

There are **many ways of classifying data,** depending on the discipline and type of research project. The following table shows **some examples of the most common classifications of data**, with a brief reference to the main requirements that their management entails.

Additional relevant information

[1] https://casrai.org/term/research-data/

CLASSIFICATION CRITERIA	TYPE OF DATA	DESCRIPTION	EXAMPLES	MANAGEMENT REQUIREMENTS
According to the format in which the information is stored	Physical	All research data on paper or other analogue formats	• Text documents and data tables on paper • Photographs, graphs or maps on paper • Inorganic material or organic tissue samples Collections, libraries, sound libraries, materials libraries, picture galleries, etc.	These require a specific and specialised storage system, depending on the type of material in question
	Digital	Data that has been produced by computer or another system with digital output	• Text documents in digital format • Videos/Audios/ Photos in digital format • Graphics/Maps in digital format • Databases • Files of numerical values generated by instruments (measurements) in digital format • Files of values derived from processing/ analysis	These require naming and organising in a storage system in folders and files, although this may depend on the format, the content, the data processing phase, etc.

CLASSIFICATION CRITERIA	TYPE OF DATA	DESCRIPTION	EXAMPLES	MANAGEMENT REQUIREMENTS
According to the level of information processing	Primary	Constitutes the starting point of a research project Data that has not been modified	Data deriving from: • Measurements • Surveys • Observations • Experiments	The production of primary data has to be justified in the Data Management Plan (DMP). It should be stored together with the details of its origin. The authors are responsible for complying with the FAIR principles and for providing sufficient information for their reproducibility.
	Secondary	Data that has undergone some type of processing (intermediate data or results)	• Processed data • Analysis results • Graphical representations • Summaries, annotations, etc.	
According to the type of content	Personal	Making reference to characteristics of people and allowing their identification	• Names and surnames • Postal addresses, telephone numbers, IPs of electronic devices • Genetic, clinical data • Physical characteristics, age • Videos, photos, images	This data is subject to special security measures, according to GDPR and laws of the country of origin
	Non-personal	Refers to people, and if the data in question is persona, it does not allow the individual to be identified	• Anonymised data • Positions or company names • Any data on plants, animals or things	The security and access measures of this type of data are established according to its confidentiality

Source: compiled by author.

RECOMMENDED ACTIONS

- **Identify the research data that each institution must handle and provide its definition and classification for the research staff of each research unit, depending on the scientific discipline.**

2.2 The *FAIR principles*

The origin of the ***FAIR*** (Findable, Accessible, Interoperable, Reusable) principles lies in the article "*FAIR Guiding Principles for scientific data management and stewardship"* published in 2016 in the journal *Scientific Data*. They are a **set of guidelines that must be followed to ensure that research data is accessible, can be understood and can be reused,** also in studies of other disciplines. Ultimately, the *FAIR* principles arose as a **response to the need to standardise and improve the management of research data,** which generally depend on the criteria of the owner of the data. Therefore, data is considered *FAIR* when curated according to these principles. The aforementioned article suggested that research data must comply with the four principles explained below:

- **Be findable:** data is findable **when it is assigned a globally unique and persistent identifier.**This identifier can be a *DOI (Digital Object Identifier)* or a *handle*, which is another type of identifier generally used in repositories for publications. Most data publishing service providers offer *DOI* allocation. Its use is also standard in publishing.

If a centre wants to perform this allocation in its own repository, it must register with a registration agency, such as *DataCite*, a specialist in *DOI* registration for data. There are other registration agencies, like *Crossref*, that specialise in publications.

Both the data and the metadata must be registered and indexed in order to be findable. To do this for a centre's own repository, it should be registered in international registries, especially in the leading portal *Re3data*[2]. If the centre publishes the data in an external service, this registration may have already been carried out.

Certain guidelines and standards should be followed to index the metadata, such as those of the *DataCites* chema, which indicate the required metadata needed to describe a data set, as well as the values or methods used to describe them.

- **Be accessible**: data is accessible **when**

Additional relevant information

[2] https://re3data.org/

the metadata can be retrieved by its identifier using a standardised communications protocol. This protocol must be open, free, and universally implementable and allow for an authentication and authorisation procedure, where necessary. Again, this accessibility may already be established by default in an external repository, but it will have to be activated if it is in the centre's own repository. Finally, metadata is accessible, even when the data is no longer available.

- **Be interoperable:** interoperability is defined **by the use of a formal, accessible, shared, and broadly applicable language for knowledge representation, both in the case of data and metadata.**The metadata uses vocabulary that follows the same *FAIR* principles and includes qualified references to other metadata.
- **Be reusable**: the data is reusable **when the metadata is described with a plurality of accurate and relevant attributes, is released with a clear and accessible data usage licence, is associated with detailed provenance and meets** discipline-relevant community standards

The concepts of **data managed according to the *FAIR principles* and open data** are not equivalent or exclusive, but **complementary**. Although in this sense, there are several initiatives such as the *Open Science* movement promoting good data management that aims to expand the use of open and *FAIR*

The *FAIR* principles do not guarantee that data is open, just as open data is not necessarily data curated under the *FAIR*principles. The level of accessibility is what determines whether the data is open or restricted to certain users. The degree of compliance with these principles is what determines whether the data is *FAIR*. The following diagram illustrates this:

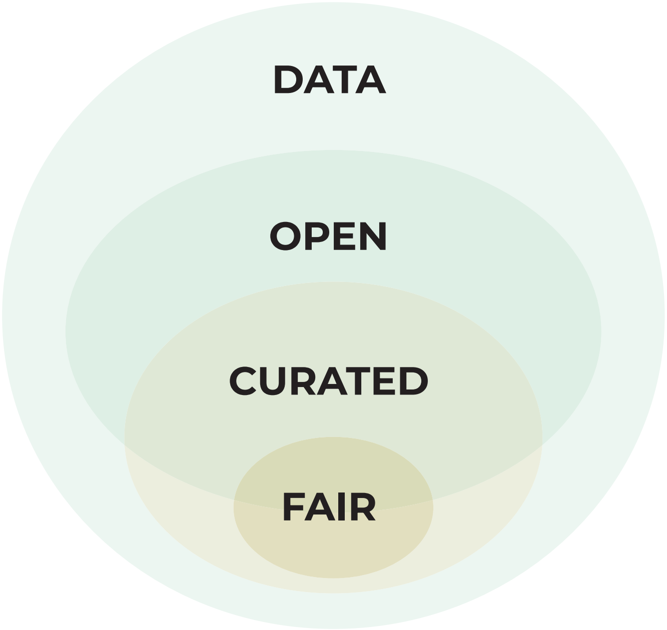


Figure 2: Open, curated and FAIR data. Source: compiled by author.

2.2.1 How to ensure that data follows the *FAIR*

One of the **main objectives required to achieve proper management of research data** is to ensure that the **institution's data comply with the *FAIR* principles to the highest possible degree.** This objective should be taken into account from the start of any research **project**, when **deciding the following:**

- What **formats** are most suitable for working with the data and preserving it.
- What **repository or infrastructure** should be used.
- What **metadata** will be used to describe the data.
- What **persistent identifiers** will be used when publishing the data.

In general, research staff do not have the necessary specific knowledge in data management to adapt the data to the *FAIR*principles, so if **the centre** adopts these principles, **it must provide its staff with:**

- **Training** in the subject.
- **Access to any tools** that may be necessary.
- **Access to specialised experts in data management** who, in addition to giving advice throughout the entire process, can provide their know-how on choosing the right metadata, formats, repositories, etc.

RECOMMENDED ACTIONS

- **Declare, in the centre's data management policy document, its intention and level of commitment to the *FAIR* principles, so that the research data managed is findable, accessible, interoperable and reusable.**

- **Provide support in terms of the resources, personnel, information, etc. needed by the research staff to comply with the application of the *FAIR* principles and inform the staff on the generic, institutional or subject-based platforms and repositories available to them, as well as on the tools they can use to publish and share data following the *FAIR* principles.**

Additional relevant information

- *"FAIR Guiding Principles for scientific data management and stewardship"*: https://www.nature.com/articles/sdata201618
- A more detailed explanation of the *FAIR* principles and sub-principles can be found on the *GoFAIR* portal: https://www.go-fair.org/fair-principles/
- *"FIP mini-questionnaire; Build your FAIR Implementation Profile"*: https://www.go- fair.org/wp-content/uploads/2021/06/FIP-mini-questionnaire.pdf
- Most used metadata schema in data repositories, *DataCite*: https://schema.datacite.org/
- *"EU-LIFE recommendations on Research Data Management in the life sciences"*: https://eu-life.eu/newsroom/news/eu-life-recommendations-research-data-management-life-sciences

2.2.2 Examples of *FAIR data*

Although it is always difficult to identify useful generic examples due to the peculiarities of each discipline or scientific field, the highlighted examples below are portals that offer data following the *FAIR* principles

UK Data Service[3]
This portal offers a data catalogue with a wide variety of metadata, including the use of a *DOI* system and the use of controlled vocabularies. In many cases, access is restricted and requires an account, although this does not preclude the data from following the *FAIR*principles.

Omics DI[4]
This platform offers the localisation of genomic and proteomic data sets, among others. It does not use a *DOI*, but its own identifier or the identifier that the dataset has at source, although, again, the data is widely described and information is provided on its origin. The portal also includes references to the publications associated with the data and offers access through a communication protocol or web service that allows the result of a query to be obtained directly, without the need to download the data.

Mendeley Data[5]
Some platforms or repositories highlight the quality of certain data, such as this portal, which offers a list of the data sets that best comply with the *FAIR* principles, selected by an external committee of researchers.

RECOMMENDED ACTIONS

- Draw up a list of examples of *FAIR* data in specific disciplines and distribute it appropriately among research staff.

2.3 The research data management cycle

Research data management is carried out in parallel with the research process, which means that tasks are carried out during the planning, execution and dissemination of *R&D* results, as the following table shows.
Subsequently, once the project is finished and the results have been published, the data must continue to be managed in accordance with the *FAIR* principles, in order to

Additional relevant information

[3] https://beta.ukdataservice.ac.uk/datacatalogue
[4] https://www.omicsdi.org/
[5] https://data.mendeley.com/collections/46769202-3a48-4097-a6fb-1d056cf0d99d

RESEARCH PHASE	RDM PHASE	DATA MANAGEMENT ACTIVITIES
Start of the R&D project	Start of the R&D project	· Choosing the editing tool · Planning for reviews and updates · Identifying existing data and the permissions/rights to access and use it
	Processing	· Choosing the infrastructure for the preservation of the data created or collected (FAIR activity) · Defining the folder structure and file naming conventions (FAIR)
Research (Analysis)	Analysis	· Choosing and documenting the data, formats and metadata (FAIR)
	Preservation	· Defining the backup and preservation mechanisms. Security and access control: institutional repositories (FAIR)
Publication of results	Publication	· Publishing data in subject or multidisciplinary repositories · Choosing licences for use, data protection and copyright
Evaluation	Data management	· Evaluating management costs · Evaluating and reviewing the management plan

Source: compiled by author.

guarantee effective compliance with these (safe storage, accessibility for other researchers, availability for reuse, etc.).

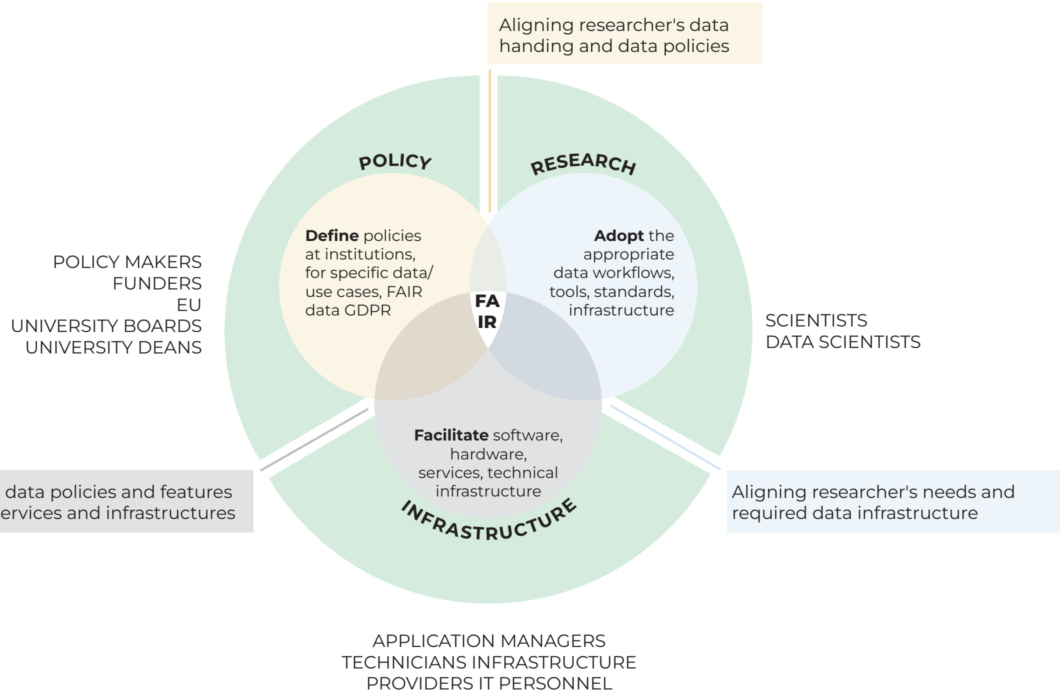
2.4 The role of the data steward

The main function of a **data steward** is to **align the interests of the various stakeholders involved in the research process, through data management.**The data steward is responsible for coordinating **three main implementation areas**:

- The **policy/regulatory area:** the data steward must be aware of the regulations and policies that apply (for example, GDPR, institutional, centre strategy or policy, etc.).
- The **area of the research itself:**the data steward must take into account **what the information flows** are so that the research process is carried out **according to the discipline** in question (for example, the need to share information with external agents), the **necessary tools** (for example, software for data processing or specific laboratory equipment), **standard formats commonly used in the discipline**, etc.
- The **area of the infrastructure and services available:** the data steward must **align the availability** of infrastructure and services both with the needs of the research community and with the identified political/regulatory constraints (for example, the way in which confidential information is processed such as anonymisation or access management, among other aspects; the choice of the type of repository for publishing the data; type of licence, etc.).

The figure below shows the three main implementation areas that fall within research data management: policy, research activity and infrastructure. The study from which this illustration derives was carried out in the field of life sciences, but it applies equally to other fields of knowledge.

IMPLEMENTATION AREAS FOR DATA STEWARDSHIP



Aligning researcher's data handing and data policies
POLICY
RESEARCH
Define policies at institutions, for specific data/ use cases, FAIR data GDPR
Adopt the appropriate data workflows, tools, standards, infrastructure
FA IR
POLICY MAKERS
FUNDERS
EU
UNIVERSITY BOARDS
UNIVERSITY DEANS
SCIENTISTS
DATA SCIENTISTS
Facilitate software, hardware, services, technical infrastructure
INFRASTRUCTURE
Aligning data policies and features of data services and infrastructures
Aligning researcher's needs and required data infrastructure
APPLICATION MANAGERS
TECHNICIANS INFRASTRUCTURE
PROVIDERS IT PERSONNEL

Figure 3: Areas of competence of the data steward. Source: https://doi.org/10.5281/zenodo.3460552

Therefore, the **data steward's job comprises the characteristics of three different roles:**

- That of **researcher**, since they need a certain degree of knowledge of the scientific discipline to be able to understand its language, as well as to know the formats and types of data used.
- That of **IT manager**, to ensure all the data is stored securely in suitable infrastructures, so it can be shared during the execution of the research project.
- That of **documentalist**, in order to be able to describe the data according to the existing metadata schemes.

The creation of the figure of the data manager at the university level is already being implemented in some countries. This is the case of *the Netherlands*, with the example of the *University of Delft (TU Delft)* which has established a data management team made up of specialised professionals in each faculty. These teams advise research staff, provide training, monitor policies, coordinate with technical teams, etc. They are key to the institution's data management[6].

To **ensure proper data management, a research centre must have qualified personnel to carry out the different tasks** that this process encompasses. With adequate training, some tasks can be carried out by the centre's own research staff, but others will require the intervention of specialised professionals. Some groups have data analysts or engineers to do this, although ideally the centre will have a data curator or steward.

Therefore, **the centre must schedule a solid training plan** that can be **complemented with more specific training actions** in certain matters depending on the characteristics of the centre or the projects (for example, how to approach ethical aspects or how to anonymise data). These training sessions can be carried out by specialised personnel, but meetings between researchers can also be organised to exchange best practices and experiences that facilitate the training of the entire organisation.

RECOMMENDED ACTIONS

- **Have staff who specialise in data management, capable of training and guiding research staff in making research data management decisions for their projects, which implies:**
  - Identifying the necessary roles for data management.
  - Create new job profiles or assign data management roles to existing profiles.
- **Have a data management training programme in place.**

Additional relevant information

For more information on the competences of data stewards in the European context of open science, the following report is recommended: *"Digital skills for FAIR and open science: Report from the EOSC Skills and Training Working Group"* https://op.europa.eu/s/pnPe

6 https://openworking.wordpress.com/2020/02/14/a-year-in-the-life-of-a-data-steward/

Chapter 3

PLANNING

3

PLANNING

Before starting any research, a plan should be put in place for the management of the data that will be used. In order to do this, a **data management plan** *(DMP)* is required, which must include the description of the research data and the data processing carried out throughout its life cycle.

This section presents a set of **guidelines so that** research staff **can draft a data management plan and so that centres can offer the support necessary to complete this task.** It also covers other relevant aspects that should be taken into account before starting the research, such as **the management of intellectual property** and the need to **design the research taking into account** the possible impact that the **use of personal data** may have on the privacy of individuals.

In terms of drafting the *DMP*, **if a centre already has defined procedures incorporated into a data management protocol, the description of the activities will be based on what has already been established. Otherwise, each data management plan for each project must contain a definition of what will be done with the data.** It is worth noting in this regard that more and more funding entities require the drafting of a *DMP* and that it is also a requirement in a centre's research data processing policy framework.

3.1 Data management plan

The data management plan (DMP) is a **living document, which undergoes updates and revisions, and** which **contains the description of the data, the management tools and any decisions chosen before and during the execution of a research project. Its preparation and review may fall to the principal investigator or another member of the research group responsible for this task**; for example, analysts or data engineers (if such roles exist in the team).

The *DMP* should not be considered merely a document that must be filled as per the requirements of the funding entity, but should be a **dynamic document that helps research staff to improve their research, designing how the data will be used, identifying possible problems that may arise and devising alternatives.** As with any planning, it may need to be modified throughout the process – for example, when the expected data cannot be obtained – so it should be reviewed periodically and again once the project is finished.

Data management plans **must be validated by the person who requested their drafting;** either the funding entity or the centre itself. From the point of view of a centre, its use may be established as an institutional policy, but personnel should still be designated to monitor it. The responsibility for this would ideally fall to the data steward.

The data management plan **should include:**

- **Basic administrative information**, including the project details (title, code and funding entity) and the principal investigator. Other data may be included, such as the details of other members of the research team and other members of the consortium, where appropriate.
- **A description of the data involved in the project**, including the formats and an estimate of the size. Data may be created during the research process, either through simulation, experimentation or observation, or existing data to which the research staff have access, may be used. In the latter case, the plan must specify how this data is accessed and what the authorised uses are.
- **A description of the storage system** that will be used during the project, as well as the **systems that will be used to analyse or process the data.** This should take into account who must have access to this area, be it internal or external personnel. If personal data will be used, the data protection officer must be notified and authorisation from the corresponding ethics committee must be obtained, respectively.
- **Access policy: who, how and when will be the data be accessed.** This access may be staggered over time, or be restricted, with certain terms of use established. The **data controller** will also be specified, in order to identify **who is able grant access to the information.**
- Finally, the plan must indicate where **the data will be stored and preserved** once the project is completed. A trusted repository must be found and the corresponding metadata must be included if the data is to be published.

It is common for research funding entities to propose their own data management plans, with templates that the beneficiaries must complete and send. For example, the *European Commission and the European Research Council (ERC)* propose the following outline:

- Summary of the data: description of the data to be used, format and size.
- *FAIR* data: explanation of the steps to be followed to ensure that the data complies with the *FAIR* principles.
- Resource allocation: identification of resources allocated to data management. These can be charged as project costs.
- Data security: description of procedures to control access to the data, especially relevant when working with sensitive or confidential data.
- Ethical aspects: identification of any ethical aspect related to the research, regulations or applicable laws.
- Other aspects: any other aspect that is considered worth mentioning.

The structure and content suggested by *“la Caixa” Banking Foundation* in its research calls is made up of five sections:

- Summary of the data
- Responsibilities
- *FAIR* data
- Data security
- Other aspects

It is worth highlighting in this model the section on "Responsibilities", where the person responsible for managing the data, granting access and the owner of any potential intellectual property rights must be indicated.

At the state level, the latest *R&D&I* Plans recommend the drafting of data management plans but do not require it, as is the case at European level, nor do they offer an official template that acts as a guide.

RECOMMENDED ACTIONS

- Provide support for research staff so that they can plan the data management of projects in accordance with the requirements of the main funding entities.
- Establish as a best practice of the centre itself the drafting of a data management plan aligned with the existing templates of funding entities, in order to ensure researchers are familiar with the main aspects that should be taken into account.
- Publish the data management plan model on the centre's intranet, along with a section on frequently asked questions regarding how the plan should be drafted.
- Organise periodic training sessions for researchers on how to complete the data management plan, using both the centre's own model and the templates offered by the main funding entities.
- Carry out systematic reviews of the data management plan, as well as a final review, once the project is completed, with personnel assigned to these tasks.

3.1.1 Tools for drafting a data management plan

The design of a *DMP* **does not require a specific application,** since it can be generated as a report using a word processing program. However, **there are online tools**, some of them collaborative, which enable the drafting of a plan, its sharing with other users and tracking of the different versions, in order to control the changes introduced during the research project. These applications usually take into account the requisites of funding entities through templates and, if they are installed on the centre's own servers, they also allow custom templates to be incorporated. Two of the most popular tools are described below:

DMP Online[7]
This is the best-known and most-used tool. Created by the *Digital Curation Centre (DCC)* in Scotland, it has been adapted and translated by numerous institutions, including the *Consorcio Madroño*[8] and the *Consorci de Serveis Universitaris de Catalunya (CSUC)*[9]. The *DCC* offers the source code to install the application locally and to manage the templates offered to users.

Additional relevant information

- *"H2020 templates: Data management plan v2.0 – 15.02.2018"*: https://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf
- "Modelo de plan de gestión de datos para proyectos de la *Fundación "La Caixa""*: https://fundacionlacaixa.org/documents/10280/1198289/modelo_de_plan_de_gestio n_la_caixa_es.pdf/685ac764-bc0b-9511-8e88-2768919637a8?t=1593414644950

[7] https://dmponline.dcc.ac.uk/

[8] https://pgd.consorciomadrono.es/

[9] https://dmp.csuc.cat/

ARGOS[10]

Offered by the *OpenAIRE* consortium, it is similar to the aforementioned tool except that it is *machine readable*. This means that the contents of the plan can be standardised and made more interoperable, and relevant information can be extracted more easily.

RECOMMENDED ACTIONS

- Provide research staff with support tools for drafting data management plans, as well as practical training on how to use them.
- Provide research staff with access to a repository, with real examples of *DMPs* and templates adapted to the characteristics of the centre.

3.1.2 Data management protocols

Data management plans are associated with specific research projects. However, if their use is to be adopted across the centre, a data management protocol must be developed. This, in turn, can form part of the institution's general policy or strategy.

A data management protocol **should include:**

- The applicable laws and regulations.
- The responsibilities regarding data management and access.
- The infrastructure and services available to the research staff throughout the research data cycle.
- A list of preferred formats that includes the most commonly used by the research communities of the members of the centre.
- Information on personal data management and the management of intellectual property.
- The data preservation and publication options offered by the centre.

Although drawing up a protocol **falls under the remit of a centre's data manager, it is recommended that a working group be set up which includes all the profiles that take part in the different activities within this process,** either directly, such as the researchers, data scientists, etc. or indirectly, such as personal data protection officers, system managers, and others.

RECOMMENDED ACTIONS

- Prepare and publish a data management protocol on the centre's intranet, aligned with the centre's general policy or strategy. In this way all staff will be aware of the guidelines that must be followed during the activities of this process.

Additional relevant information

Proposal for discipline-specific data management protocols from *Science Europe:*
https://www.scienceeurope.org/media/nsxdyvqn/se_guidance_document_rdmps.pdf

[10] https://argos.openaire.eu/splash/

3.2 Aspects related to personal data

According to the ***General Data Protection Regulation*** (*GDPR*)[11] **, personal data is any data that allows the identification of a person,** including: names, dates, physical characteristics, personal images, physical or electronic mail addresses, *IP (Internet Protocol)* addresses of devices, identification numbers associated with a database through which data that identifies a person can be found, etc.

This personal data also includes **"specially protected"** or sensitive data, such as data **related to the ethnic origin, beliefs, political opinions, sexual orientation or trade union membership of a person**, for which secure, adequate infrastructures are required for processing and storing said information. There are also **"special categories"** of data, which refer to **genetic information, biometric or health data.** These must also be processed with care, so as not to put the rights of individuals at risk (this is blocked data which must be stored in a secure place and which must be carefully monitored for any security breaches).

Research projects often involve the use of personal data. It is very important that the centre establishes the procedures that must be followed in these cases in order to comply with the regulations and with its own code of conduct. Moreover, the research staff must be provided with all necessary information for its correct use.

If the project involves the processing of personal data, the data protection officer must be involved from the start of the planning phase. This figure is responsible for supervising the planned processes **in order to protect the fundamental rights of individuals. They can provide advice to mitigate or eliminate any potential impact on privacy.** It is a different profile from that of a manager, since the person is conferred a legal status and is an expert in the management of personal data in the field of research, but also in other fields such as quality surveys, associate surveys, etc.

Another consideration is **whether the requested data is essential for the research** and whether participants have been adequately informed about the data that will be used and how it will be processed.

The centre must also have a record of all the different processes that it carries out using personal data and, if there is a **plan to publish the data at some point, it must be previously anonymised.** If a fundamental right may be put at risk, Privacy Impact Assessment (PIA) should be carried out, drawn up by the data protection officer in collaboration with the research staff.

Finally, another aspect to take into account is the **ethical validation that a research project** may require when it is carried out using individuals. In this case, even if personal data is not processed directly, validation will be required (in the case of surveys, recordings, etc. of people). This type of validation **is carried out by internal and/ or external ethics committees,** in response to the institutional codes of conduct themselves, at the request of funding entities, or by legal imperative, as occurs, for example, in the field of biomedical research[12] or animal testing. The legislation establishes the requirement of assessment by a recognised ethics committee, but does not establish any formal requirements. These ethics committees can provide support and examples when drafting

Additional relevant information

[11] https://gdpr.eu/

[12] https://www.boe.es/eli/es/l/2007/07/03/14

informed consent documents. Again, it is important that the centre's research staff be aware of the procedures established for this purpose.

RECOMMENDED ACTIONS

- Put in place procedures and personnel (data protection officer and ethical committees) to minimise impacts and risks when processing personal and sensitive data. The research staff must follow, at all times, the instructions of the centres in this regard.

3.3 Management of intellectual property

Intellectual property refers to the rights over any creation, such as artistic and scientific works, as well as databases, **which correspond to the authors,** as a result of their intelligence and creativity, **and to other holders.** The term intellectual property **is often extended to the concept of industrial property,** which includes industrial designs, patents and utility models, trademarks and trade names, as well as semiconductor topographies[13]. **In Spain these two concepts are regulated by different laws that grant** different conditions and duration to the holders of the rights.

The management of intellectual property must be taken into account right from the planning stage, identifying what rights[14] are generated (exploitation, recognition, etc.), **who the holder is**, whether the **data generated can be reused** and under what **conditions**.

In many cases the individual data cannot be protected in this way as it is not an original creation that can be considered someone's

Additional relevant information

• Data protection impact assessments, Spanish Data Protection Agency (AEPD): https://www.aepd.es/es/derechos-y-deberes/cumple-tus-deberes/medidas-de- cumplimiento/evaluaciones-de-impacto

• Practical guide for data protection impact assessments subject to the GDPR, Spanish Data Protection Agency (AEPD): https://www.aepd.es/sites/default/files/2019-09/guia-evaluaciones-de-impacto- rgpd.pdf

• "Ethics and data protection". European Comission, 14 November 2018: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h 2020_hi_ethics-data-protection_en.pdf

•"Consentimiento informado del participante (modelo orientativo)". *Bioethics Committee of the University of Barcelona:* http://www.ub.edu/comissiobioetica/sites/default/files/documents/pagina-basica/model_orientatiu_consentiment_es.pdf

• "El consentimiento informado en la investigación en enfermedades raras". Teresa Pàmpols, Carmen Ayuso and Guillem Pintos: https://www.ciberisciii.es/media/602614/14-cap-9.pdf

• Network of Ethics Committees of Universities and Public Research Bodies (RCE): https://redcomitesetica.es/

[13] https://www.oepm.es/es/propiedad_industrial/propiedad_industrial/

[14] See Chapter III, articles 14 to 25, of Royal Legislative Decree 1/1996, of 12 April, approving the consolidated text of the Law on Intellectual Property, regulating, clarifying and standardising the current legal provisions related to this matter.

own work. This is the case, for example, of values arising from an experiment, the temperature data observed or dates. With regard to **compilations or databases, the intellectual property law** in force[15] in Spain, as well as most of the laws in force in Europe[16], include the **recognition of the "*sui generis*"** right, which grants the exclusive right to extract and reuse the contents to the creator of a database. Therefore, that right holder can grant the necessary permissions to reuse the data contained in a database regardless of whether the individual bits of data are also protected by intellectual property. This aspect is important for centres, both because they generate new data and because of the reuse of existing data in their research. It is also necessary to take into account whether the research being carried out will lead to results that are the subject of industrial property in which case the necessary mechanisms must be established so as not to disclose any information. In this case, any data that may contribute to generating a patent must remain closed for a reasonable period of time to ensure there is no risk to the patent being awarded.

RECOMMENDED ACTIONS

- **Provide support on intellectual property matters for research staff, providing guidance on the centre's policy and providing access to specialist personnel to analyse situations and make decisions.**

- **Establish a clear policy regarding ownership of the rights and income that the creation of data could generate, as well as its compilation, and disseminate said policy among the research staff, so that they know the conditions to which they are subject (if they are permitted to share the data at their discretion, etc.).**

- **Clearly define who is the rights holder in each situation (the research staff, the centre or its parent entity, the consortium or the funding entity), as this is essential for determining what licences to give and what data uses to allow.**

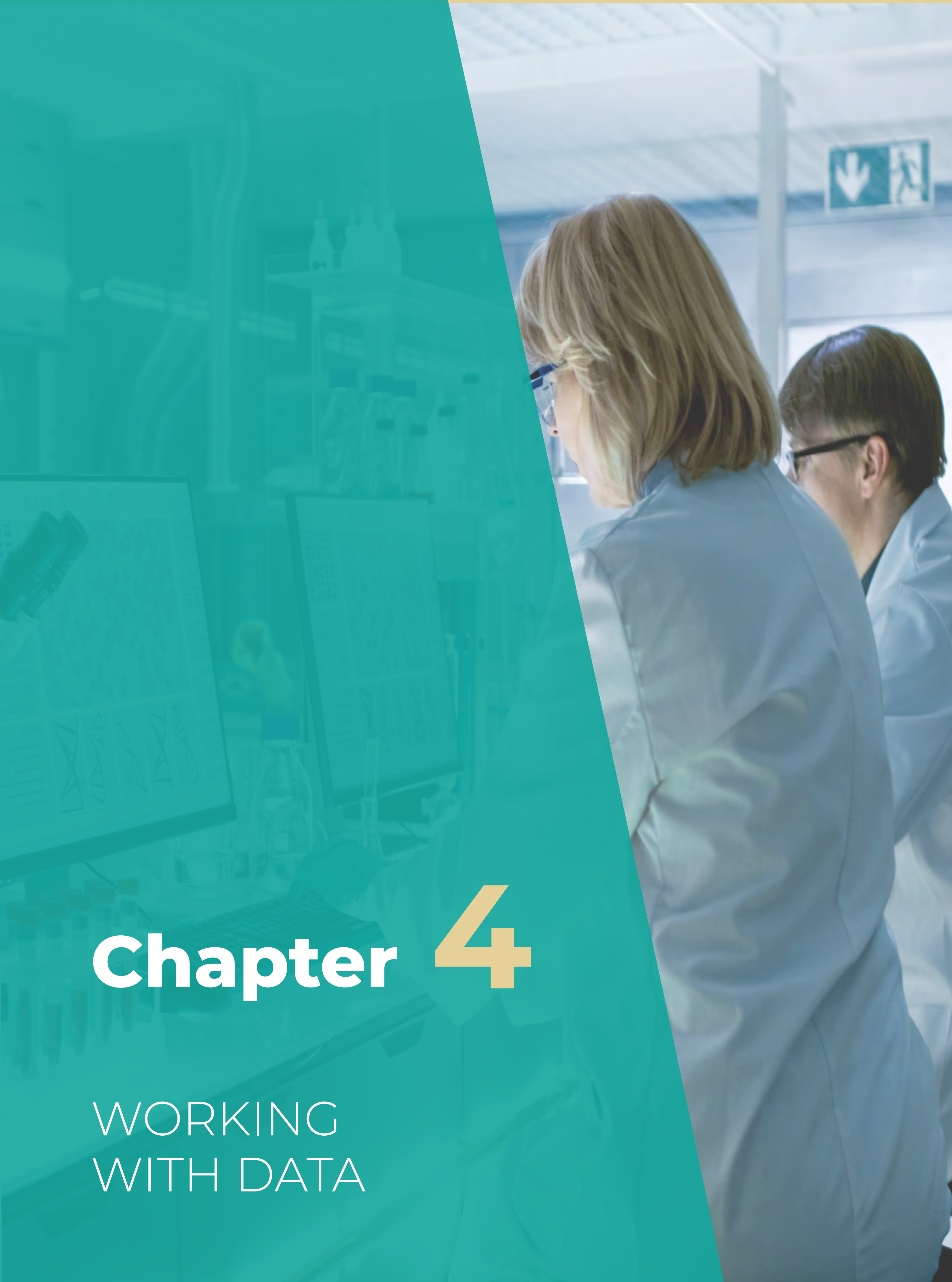
Additional relevant information

It is common for research institutions to have clear policies, standards and procedures to deal with industrial property, but they do not always exist in the case of intellectual property and, where there do exist, they often only include certain types of works, such as software and databases. In most cases, the ownership of the exploitation rights rests with the institution and any income derived from its commercial exploitation will be divided between the researchers and the institution to which they belong.

If the centre or its research staff participates in a project funded by an external organisation, they should check whether there are any requirements regarding ownership and always try to find a balance between these requirements and the centre's standards. Any occasional contract to bring in research staff to collaborate in a project that generates data and that is capable of producing rights must include a section referring to the centre's regulations, since a researcher could otherwise collect data and then use it in other projects without the knowledge of the centre.

[15] Royal Legislative Decree 1/1996, of 12 April, approving the consolidated text of the Law on Intellectual Property, regulating, clarifying and standardising the current legal provisions related to this matter.

[16] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.



Chapter 4

WORKING WITH DATA

4

WORKING WITH DATA

When designing a research project the type of data that is needed is first identified and the most appropriate source is determined, either reusing existing data or generating new data. After deciding this, the next thing to consider **is the formats that are going to be used, how the data is going to be organised, where and how it is going to be stored and published.**

4.1 Reuse of existing data

Research staff sometimes start a project using existing data, which **may be freely accessible or require specific access permission.** This practice is common in certain disciplines, for example in the social sciences when statistical data is used or when conducting epidemiological studies based on health data.

The first step **when accessing external data** is, therefore, to analyse **its terms of access, use and publication.** It is important to be aware of the difference between public and open data, terms which are often confused. **Data is public if it can be accessed for consultation without any type of registration, for example, when it is published on a data portal.** However, **for the data to be reused, a licence or a notice** indicating **what uses are allowed, and under what conditions, is required.** The data can then be considered open data. **In the absence of a licence or notice the data must be considered to have *"all rights reserved."*** Meanwhile, although the individual data might not be protected by intellectual property rights, the database that contains the data could be, as we have already seen. The extraction and reuse of all, or a substantial part, of the contents of a database requires a specific authorisation, unless it has already been previously granted through its licence of use.

The “open data” portals offered by public administrations are a clear example of open data. This data is usually published under licences that sufficiently allow their reuse, such as the *CC BY*[17] and *CC0*[18] licenses, which will be looked at in the next chapter. This is the case of the *National Statistics Institute (INE),* which allows data to be consulted without any type of registration (public data) and also has an open data section[19] in which the terms for reuse are specified. The *Xunta de Galicia*[20] and the *Spanish Government itself*[21] also have their respective open data portals.

In other cases, access to the data is more restricted and the reuse and dissemination

Additional relevant information

[17] https://creativecommons.org/licenses/by/4.0/
[18] https://creativecommons.org/share-your-work/public-domain/cc0/
[19] https://www.ine.es/datosabiertos
[20] https://abertos.xunta.gal/
[21] https://datos.gob.es/

of such data, initially, is not allowed or has limitations. This is the case of *Eurostat*microdata, which are only accessible for scientific purposes and require registration[22]. In any case, the same institution may be offering data simultaneously under different terms of use, as is the case of the INE, which specifies on its website that the information contained comes from multiple sources, authorising the reuse of information where the original source is the institution itself.

When data is generated with a view to its future reuse or when existing data is reused, one aspect to take into account is its citation in research sources. At the initiative of *FORCE11*, a community of academics, librarians, editors and research funders, some principles[23] have been proposed for the correct citation of data, which include best practices for the recognition of authorship, access to data and its verification. This community also claims that the data should have the same consideration as any research findings. Here are examples of citations that follow these recommendations:

Examples

- King, G.; Zeng, L. 2007. "Replication data for: When Can History be Our Guide? The Pitfalls of Counterfactual Inference". https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EK886K

- Barnett, C.L.; Wells, C.; Thacker, S.; Guyatt, H.J.; Fletcher, J.M.; Lawlor, A.J.; Winfield, I.J.; Beresford, N.A. (2015). "Elemental concentrations in fish from lakes in Northwest England". NERC Environmental Information Data Centre. (Dataset). https://doi.org/10.5285/ed90df1b-462c-46bb-afbd-59794fb03f6b

RECOMMENDED ACTIONS

- **Analyse how the data can be reused and if it can be published later (what uses are allowed and under what conditions).**

- **Include the following main elements in the data citation: authors, dataset title, repository or file, year, version, and persistent identifier.**

Additional relevant information

[22] https://ec.europa.eu/eurostat/web/microdata/overview

[23] https://www.force11.org/datacitationprinciples

4.2 Format and organisation

If the data is to be considered *FAIR*, the format used must be carefully chosen. Sometimes the format will be dependent on the device with which the research is being carried out, but on other occasions the most convenient format can be chosen. The **use of standard formats which are not dependent on proprietary programs or specialised software**, is always recommended, where possible.

The choice of format is therefore **a crucial aspect when it comes to sharing data and promoting its reuse,** since it may require later transformation if the format does not meet the storage requirements of the repository where it is to be archived. If the data can only be saved in a proprietary format, the program and the specific version used to generate this data should be specified.

A closed list of formats cannot be established at the outset, as each discipline may use different formats, established and consolidated by the community. However, a list can be drawn up of recommendations or preferred formats, which may depend on the preservation policy. For example, *Data Archiving and Networked Services (DANS),* the Dutch national centre for research data, establishes preferred formats for different types of data[24]. The following table lists some of the data formats recommended by DANS.

TYPE	PREFERRED	NON-PREFERRED
TEXT	PDF/A, ODT	DOC, DOCX, PDF (not PDF/A)
BRAND LANGUAGE	XML, HTML	SGML, MD
SPREADSHEET	ODS, CSV	XLS, XLSX, PDF/A
STATISTICAL DATA	DAR, SPS, DO, R	POR, SAV, DTA
VECTOR IMAGES	SVG	AI, EPS, WMF

Source: DANS

Additional relevant information

[24] https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/before-depositing/file-formats

Furthermore, there are formats that may be recommended for a certain type of data, but which can make it difficult to reuse (for example, a table in PDF) format. For this reason, **when selecting a format, both the time required and the problems that its use may entail should also be considered.**

Data organisation, meanwhile, refers to the way it is stored in the workspace, and it must be structured and consistent. This aspect is important when the data is to be shared or when it is necessary to facilitate access to common spaces or shared folders to people from the same or different organisations.

Moreover, file names should be self-explanatory, in order to facilitate the organisation of data. In the case of **dynamic data**, version control should be implemented and the current version should always be indicated. This indication can be maintained when the data is published in the corresponding repositories.

RECOMMENDED ACTIONS

- **Select and inform staff about the centre's recommended formats, as well as its methodology for organising the data, file naming and assigning versions.**

4.3 Storage and access

When working with data, it is very important **to determine the types of devices that will be used for its storage during the execution of the project.** A wrong choice could lead to data loss, with disastrous consequences for the research.

Research centres should offer researchers a workspace where they can store data with access control, which allows setting permissions and therefore **configuring access profiles** for working with data that is not yet suitable for dissemination. Shared access should be considered for both internal and external users, since it is common for research projects to work in collaboration with other entities.

A backup process should also be established to ensure that data is not lost, even in exceptional circumstances.

Likewise, research staff must know how to properly handle data, in order to avoid any loss of information. Be aware that devices such as *USB* sticks or personal computers are quick and easy-to-use tools for storing data, but not the most suitable for information security purposes.

Additional relevant information

For more information on data organisation:

• *UK Data Service:* https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/organising/

• *Research Data Netherlands (RDNL):* https://datasupport.researchdata.nl/en/start-the-course/iii-research-phase/organising-data

Meanwhile, the storage spaces used during the research project can be connected to the repositories where the data will be deposited once the project is finished, which can in turn be used to publish and preserve the data.

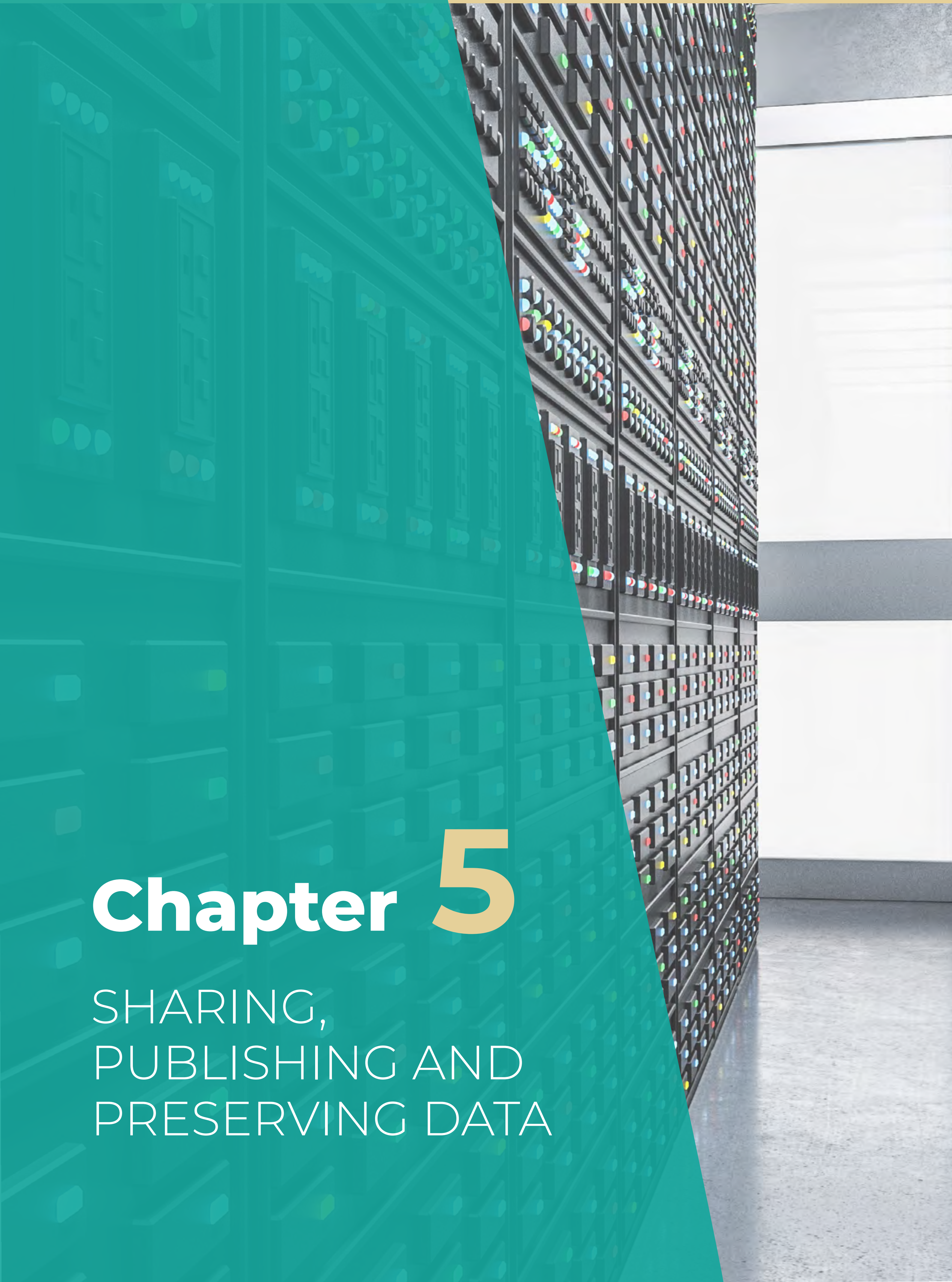
The *University of Utrecht* provides an example of this, offering researchers free and paid solutions for storing data. To help them choose, they have developed a dynamic questionnaire[25] that rules out options as the user goes through it, finally arriving at the most appropriate solution.

RECOMMENDED ACTIONS

- Decide which devices should be used to work with the data.
- Set up a workspace for storing the data with access control, which allows permissions to be established and define user roles for working with the data.
- Define a data backup system.
- Define action guidelines for data management. Inform, train and educate the centre's staff on the importance of monitoring.
- Select possible alternative solutions for data storage and provide the research staff with information on how to select the most appropriate solution in each case.

Additional relevant information

25 https://tools.uu.nl/storagefinder/



Chapter 5

SHARING, PUBLISHING AND PRESERVING DATA

5

SHARING, PUBLISHING AND PRESERVING DATA

This chapter looks at the **final phase of the data management cycle: how to share, publish and preserve data,** addressing in detail the choice and role of the different types of existing repositories.

The following diagram shows the process which primary data goes through up until its publication in an open access or restricted access repository:

- The process starts with **primary data** or with the reuse of existing data.
- After this data is obtained, **secondary data** is generated through the treatment or processing of the primary data.
- The secondary data may be **discarded, archived, or published** in a repository.
- This could be a **subject, institutional or consortium repository.**
- **The metadata** of the data published in the repository must always be open to everyone so that the information can be identified. If the data is offered in an open format, it must be accompanied by a **licence of use.** However, **when access to the data is restricted,** it should **be accompanied with information on how it can be used,** in the event it follows the *FAIR* principles. One option is to provide a standard document *(Data User Agreement, DUA)* that the user must sign and send to the owner of the data in order to use it. In any case, publication in a repository should entail the assignment of a persistent identifier (*PID*) such as a DOI.

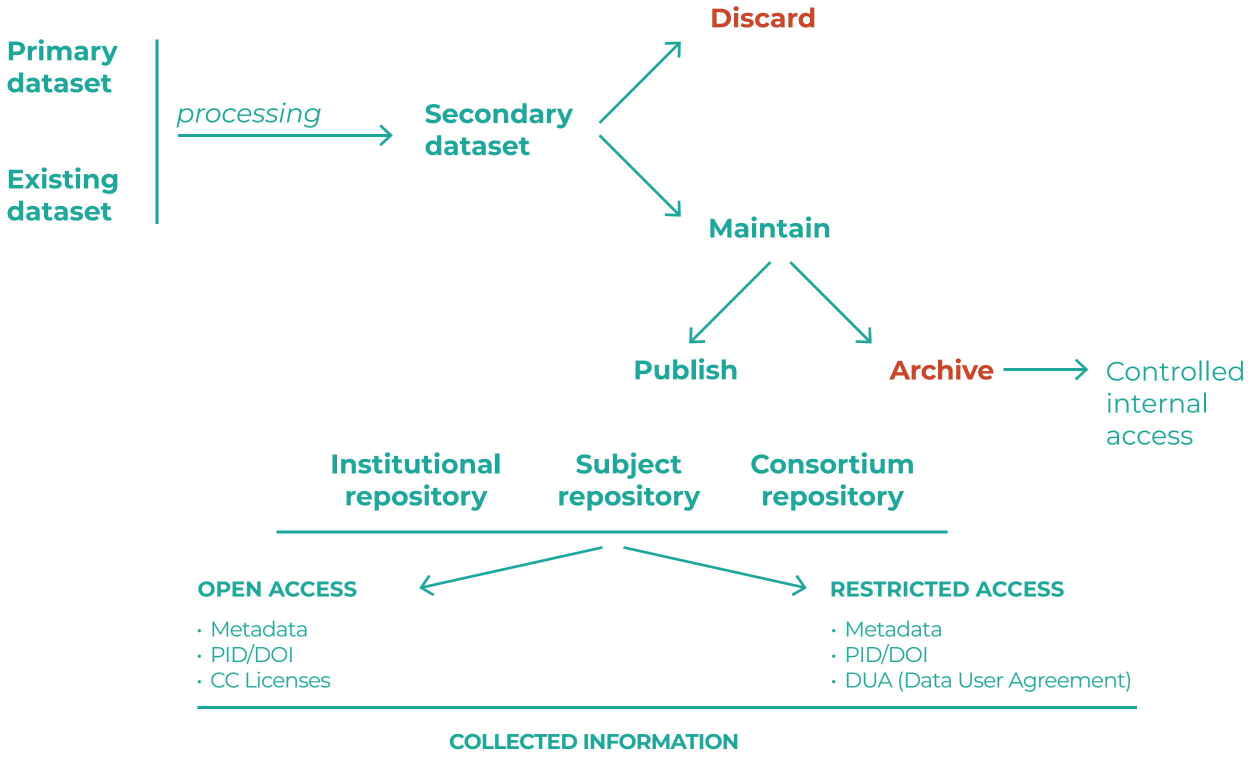


Figure 4: Process which primary data goes through in an open access or restricted access repository until its publication.

Source: compiled by author.

As mentioned throughout this guide, the *FAIR* principles offer guidelines on how to properly manage research data. **If data is to be shared publicly, a repository that also follows the *FAIR*** principles should be chosen. This should offer the following characteristics, among others:

- Use a unique identifier.
- Employ a standard metadata schema.
- Allow the use of data licences.
- Use a preservation system.

The **application of the reuse principle** established by the FAIR principles **requires open and closed data to be processed differently.** As a result, the first few sections of this chapter focus on this aspect. First, **the different licences that can be used to allow said reuse of data are described.** It then goes on to explain **how closed data can be processed so that it can be shared in an aggregate or anonymous manner and how to inform users of a restricted reuse subject to conditions.**

5.1 Sharing data

5.1.1 Open data

As explained previously, **open data is data that does not require authorisation to be accessed, examined, analysed and reused.** Data can also be considered open when the **access conditions are minimal** (for example, after prior registration or the acceptance of very lax terms of use). **The type of control or authorisation required to access the data** determines whether it can be considered "open". This concept **does not necessarily offer guarantees about the quality or scientific value of the data,** or about its possible future usefulness.

If the data is to be shared openly, that is, **allowing its reuse without users having to ask for permission, some type of licence is required.** The most popular open content licences are probably those offered by the *Creative Commons*[26] organisation. Since 2002, this organisation offers a set of licences that are used to authorise the reuse of data with different options. These licences were not created specifically for data, but from version 4 onwards, they contain a clause that indicates how they can be applied to database rights, thus:

- If the licence limits commercial exploitation, the contents of a database can be reused to create another database that cannot be commercialised.
- If derivative works are limited, databases cannot be created with the elements from the database subject to the licence. A derivative work is understood to be one that contains all or a significant part of the elements of an existing database.
- If the *copyleft* condition is included, in other words, the licence is maintained in any possible derivative works, the original licence must be maintained in any database created.

The following table lists the standard Creative Commons (CC) licenses.

Additional relevant information
To learn more about the technical characteristics of the DOIs/PIDs generation service, see this article: https://www.biblogtecarios.es/rafaelibanez/pid-identificadores-persistentes/

[26] https://creativecommons.org/licenses/

cc BY	Attribution	This allows the copying, distribution and public communication of the original work, as well as the creation of derivative works, without any restriction so long as attribution is given to the creator and the licence notice is maintained.
cc BY SA	Attribution-ShareAlike	This allows the copying, distribution and public communication of the original work, as well as the creation of derivative works, without any restriction so long as attribution is given to the creator and the licence notice is maintained. In the case of derivative works, these must be disseminated with the same licence.
cc BY ND	Attribution-NoDerivatives	This allows the copying, distribution and public communication of the original work without any restriction so long as attribution is given to the creator and the licence notice is maintained. No derivatives of the work are permitted.
cc BY NC	Attribution-NonCommercial	This allows the copying, distribution and public communication of the original work, as well as the creation of derivative works, as long as it is for non-commercial purposes only. Attribution must be given to the creator and the licence notice maintained.
cc BY NC SA	Attribution-NonCommercial ShareAlike	This allows the copying, distribution and public communication of the original work, as well as the creation of derivative works, as long as it is for non-commercial purposes only. Attribution must be given to the creator and the licence notice maintained. In the case of derivative works, these must be disseminated with the same licence.
cc BY NC SA	Attribution-NonCommercial NoDerivatives	This allows the copying, distribution and public communication of the original work as long as it is for non-commercial purposes only. Attribution must be given to the creator and the licence notice maintained. No derivatives of the work are permitted.

Source: compiled by author.

Another legal instrument offered by *Creative Commons* is *CC0*[27]. This tool is usually presented as a waiver of intellectual property rights, but it is actually a bit more complex. A more detailed look at the legal code reveals that there are three levels of action:

1. The first is the waiver: The holder of any rights waives them all, but since Spanish legislation (among others) establishes certain inalienable and non-transferable rights, the waiver applies to all those rights that the owner is able to waive.
2. This is where the second level comes in: The holder grants the user what could be called an unconditional licence over all those rights that the former is unable to waive.
3. Finally, the third level comes into play: the holder agrees not to pursue the infringement. Therefore, the holder of the rights will not take any legal action if authorship is not recognised, even if it is an inalienable moral right. This shifts the focus more towards the rules of the research community than on strict compliance with intellectual property law.

Another option for licensing data is to use the set of licences offered by the *Open Data Commons (ODC)*[28] *initiative*. This initiative promoted by the *Open Knowledge Foundation* offers three licences:

1. ***Open Data Commons Open Database License (ODbL):*** includes the *copyleft* requirement, equivalent to what is currently required under the *CC BY-SA license*.
2. ***Open Data Commons Attribution License:*** only requires an acknowledgement, like the *CC BY license*.
3. ***Open Data Commons Public Domain Dedication and License (PDDL):*** transfer to the public domain equivalent to the *CC0* option from *Creative Commons*.

The following table shows the ODC licenses with their Creative Commons equivalent.

ODC License	ODC ODbL	ODC Attribution License	ODC PDDL
Equivalence in CC	CC BY-SA	CC BY	CC0

Source: compiled by author.

Finally, it is worth mentioning that some countries have chosen to generate their own national licences, as is the case of *France*[29] and the *United Kingdom*[30].

Additional relevant information

[27] https://creativecommons.org/choose/zero/
[28] https://opendatacommons.org/
[29] https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf
[30] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

RECOMMENDED ACTIONS

- Analyse the ownership and licences associated with all the data used in a research project.
- Clearly define who is the holder of all the rights (a third party – reuse – the researcher, the centre or the funding entity), since it will be the holder who can authorise the use of this data.

5.1.2 Closed data

Although there is a trend to promote the opening of research data, **there are also many situations in which data must be kept closed**, for example, **when working with personal or confidential data or under security regulations.** It is important to stress that the *FAIR* principles do not require that the data be open, but that users be informed about how it can be reused.

As already mentioned, a ***Data Use Agreement (DUA)*** can be established, which **indicates the permitted uses and under what conditions,** taking into account any applicable legal aspects and legislation, such as that referring to the use of personal data (European General Data Protection Regulation[31] and national legislation[32]). The agreement would need to be signed by both parties before access to this non-public data is granted.

Other options for sharing data that cannot be made public include **anonymisation**, **aggregation** and the use of techniques such as **differential privacy**. These must be implemented correctly, so as not to jeopardise the confidentiality, security or privacy of the data and the subjects involved.

- **The anonymisation of personal data** is the process through which **the connection between personal data and the natural person to whom it refers is broken,** thus preventing their identification. The resulting data is said to be "anonymised" and does not fall within the cases protected by the *General Data Protection Regulation (GDPR)* or other applicable regulations. When the element (software, table of values, etc.) that allows the re-identification of the person to whom the data refers, is simply isolated and detached from the data, and not destroyed, the data is said to be "pseudonymised", and the data is still considered to be personal data. The *Spanish Data Protection Agency* provides a set of guidelines in this regard to ensure that personal data anonymisation procedures are followed correctly[33].

Additional relevant information

[31] General Data Protection Regulation (Regulation EU 2016/679) https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

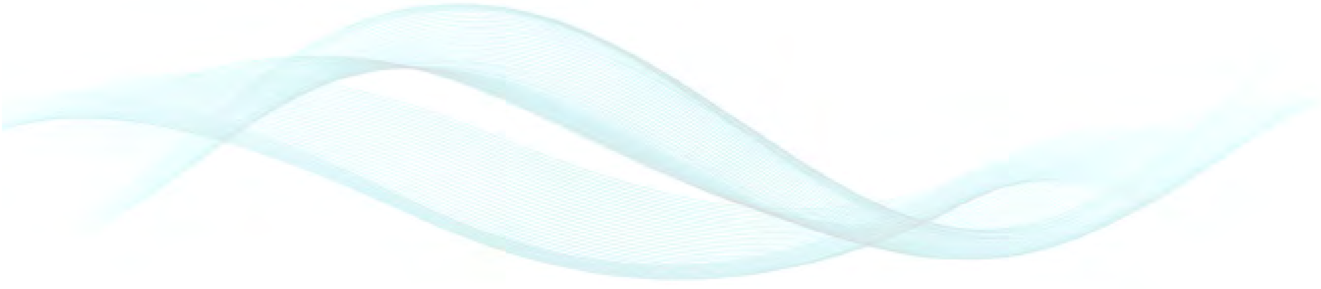
[32] In Spain, Organic Law 3/2018, of 5 December, on the Protection of Personal Data and guarantee of digital rights https://www.boe.es/eli/es/lo/2018/12/05/3

[33] https://datos.gob.es/es/documentacion/orientaciones-y-garantias-en-los-procedimientos-de-anonimizacion-de-datos-personales

- **Aggregation consists of compiling data, and eliminating the individual information from each item of data.** An example might be obtaining information about how many people have made a certain journey, without revealing which individuals have done so.
- **Differential privacy[34],** meanwhile, consists of a set of systems and practices that **allow conclusions to be drawn about a group, while maintaining the privacy of the data of each of its members at all times.** An example would be to introduce errors or noise into census data without altering the overall data[35].

RECOMMENDED ACTIONS

- **Draft restricted data use agreement *(DUA).***
- **Publish the metadata corresponding to the closed data, as well as allow restricted access to it.**
- **Set up the licences that can be used in the centre to share data.**
- **Offer a data anonymisation service.**



Additional relevant information

34 https://en.wikipedia.org/wiki/Differential_privacy

35 https://www.technologyreview.es/s/11974/tr10-privacidad-diferencial

"Guidance and guarantees in the process of personal data anonymisation" by the *Spanish Data Protection Agency:* https://datos.gob.es/en/documentacion/guidance-and-guarantees-process-personal-data-anonymisation.

5.2 Repositories

There are many **different repository options:**

- The first option is **to select a repository dedicated to a specific discipline or field of knowledge,** if one exists and is well-established.
- The second option involves choosing **an institutional repository or a consortium repository.**
- Finally, there is also the possibility of opting for a **multidisciplinary repository.**

The volume of files, format and preservation conditions can also influence the choice of the most suitable repository. Each of these options is detailed below.

5.2.1 Subject repository

Some disciplines offer repositories dedicated to sharing data. Some well-known examples include the *GenBank*[36], repository of *DNA* sequences, or the *GBIF*[37] repository of biodiversity data, but other repositories are not as well known even to researchers themselves. To aid in the identification of the right repository, the *re3data*[38] registry offers a list of research data repositories and can be searched using different parameters such as scientific scope, type of access and licences of use, among others. Another alternative would be to use the *FAIRsharing*[39], portal, which allows various filters to be applied in order to locate the most suitable repository.

The first option should always be to deposit the data in a subject repository that is well-established among the scientific community of a certain field or discipline. The research centre should analyse the alternatives and provide researchers with a list of repositories, depending on the areas in which they work.

Depositing the data in subject repositories **does not mean that the centre breaks its link with the data. It could create a data catalogue with the corresponding description and a link,** in the same way that other research results such as publications are compiled. This catalogue would contain both the data deposited in external repositories and in internal repositories.

Additional relevant information
Recommendations for selecting a repository for storing research data: Version 5, October 2020": http://hdl.handle.net/2072/377586
[36] https://www.ncbi.nlm.nih.gov/genbank/ 37 https://www.gbif.org/
[37] https://www.gbif.org/
[38] https://www.re3data.org/
[39] https://fairsharing.org/databases/

5.2.2 Institutional repository

Another alternative is to use the **centre's own repository or a repository of a related institution,** such as a university. Another option in this case is the creation of a **consortium-based institutional repository**, in which different institutions participate. **The main advantage here lies in the optimisation of resources and services.**

In any case, **the centre must designate a person to coordinate the depositing of data in the repository.** Furthermore, if the centre chooses to create its own repository, **it will need to ensure there is technical support in place and perhaps even assign a person to maintain and update the infrastructure.**

Common working guidelines are normally established in consortium repositories, but each centre or institution must be responsible for its space and for contacting its researchers to deposit the data, though individual technical maintenance is not necessary.

Many institutions that already have a consolidated institutional repository, initially destined for documents, have chosen to also use it as a data repository. They have created specific collections for data and even for other types of results such as software. This is the case of *Digital CSIC*, the repository of the *Spanish National Research Council (CSIC)*, which includes a specific collection of data sets for its centres in each of its sections[40].

Other institutions have chosen to create a repository dedicated exclusively to data. This is the case of the *University of Edinburgh* through its *Edinburgh DataShare*[41] repository and the *University of Amsterdam*[42]. The first example is a repository which is installed in the university itself, while in the case of the latter, the infrastructure has been outsourced to an external provider.

There are also multiple examples of consortium repositories, such as the Madrid-based *e-cienciaDatos*[43], which groups together the universities that are part of the *Madroño Consortium*, or the national repositories of the *Netherlands -DataverseNL*[44]- and *Norway -DataverseNO*[45]-.

When opting for an institutional repository, whether independent or part of a consortium, a preservation policy must be defined. Things to consider include a distributed backup strategy should be considered and also the period of time during which the data will be preserved. In many cases, this period is set at 10 years, although there may be an obligation to keep the data for longer periods of time.

5.2.3 External multidisciplinary repository

If there is no possibility of using a subject repository or an institutional repository, the data can be deposited in an external multidisciplinary repository. **This alternative,**

Additional relevant information

40 https://digital.csic.es/handle/10261/166574 41 https://datashare.ed.ac.uk/

41 https://datashare.ed.ac.uk/

42 https://uvaauas.figshare.com/uva

43 https://edatos.consorciomadrono.es/

44 https://dataverse.nl/

45 https://dataverse.no/

sometimes used individually by research staff**, is often an attractive option because it is initially free, but the available space and the size of the data sets are generally limited.** These restrictions can be eliminated **by using a paid service**, although this decision requires an overall assessment of both the needs of the centre's researchers and the benefits offered by the repository and other alternatives.

Some repository providers offer global solutions to institutions to become a stable resource and host the data collection as an institutional repository. This is the case of the *University of Amsterdam*, mentioned previously.

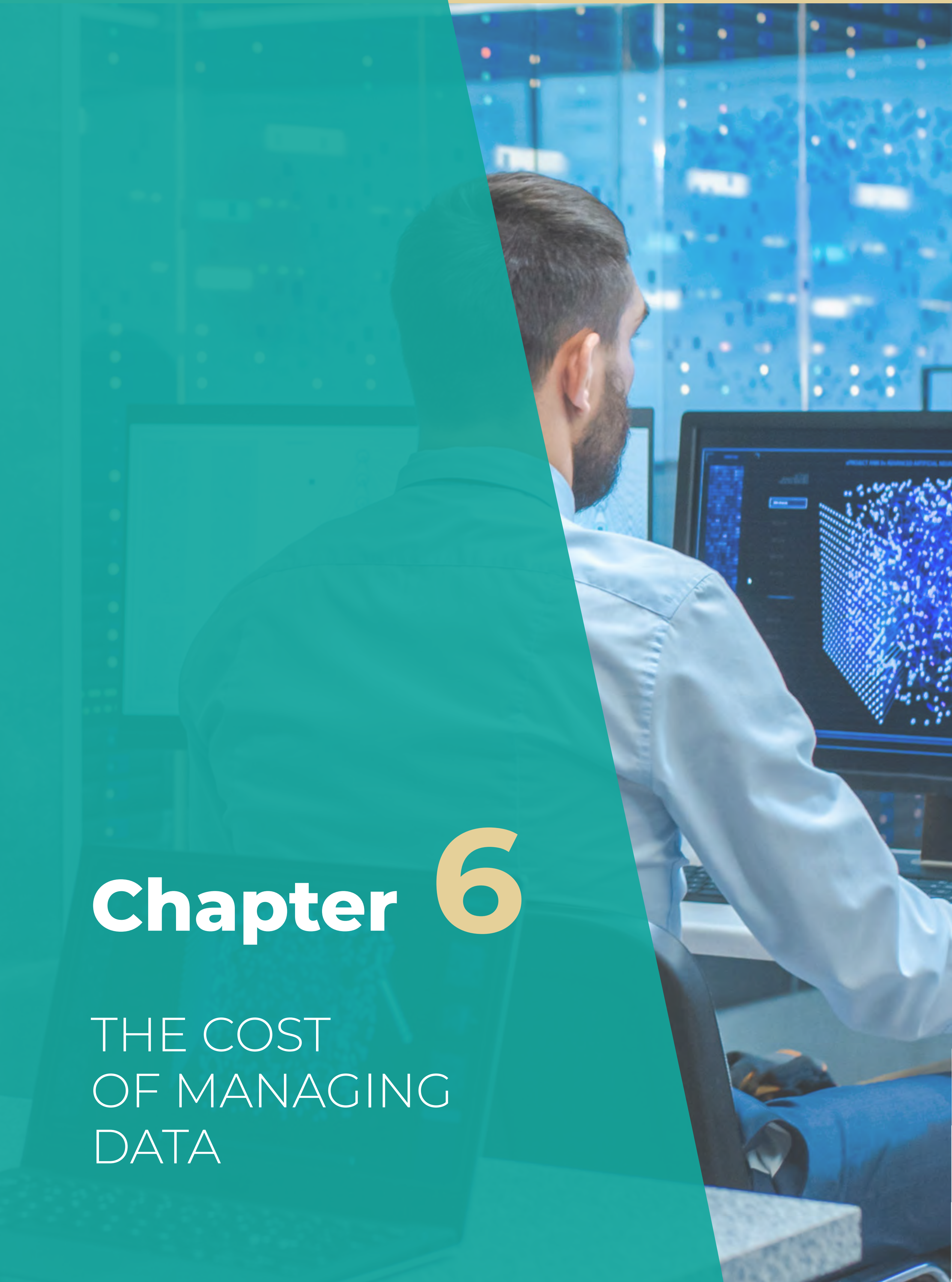
The research support working group of the *Consortium of University Services of Catalonia (CSUC)* has a list with the different characteristics of these multidisciplinary repositories. The latest version dates from October 2020[46]. The following table provides a sample of some of these repositories and their characteristics.

Repository	Figshare	Mendeley Data	Zenodo
Provider	Digital Science	Elsevier	CERN
URL	https://figshare.com	https://data.mendeley.com/	https://zenodo.org/
Size of files	Up to 5GB per file. Up to 20 GB of private space	Up to 10GB for a data set. Expandable if the institution subscribes to Mendeley Data	Up to 50GB for a data set. Can be contacted for larger sizes
Licence of use	By default. CCO for data and CC BY for other types of material	Allow the use of different data licences: CC0, CC BY, CC BY-NC	Allows a variety of licences, although they recommend open licences
Persistent identifier for the data	DOI	DOI	DOI
Cost	Free. But there is an option to pay for more space	Free	Free, but they do not rule out having to pay in the future if a certain limit is exceeded
Versioned	Allows different versions of a file already published	Allows the comparison of different versions of a file	Allows different versions of a file

Source: Consortium of University Services of Catalonia (CSUC), extracted from http://hdl.handle.net/2072/377586.

Additional relevant information

[46] http://hdl.handle.net/2072/377586



Chapter 6

THE COST OF MANAGING DATA

6

THE COST OF MANAGING DATA

When estimating the costs of managing research data, **all the processes involved must be taken into account.** This includes the **hours employed by the research staff and by the expert staff** dedicated to guiding and supervising the work, but also the **costs of the associated infrastructures and services.**

The planning process should include an estimate of **the hours dedicated to preparing the data management plan** by the research staff and expert advisory staff. The first plans will always be more time consuming, but implementing a protocol that follows pre-established practices will help expedite the process.

The estimated cost of the data collection process varies depending on different factors. In some projects this may take into account the **cost of access to external databases**, while in others the **cost of generating the data will have to be estimated,** either with the means available to the centre, by hiring external resources or through a combination of methods. This process may also include **costs associated with transcription, normalisation, anonymisation or management of informed consents.** It should also assess the **cost of the infrastructures involved in the collection of data.**

Data documentation requires staff to describe the data sets and add the necessary metadata for later publication or exchange with other research staff. **Once the data is documented, it will be stored in an infrastructure whose cost may vary** depending on the required level of security and preservation.

The cost of the data publishing process should also be estimated. There may be costs associated with publishing to an external repository, maintaining an owned or consortium repository, or simply costs associated with granting persistent identifiers to data sets. There may also be **costs associated with managing rights.**

As already mentioned, the estimate must also include the costs of the specialised staff who help the research staff throughout the data management process.

RECOMMENDED ACTIONS

- **Identify all the processes involved in the centre's data management.**
- **Estimate the cost of the centre's data management.**
- **Charge, when possible, the data management expenses to the research projects.**

Additional relevant information

Uk Data Service tool to estimate the cost of data management: https://www.ukdataservice.ac.uk/media/622368/costingtool.pdf

• Economic estimate of the cost of data management carried out by the *University of Utrecht:* https://www.uu.nl/en/research/research-data-management/guides/costs- of-data-management

• Economic estimate of the cost of data management carried out by a British research group to justify the costs of a European project.

Chapter 7

A DATA MANAGEMENT POLICY MODEL

7

A DATA MANAGEMENT POLICY MODEL

The data management policy of a centre is the set of **fundamental and generic principles** that it follows in this area. This policy is implemented through **specific activities and choices depending on the project, the scientific discipline and the type of data.** It should aim **to improve data management**, ensuring its preservation, so that it is available for both internal and external reuse.

The data management policy **should contemplate the creation and maintenance of a data catalogue for the centre,** so that location of the data can be known even if it is not publicly accessible. But the policy should not include the obligation to publish the data as open data in certain repositories, for various reasons. First, because, as has already been mentioned, not all data can or should be published openly. Second, because there are already well-established subject repositories, that house data from certain disciplines or areas of knowledge, which may not meet this obligation. The policy itself should provide for its own periodic review.

A good example to follow is the policy suggested by the *LEARN*[47] project, funded by the *European Commission*. The project openly published different materials related to data management:

- A survey to find out if an institution is ready for data management.
- A list of best practices.
- A data management policy example form.
- A series of indicators[48] for monitoring the data management policy.

The suggested form has the following structure:

- An introduction explaining the reasons for the policy.
- The establishment of the jurisdiction that governs the policy.
- The applicable intellectual property regulations, as well as the ownership of the derived rights.
- The data management.
- The responsibilities, duties and rights of the parties involved.
- The validity of the policy, as well as its future revisions and updates.

A data management policy also requires the **creation and use of indicators** to monitor the progress of the organisation towards the objectives established during its implementation. If the established objectives are not achieved, then changes are required to the policy or to the actions being implemented.

Additional relevant information

47 http://learn-rdm.eu/

48 http://learn-rdm.eu/wp-content/uploads/KPIs_MD.pdf

The *LEARN* project proposes a list of indicators that should help institutions measure the success of the implementation of the data management policy. The first set of indicators is intended to prepare the institution for data management. This means setting up a group to lead the centre's data management, and having adequate services, profiles and a training programme in place. The second set of indicators should be used to measure the implementation of the data management policies. This would include, for example, the number of researchers who use the services and the possible infrastructure that the centre can offer, the number of stored and published data sets or the number of training sessions carried out and the level of participation.

7.1 Technical aspects

Each centre must inform its research and support staff of the **principles that must be followed when making decisions on technical aspects.** Generally, this information is included in a "data management regulation" that implements the institution's data management policy. In the event that there is no defined position within the centre, decision-making must be delegated to the qualified staff or research staff, on whom the responsibility will fall.

The technical issues on which institutions must express their position are the following:

- Method of publishing data produced and licences offered.
- Degree of reuse of existing data and processes for requesting access to restricted data repositories.
- Preferred formats and internal organisation of data in local infrastructure, if applicable.
- Available or preferred tools for data storage and access.
- Medium and long-term data preservation policy and responsibilities regarding data.
- Degree of obligation regarding the access, use and depositing of new data produced in subject, multidisciplinary and institutional data repositories.
- Incentives or limitations of the institution for the publication of open data or mechanisms to limit access to data produced by groups associated with the institution.
- Cost of data management and the option for the research groups or the centre itself to pass this cost on to the projects.

Additional relevant information

- *"Política de gestió de dades de recerca de la Universitat de Barcelona"*: http://hdl.handle.net/2445/142043.
- *"University College London (UCL) Research Data Policy"*: https://www.ucl.ac.uk/isd/sites/isd/files/ucl_research_data_policy_v6.pdf
- *"Centre National de la Recherche Scientifique (CNRS) Research Data Plan"*: https://www.science-ouverte.cnrs.fr/wp-content/uploads/2021/04/Cnrs_Research- Data-Plan_mars21.pdf

Glossary 8

8

GLOSSARY

• **Anonymisation of personal data:** the process through which the connection between personal data and the natural person to whom it refers is broken, thus preventing their identification. The resulting data is said to be “anonymised" and does not fall within the cases protected by the *General Data Protection Regulation (GDPR)* or other applicable regulations. When the element (software, table of values, etc.) that allows the re-identification of the person to whom the data refers, is simply isolated and detached from the data, and not destroyed, the data is said to be “pseudonymised", and the data is still considered to be personal data.

• **Catalogue**: list of objects sorted according to their characteristics or metadata. Objects can be files, data sets, articles, documents, etc. characterised by metadata or keywords that should be standard as far as possible and follow the vocabulary of the discipline. The objects in a catalogue can be searched both manually and with automatic search engines.

• **Curated data:** data accompanied by metadata which has gone through a documentation and archiving process. The curation process may fully or partially follow the *FAIR* principles.

• **Database:** structured set of data and metadata stored digitally. It allows data to be consulted manually, through a graphical user interface or *GUI*, and automated, through application programming interfaces or *APIs*. A distinction is usually made between relational and non-relational databases, depending on the characteristics of the metadata stored and the operations that need to be carried out. Relational databases use *SQL (Structured Query Language)* as their programming and communication language, while non-relational databases *(NOSQL)* allow a more flexible system to be used to store information.

• **Data accessibility:** refers to the mechanism by which data can be viewed or downloaded. Access to the data can be totally open or restricted to some type of registration or permission. For example, the type of authorisation granted by the owners of the data should be taken into account when sharing personal data, and its access should be carefully regulated and monitored, following European and national legislation on personal data protection.

• **Data citation:** unique reference to published data sets that have been reused. The citation of the data must at least include unique identifiers such as the *PID (Persistent Identifier)* or the *DOI (Digital Object Identifier)* although it usually includes more information (author, title, date, etc.).

• **Data confidentiality:** limitation of use and/or distribution of the data defined by its owner which implies putting adequate security mechanisms into effect.

- **Data management plan *(DMP)***: a living document, with updates and new versions, which contains the description of the data, the management tools and decisions chosen before and during the execution of a research project. The person responsible for the data management plan of a project can be the principal investigator or anyone else responsible for gathering the necessary information to be edited and reviewed.

- **Data management policy:** set of fundamental and generic principles that an organisation undertakes to follow with respect to data management. Data policies are implemented through specific activities and choices depending on the project, the scientific discipline and the type of data being processed.

- **Data organisation:** way in which data is stored in the workspace, in a structured and consistent way.

- **Data repository:** characterised by the ability to safely store data and display metadata so that data can be searched and found. Data repositories can be generically classified as: institutional, if they collect data from the same institution (for example, from a university or a research centre); consortium-based, if they collect data from institutions belonging to a consortium; subject-based, if they collect data from the same discipline or branch of research. There are also generic or multidisciplinary repositories for data from disciplines that do not have a specific subject repository.

- **Data reuse:** operation that consists of using data generated and published by other processes (research, measurement data, etc.) for the purposes of a new project and, based on it, to generate new scientific results. Data reuse avoids duplication of efforts in generating existing data. This is possible if the data is curated according to the *FAIR* principles.

- **Data security:** technical, administrative and physical mechanisms that ensure compliance with the restrictions of access, use and distribution of data, in compliance with the current and applicable legislation and confidentiality agreements.

- **Data Steward:** role specialising in research data management and in compliance with the *FAIR* principles. The specific competences are mainly of a scientific, technical and administrative nature. Depending on the level of expertise of the organisation's staff, specialist teams can be created for managing research data.

- **Dynamic data:** research data that is periodically updated. It is important to identify each update with version control.

- **FAIR principles (Findable, Accessible, Interoperable, Reusable):** set of principles that must be followed to ensure that research data is accessible, can be understood and can be reused, also in interdisciplinary research. These principles can be extended to software and, in general, to any other research product. The concepts of *FAIR* data and open data are not exclusive nor equivalent, but rather complementary. Certain *FAIR* data might not be open, if some kind of authorisation is needed to access and reuse it.

• **Indexing:** process that adds a field with an index (a number) in a table to make searches faster. It is used especially for databases that contain a large number of entries.

• **Intellectual property:** refers to the rights over any creation (such as artistic and scientific works, as well as databases), which corresponds to the authors, as a result of their intelligence and creativity, and to other holders. The term intellectual property is often extended to the concept of industrial property, which includes utility models, patents, and trademarks. In Spain these two concepts are regulated by different laws that grant different conditions and duration to the holders of the rights.

• **Interoperability:** characteristic of the data that means it can be read and used with different tools, by different groups or by scientific disciplines other than those that generated them. Interoperability is achieved through the use of widely accepted and used or properly described standards.

• **Licence of use:** a legal tool that defines people's rights to use a certain product that is protected as intellectual property (data, software, articles, books, etc.). The holder of the exploitation rights decides on which licence to assign to their products, according to current regulations and the agreements they have adopted (for example, with the journal in which they publish their results, with the institution for which they work, etc.).

• **Open data:** data that does not require authorisation to be accessed, examined, analysed and reused. Data can also be considered open when the access conditions are minimal (for example, after prior registration or the acceptance of very lax terms of use). The type of control or authorisation required to access the data determines whether it can be considered “open”. Otherwise, the data is considered "not open" or "closed". This concept does not necessarily offer guarantees about the quality or scientific value of the data, or about its possible future usefulness.

• **Persistent identifier (PID)**: reference that uniquely identifies a digital object (such as articles, books, but also files, data sets, etc.) and that at the same time allows the object in questions to be reached through a *URL (Uniform Resource Locator)*, independent of its physical location, thanks to an organisation that keeps it operational. Examples of persistent identifiers are the *DOI (Digital Object Identifier)* and the *handle*, which is another type of identifier generally used in repositories for publications.

• **Personal data:** According to the *General Data Protection Regulation (GDPR)*, personal data is any data that allows the identification of a person. The following is considered personal data: names, dates, physical characteristics, personal images, physical or electronic mail addresses, *IP (Internet Protocol)* addresses of devices, identification numbers, etc., when these are associated with a database through which data that identifies a person can be found.

• **Primary data:** the starting point of a research project, without any modification (measurements, surveys, observations, etc.). It should be stored together with the details of its origin.

• **Metadata:** data that describes or classifies the research data to which it refers. Metadata, properly organised and indexed, contributes to the proper management of the data, so that it is findable and reusable.

• **Personal data protection:** the European general data protection regulation (*GDPR*) contains the regulations that each specific national legislation on the use and processing of personal data must comply with.

• **Secondary data:** data that has undergone some type of processing (intermediate data or results).

• **Research data:** set of information, digitalised in files or on any physical medium, which in the research process contributes to the production of a scientific result. Reference is also made throughout the document to *data sets*

- **Research data management (RDM)**: all the actions involved in collecting, organising, preserving, accessing and publishing data.

- **Research data life cycle:** the set of phases and processes through which the data of each project passes during and after the research activity. The creation of new data feeds the cycle, which moves through the phases of processing, analysis, curation and publication of the data. The reuse of data created in other projects closes and reactivates the cycle.

- **Vocabulary:** set of terms used to describe objects in certain contexts. Choosing the appropriate vocabulary is essential to prevent data from being misinterpreted or used incorrectly. There are vocabularies that contain standard terms and definitions for many contexts.

XUNTA DE GALICIA

Institució CERCA Centres de Recerca de Catalunya

Xacobeo 21·22

FEUGA FUNDACIÓN EMPRESA-UNIVERSIDAD GALLEGA