

Guide on Best Practices for Data Sharing and Usage Trackability	
Work Package	2. Data Curation and Sharing
Task	--
Delivery Date	November
Dissemination Level	Public
Authorship	Cátia Carvalho Ana Inácio Daniel Alves Bruno Almeida Rui Araújo
Reviewers	João Leitão and Juliana Monteiro
Conclusion	December

Acknowledgements:

The NOVA.ID-RDM-CC Competence Centre is funded under the PNCADAI - National Programme for Open Science and Open Research Data, as part of Measure RE-C05-i08 - More Digital Science of the PRR - Recovery and Resilience Programme.

Summary

Índice

SUMMARY	2
INTRODUCTION	4
OBJECTIVES AND TARGET AUDIENCE	4
ALIGNMENT WITH FAIR PRINCIPLES	4
GUIDING PRINCIPLES FOR COLLECTING AND SHARING RESEARCH DATA	5
SOURCES AND FURTHER INFORMATION.....	5
BEST PRACTICES FOR DATA SHARING	7
SELECTING AND PREPARING DATA FOR SHARING	7
FILE FORMATS FOR PRESERVATION.....	7
DOCUMENTATION AND METADATA	8
DOCUMENTATION	8
METADATA SCHEMAS AND OTHER FRAMEWORKS FOR ENHANCING DISCOVERABILITY	9
FINDING TRUSTED REPOSITORIES	12
LICENSING AND REUSE TERMS.....	15
SOURCES AND FURTHER INFORMATION.....	16
BEST PRACTICES FOR ACCESS CONTROL	17
PROTECTING SENSITIVE DATA	17
GENERAL PRINCIPLES	17
TECHNIQUES FOR PROTECTING SENSITIVE DATA.....	18
IMPLEMENTING ACCESS CONTROL AND RESTRICTIONS	20
SOURCES AND FURTHER INFORMATION.....	22
BEST PRACTICES FOR TRACKING DATA USAGE	23
CITATION AND ATTRIBUTION.....	23
PERSISTENT IDENTIFIERS	23
DATASET CITATIONS AND REFERENCES	23
DATA USAGE METRICS	24

COMMON DATA USAGE METRICS IN REPOSITORIES 24
TOOLS AND INITIATIVES FOR TRACKING DATA USAGE 25
SOURCES AND FURTHER INFORMATION..... 25

Introduction

NOVA.ID Research Data Management Competence Centre (NOVA.ID-RDM-CC) aims to establish an infrastructure and set of best practices for research data management (RDM) starting from multiple disciplines within natural and exact sciences and making interdisciplinary connections to other scientific fields (economics, medicine, public health, law, social sciences, humanities and heritage), aiming to be a leader in the promotion of open science at the national level.

Objectives and target audience

This guide puts forward best practices for sharing research data, managing access control and ensuring usage traceability based on a review of relevant guidelines. It is mainly intended for researchers and research support staff involved in projects where the creation and sharing of research data are significant activities. The contents of this guide are useful for RDM in several disciplines, from natural and exact sciences to social sciences and humanities.

Alignment with FAIR principles

The FAIR principles are a collection of guidelines to promote the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of research data, focusing on the capacity of computational systems to find, access, interoperate, and reuse data with as little human intervention as possible. The recommendations compiled in the present guide are compatible with the creation and sharing of data according to the FAIR principles. This means that, for shared data to be **findable**, the use of persistent identifiers (e.g. DOI, Handle) is paramount, along with rich metadata descriptions, and visibility in search engines and/or data registries. For data to be **accessible**, it needs to be retrievable by its identifier using an open and free standardized communications protocol (e.g. HTTP, FTP), and the metadata description needs to always be accessible, even if the data are no longer available. For data to be **interoperable**, it needs to be structured using a shared formal schema or ontology (e.g. RDF, Dublin Core), containing meaningful links to other data sources (e.g. links to publications drawing conclusions from the dataset). The vocabularies (e.g. thesauri, taxonomies) used in the metadata descriptions need to be available as FAIR data, for instance using persistent identifiers for terms or concepts. Lastly, for data to be **reusable**, its associated metadata needs to describe the context in which the data was generated, including licensing and usage terms, and provenance (e.g. how to cite the data). Domain-relevant standards for

creating data and metadata (e.g. MIAME for microarray experiments in the life sciences) need to be followed to increase reusability within the research community.

The following chapters of this guide describe the FAIR principles in greater detail with regard to sharing research data, implementing access control, and tracking usage. At this point it is useful to stress the distinction between **FAIR data** and **open data**, as they are quite different concepts. Open data, as a key component of open science, implies that data should be openly accessible to the community, while the FAIR principles do not imply public access to research data. Several reasons may lead a researcher to restrict access to his or her data, including the protection of personal privacy, trade secrets, or national security interests. Restricted or even closed data may remain accessible, according to the FAIR principles, through its associated persistent identifier and rich metadata description.

Guiding principles for collecting and sharing research data

Collecting and sharing data falls within the scope of research ethics and, therefore, should comply with all relevant regulations and ethics codes, including those of the host institution, of partner institutions and of the funding agencies, if applicable. In general, the following principles should be followed when collecting and sharing research data:

- **Integrity:** Data should be collected and shared to ensure the quality, accuracy and reliability of the research results.
- **Fairness:** The rights of individuals and groups, human dignity and freedom should be respected, avoiding any kind of bias, discrimination and unfair treatment.
- **Transparency:** The purpose of data collection, the methods used, and the level of access to the data should be clearly stated, and explicitly and voluntarily authorized by participants.
- **Accountability:** Clear roles should be defined regarding responsibility and accountability for data management, keeping records of all relevant documentation.
- **Openness:** Data should be made available as openly as possible and as closed as necessary to ensure that sensitive information is protected while following open science practices.
- **Reproducibility:** The available (meta)data and code should allow other researchers to reproduce or analyze the results described in other research outputs, such as journal articles.

Sources and further information

ALLEA. (2023). *The European code of conduct for research integrity* (Revised Edition). <http://www.doi.org/10.26356/ECOC>

Despacho n.º 15464/2014 [Código de Ética Da Universidade Nova de Lisboa], Diário da República n.º 245/2014, Série II (2014). <https://diariodarepublica.pt/dr/detalhe/despacho/15464-2014-65953794>

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, R., Tait, J., & Zijlstra, T. (n.d.). *Open data handbook*. Open Knowledge. <http://opendatahandbook.org/guide/en/>

European Research Council. (n.d.). *Ethics guidance*. Retrieved September 16, 2025, from <https://erc.europa.eu/manage-your-project/ethics-guidance>

GO FAIR. (2022, January 21). *FAIR principles*. <https://www.go-fair.org/fair-principles/>

UKCORI. (2025). *The concordat to support research integrity*. UKCORI. <https://ukcori.org/wp-content/uploads/2025/04/The-Concordat-to-Support-Research-Integrity-2025.pdf>

Wilkinson, M. D. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci Data*, 3(160018). <https://doi.org/10.1038/sdata.2016.18>

Best practices for data sharing

Selecting and preparing data for sharing

Sharing and archiving data in a trusted repository ensures that it will remain accessible in the long term, either by the same researcher or by other researchers. This is also important for verification purposes, e.g. as part of a peer-review for publishing a paper, or of an evaluation process of a research center.

Not all data created throughout a research project needs to be shared with the community, since it requires a significant investment in terms of effort and costs. Only a selection of data should be prepared for long-term preservation and publication. This typically includes:

- Data which funders', publishers' or academic institutions' policies require to be published and preserved.
- Data preserved to comply with legal or ethical requirements (e.g. clinical trial data).
- Unique data or data that is difficult to re-generate (e.g. raw data, analysis workflows).
- Quality or high-value data with potential to be reused in the future.

Selected data needs to be prepared for sharing and preservation. Researchers should ensure that quality control measures have been followed and that their datasets are functionally usable, so that they can be reused in future projects:

- Data should be made available in standard open formats and/or in widely used formats in the research community.
- Datasets should be described by self-explanatory documentation, e.g. README files.
- Datasets should include information about provenance, i.e. the origin, collection methods, processing and further context about the dataset.
- Datasets should include information about licensing or usage agreements.
- The files and folders in a dataset should be well organized and follow a consistent naming convention.

File formats for preservation

The choice of file formats is an important consideration when archiving and sharing data in repositories. Working with proprietary formats may be necessary during the active phase of research, especially in more technology-driven fields (e.g. omics, materials science). However, open formats should be used for long-term preservation, access and interoperability of research datasets in repositories. Whenever possible, proprietary formats should be converted or exported to open formats. The table below includes examples of preferred, acceptable and deprecated formats for several types of data.

Type of data	Preferred	Acceptable	Deprecated
Tabular data with extensive metadata	CSV, HDF5	TXT, HTML, TEX, FASTQ, POR	
Tabular data with minimal metadata	CSV, TAB, ODS, SQL, TSV	XML, XLSX	XLS, XLSB
Rich text	ODT, MD, TEX, PDF/A	DOCX	DOC, PDF (OTHER THAN PDF/A)
Plain text	UNICODE TXT		NON-UNICODE TXT
Code	M, R, PY, IYPNB, RSTUDIO, RMD, NETCDF	SDD	MAT, RDATA
Raster images	TIFF, DICOM	CZI, LIF, NEF, PNG	JPEG, PS, EPS, BMP
Vector images	SVG	PS, EPS	AI, WMF, EMF, CDR
Audio	MKA, FLAC	WAV, MP3	
Video	MKV	MPG, MP4, MJPEG	AVI, MOV, QT
3D data	X3D, X3DV, X3DB, PDF3D, POV, PDBML	DWG, DXF, PDB	PXB
Geospatial data	NETCDF, GIS, SHP, SHX, DBF, PRJ, SBX, SBN, POSTGIS, TIF, TFW, GEOJSON	OGM, WEBM	
Generic data	XML, JSON, RDF		

Documentation and metadata

Documentation

Data documentation is a contextual description of a dataset, including all necessary information to understand and reuse the data, and to verify and reproduce research results. In larger research projects, spanning several years, documentation is often required for onboarding new team members. It is good practice to start documenting as soon as possible, and to maintain accurate and comprehensive documentation throughout the research process. This will save time in the long run, as it typically will be required for publishing datasets in repositories.

There are two levels of documentation to consider when preparing data for sharing and preservation:

- **Project-level documentation** describes the aims of the study, authors, institutions involved, funders, grant numbers, types of included data (e.g. interviews, images, statistical data), methods and instruments for data collection and processing,

licensing or usage terms, identifiers for each dataset, folder structure, file naming conventions and other general information.

- **Data-level documentation** provides information about individual items in a dataset, including the meaning of each variable name, label or ID, datatypes (e.g. string, numeric, date), units of measurement (e.g. cm, g, °C), terms from controlled vocabularies or ontologies accepted as values for data fields (e.g. DCMI Type Vocabulary, Gene Ontology), codes for missing values (e.g. NA, blank cells), etc.

Both project- and data-level documentation may take several forms, depending on the type of data and the research field. Some common forms of documentation include:

- **README file:** A document describing the data collection, processing and analysis. Repositories often require for datasets to be accompanied by README files for publication.
- **Codebook:** A document including all essential information about a dataset, including information about the study, the files in the dataset, variables, labels, categories, etc.
- **Data dictionary:** A document outlining the structure, content and variable or field definitions of files in a dataset (e.g. ID, label, field type, unit).
- **Data list:** A document listing each variable or field in a dataset (e.g. interview ID, date of interview, age, gender).

Documentation files may be generated in several formats depending on the tools used and on the specific needs of the study. Whenever possible, open formats should be used for long-term preservation over proprietary formats. Following the FAIR principles, structured and machine-actionable formats (e.g. XML, JSON, RDF, CSV) are preferable over text formats (e.g. PDF, ODF, TXT). README files, which are usually created in simple text format (TXT), are an exception, since they are human-readable files which are not meant to be machine-actionable.

Metadata schemas and other frameworks for enhancing discoverability

Metadata are structured descriptions of data used for cataloging and discovery purposes, allowing users to find data, determine its reusability and cite it. Metadata schemas are sets of rules and defined elements for describing data. In a metadata schema, some elements are mandatory while others are recommended or optional. Metadata schemas are often associated to controlled vocabularies or ontologies, which supply lists of possible values for its elements in a database. For example, a Type metadata element could be linked to a controlled vocabulary of resource types (e.g. text, image), such as the [DCMI Type Vocabulary](#).

There are several types of elements in a metadata schema. The primary types of metadata are the following:

- **Descriptive metadata:** Elements that allow to identify data for discovery and citation purposes (e.g. Title, Creator, Subject, Identifier).
- **Structural metadata:** Elements indicating how data components are organized and related (e.g. files and folders included in a dataset).
- **Administrative metadata:** Elements for managing a data resource (e.g. creation and modification dates, file types, access permissions, rights information).

Some metadata schemas are generic, while others are discipline specific. The following sections list some of the more cited examples of each category. These examples also include frameworks that, while not metadata schemas, are nevertheless relevant for creating or documenting FAIR data. Researchers are encouraged to use directories, such as [FAIRsharing](#), [RDA Alliance](#) or [Linked Open Vocabularies](#), for finding metadata schemas and other standards for their projects.

Generic metadata schemas and other frameworks

[DataCite Metadata Schema](#). List of metadata properties for consistent identification, citation and retrieval of research outputs and other resources. DataCite Metadata Schema is widely supported by data repositories for metadata descriptions.

[Data Catalog Vocabulary \(DCAT\)](#). RDF vocabulary for describing datasets and for interoperability between data catalogs in the web.

[DCMI Metadata Terms \(Dublin Core\)](#). General purpose metadata schema for describing any kind of digital or physical resource. DCMI Metadata Terms includes the original Dublin Core elements (e.g. Creator, Title, Identifier) along with several more properties, classes and encoding schemes.

[Resource Description Framework \(RDF\)](#). Framework for representing information in the web, used for creating ontologies, controlled vocabularies, knowledge bases and other types of resources.

[Schema.org](#). RDF vocabulary for enhancing the findability of structured data in the web, helping search engines and other applications to make sense of the contents of a website.

Disciplinary metadata schemas and other frameworks

Natural sciences, medical and health sciences

[BioSchemas](#). Extension of Schema.org for biomedical resources.

[Darwin Core \(DwC\)](#). Set of metadata standards for compiling and sharing biodiversity data, originally developed as an extension of Dublin Core.

[ECRIN Metadata Schemas for Clinical Research](#). ERCIN (European Clinical Research Infrastructure Network) maintains a metadata schema for FAIR data regarding studies and data objects in clinical research.

[Fast Healthcare Interoperability Resources \(FHIR\)](#). Standard for clinical and patient health data exchange published by Health Level Seven International.

[International Virtual Observatory Alliance Technical Specifications \(IVOA Standards\)](#). Set of standards for interoperability and integration of astronomical archives into an international virtual observatory. These standards include metadata schemas for several data types, including photometry data (PhotDM), simulation data (SimDM) and observational data (ObsCoreDM).

[ISA Framework](#). Schema for describing experimental data and workflows in the life sciences.

[ISO 19115 – Geographic Information](#). International standard for describing geographic information and services. The standard is published in 3 parts: ISO 19115-1:2014 contains the fundamental concepts of the standard, ISO 19115-2:2019 describes extensions for imagery and gridded data, and ISO 19115-3:2023 provides an XML schema implementation.

[Minimum Information Standards \(MI Standards\)](#). Set of guidelines and formats for reporting experimental data generated by high-throughput methods in the life sciences. MI Standards are available for a wide range of experiment types, including microarray (MIAME), RNA-Seq (MINSEQE), proteomics (MIAPE) and plant phenotyping (MIAPPE).

[NFDI4Health Metadata Schema](#). Schema for the structured collection and description of metadata in health, clinical, and epidemiological studies.

Social sciences, arts and humanities

[Conceptual Reference Model \(CIDOC-CRM\)](#). Ontology for describing cultural heritage as linked data developed by the International Council of Museums. Several models extend CIDOC-CRM to specific applications, e.g. CRMarchaeo for archaeological excavations and CRMsoc for social phenomena.

[Data Documentation Initiative \(DDI\)](#). The DDI Alliance promotes the development of several metadata standards for describing data produced in surveys and other observational methods in the social, behavioral, economic and health sciences. DDI-Codebook, the more relevant standard for the present guide, defines a set of elements for documenting the

content, meaning, provenance and access for a dataset. DDI-Lifecycle is a more comprehensive schema, designed for larger, linked and complex datasets.

[Text Encoding Initiative \(TEI\)](#). Guidelines for encoding, describing and exchanging digital texts in XML, widely used in digital humanities projects. The TEI guidelines include modules for specific types or modalities of text, including, manuscripts, other primary sources and language corpora. Descriptive metadata may be encoded within the TEI files.

[VRA Core](#). Metadata schema for the description of visual works and the images documenting them, including paintings, drawings, sculpture, architecture, photographs, decorative and performance art.

Finding trusted repositories

Publishing research data in repositories is presently a requirement or recommendation of funders and publishers in most domains. Making your data available is a good scientific practice, as it allows the research results to be verifiable and reproducible. Repositories also provide significant advantages over other data sharing practices, such as publishing data as supplementary materials in journals, as repositories are designed for long-term preservation of complete datasets.

Trusted repositories are reliable digital storage services, providing long-term access to datasets and following strict organizational, technical, and procedural standards. They are often certified by an independent authority, such as CoreTrustSeal, although certification is not strictly necessary for assessing the trustworthiness of a repository. The TRUST principles provide a broader framework for this assessment:

- **Transparency.** Clear and relevant information is provided about the repository's services, target user community, terms of use and policies.
- **Responsibility.** The repository ensures the integrity and persistence of its holdings.
- **User focus.** The repository supports the relevant metadata standards, file formats, controlled vocabularies and ontologies for its target users.
- **Sustainability.** The repository ensures uninterrupted access to its services and the long-term preservation of its data holdings.
- **Technology.** The repository supplies the necessary infrastructure for supporting its services and the security and integrity of its data holdings.

Choosing a trusted repository will depend on several factors, including the type of data to be published, the licensing and type of access to the data and the existence of any contractual requirements by funders or publishers. The institution in which the study is carried out might

also recommend your data to be archived in its own data repository. In general, the following steps are recommended:

1. Use a disciplinary repository for your type of data.
2. Alternatively, use an institutional repository, if available and if it assures long term preservation.
3. Use a trusted general-purpose repository.
4. Use an established registry, such as [re3data](#) or [FAIRSharing](#), for searching trusted repositories for your domain and type of data.

As a reference, the following sections include examples of trusted repositories, both general-purpose and disciplinary. The following list does not include institutional data repositories.

General-purpose repositories¹

[Dryad](#). Repository service managed by a non-profit organization registered in the United States, with several academic and publishing partners. Although Dryad originated from the need to share data underlying publications in the life and medical sciences, it has since expanded to other domains. The service requires data publishing charges to be paid either by the authors or through partner institutions.

[Figshare](#). Commercial repository software provider which offers both free and premium repository services. Figshare is supported by Digital Science, a subsidiary of Springer Nature, and is hosted by Amazon Web Services.

[Zenodo](#). Repository hosted in the CERN Data Center and funded by the European Commission through several OpenAIRE projects, as well as by other institutions. Zenodo is based on the open-source InvenioRDM software.

[Open Science Framework \(OSF\)](#). Platform for collaborative work in research projects which allows the sharing and long-term preservation of manuscripts, registered materials and datasets.

Disciplinary repositories

Natural sciences, medical and health sciences

¹ FCT-FCCN is currently implementing a national data repository service, POLEN DataHub, to be launched in the second quarter of 2026. When available, POLEN DataHub should become a trusted repository for research data in all domains. More information available at <https://polen.fccn.pt>.

[Cambridge Structural Database \(CSD\)](#). Repository for chemical crystallography data of organic and metal-organic compounds. CSD is maintained by the Cambridge Crystallographic Data Center.

[European Nucleotide Archive \(ENA\)](#). Repository providing a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. ENA is managed by the European Bioinformatics Institute and is part of the International Nucleotide Sequence Database Collaboration, which also includes GenBank.

[GenBank](#). Genetic sequence repository of all publicly available DNA sequences for hundreds of thousands of known species. GenBank is managed by the National Center for Biotechnology Information in the United States and is part of the International Nucleotide Sequence Database Collaboration, which also includes the European Nucleotide Archive (ENA).

[Global Biodiversity Information Facility \(GBIF\)](#). International network and infrastructure providing open access to biodiversity data, funded by the world's governments.

[DMPortal](#). Repository managed by BioData.pt, the Portuguese distributed infrastructure for Life and Health data, which integrates ELIXIR ERIC, the European life sciences infrastructure. DMPortal is open to all research fields in the life sciences but focuses on human health and plant science data.

[European Genome-phenome Archive \(EGA\)](#). Repository for archiving and sharing personally identifiable genetic, phenotypic, and clinical data generated for the purposes of biomedical research projects or in the context of research-focused healthcare systems.

[Materials Cloud](#). Repository for sharing and disseminating data in computational materials science, offering educational, research, and archiving tools as well as simulation software and services.

[PANGAEA](#). Repository for archiving and publishing data in earth and environmental sciences.

Social sciences, arts and humanities

[Arquivo Português de Informação Social \(APIS\)](#). Repository for the preservation and dissemination of social science data. APIS integrates CESSDA ERIC, the European infrastructure of social science data archives.

[PORTULAN CLARIN repository](#). Repository managed by the Portuguese Research Infrastructure for the Science and Technology of Language, part of the international research infrastructure CLARIN ERIC.

Licensing and reuse terms

A license defines the terms of use of a dataset, including information regarding intellectual property, and information regarding personal rights in the case of data describing human beings. Licensing is a key aspect of the FAIR principles, as it determines if and how a dataset will be reusable. Researchers should confirm data ownership before publishing data, as there might be rights belonging to a third party, e.g. in projects involving contributions from several institutions. Once ownership has been established, an appropriate license needs to be selected for sharing the datasets. Which license to choose will depend on the type of data and rights involved and should be compatible with funders' mandates and institutional policies. The following is recommended:

- **Choose the least restrictive license as possible.** If the dataset can be published as open data, then a permissive license should be chosen, ideally a public domain license such as [Creative Commons CC-0](#) or [Open Data Commons Public Domain Dedication and License](#). More restrictive licenses may be required for datasets with creative content or that are otherwise covered by copyright (e.g. photographs or drawings).
- **Avoid overly restrictive licenses.** Restrictive licenses, e.g. prohibiting the creation of derivative works, place legal barriers on dataset reuse, whose terms may not even be applicable to all data, as they were designed specifically for creative works.
- **Choose an appropriate license for code.** Code should be licensed separately from data using a software-specific license. Open-source licenses, i.e. approved by the [Open Source Initiative](#), are recommended for code underlying research datasets.

There are several tools for helping researchers choose a license for their data, software and other works:

- [Public License Selector](#). Online tool integrated in EUDAT's services for selecting a license for data or software.
- [Creative Commons License Chooser](#). Online tool for selecting a Creative Commons license for creative content.
- [Choose a License](#). Online tool managed by GitHub for choosing an open-source license for code.
- [Licenses compatible with the Open Definition](#). List of licenses put forward by the Open Definition Initiative that are applicable to data or creative content.

- [Licenses approved by the Open Source Initiative](#). List of open-source licenses for code.

Sources and further information

Borba, F., Costa, L. da, Moura, P., Gomes, A. C., Pereira, A. A., Carvalho, J., Vieira, A., Miranda, P., & Príncipe, P. (2022). *Como preparar dados para depositar e publicar no repositório*. Zenodo. <https://doi.org/10.5281/zenodo.7082551>

CESSDA Training Team. (2020). *CESSDA Data Management Expert Guide*. CESSDA ERIC. <https://doi.org/10.5281/zenodo.3820472>

ELIXIR. (n.d.). *RDMkit: The ELIXIR Research Data Management toolkit for Life Sciences*. <https://rdmkit.elixir-europe.org/>

Grupo de trabalho - Repositórios de Dados: Tecnologia, organização e certificação. (2023). *Como identificar um repositório confiável: Dicas práticas para investigadores, gestores de dados e de ciência*. Zenodo. <https://doi.org/10.5281/zenodo.10092359>

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J. (2020). The TRUST Principles for digital repositories. *Scientific Data*, 7(1), 144. <https://doi.org/10.1038/s41597-020-0486-7>

OpenAIRE. (n.d.). *Data formats for preservation*. OpenAIRE. <https://www.openaire.eu/data-formats-preservation-guide>

OpenAIRE. (n.d.). *How to find a trustworthy repository for your data*. OpenAIRE. <https://www.openaire.eu/find-trustworthy-data-repository>

OpenAIRE. (n.d.). *Research Data Management Handbook*. OpenAIRE. <https://www.openaire.eu/research-data-management-handbook>

Tóth-Czifra, E. (2019). *DARIAH Pathfinder to Data Management Best Practices in the Humanities*. Version 1.0.0. DARIAH Campus [Pathfinder]. <https://hdl.handle.net/21.11159/019595b2-cca1-70ad-b1e3-d088b4409de5>

UK Data Service. (2025). *Research data management. Documenting data*. <https://ukdataservice.ac.uk/learning-hub/research-data-management/#document-your-data>

Best practices for access control

Protecting sensitive data

General principles

Sensitive data are data that needs to be protected against unwanted disclosure and whose access should be safeguarded. Protection of sensitive data may be required for legal or ethical reasons, including questions pertaining to personal privacy and proprietary considerations. Types of sensitive data include:

- **Personal data.** Names, identification numbers and other identifiers or sensitive metadata; physical, physiological, genetic or mental characteristics; economic, social or cultural characteristics; location data from GPS devices or mobile phones.
- **Confidential data.** Trade secrets, investigations, data protected by intellectual property rights; security data, such as passwords, financial information, information affecting national security, military intelligence.
- **Biological data.** Data pertaining to endangered species when their survival depends on the protection of their location.

When handling sensitive data, special attention needs to be placed on collecting, processing and storing data throughout the research life cycle. This is particularly relevant in the case of personal data in which a living person is directly or indirectly identified. For example, a person may be directly identified through identifiers, such as their name, address, phone number or email, or through a photo or an interview. A person may be indirectly identified based on other characteristics, e.g. their employer, or whether they belong to small or unique subgroups, geographical areas, or through a combination of otherwise common attributes (e.g. date of birth, gender and rare medical condition).

The following principles are recommended for research activities involving personal data:

- **Limit the handling of personal data.** Gather only the strictly necessary data for answering the research questions. This principle can be implemented in several ways, e.g. limiting the number of participants, the number of attributes or questions, the granularity of the data (e.g. by using age groups instead of exact ages or birth dates), the transmission and internal access to the data.
- **Ensure that informed consent was obtained.** The participants must have been previously informed of and agreed to the purpose of the data collection, which entities are responsible for the processing of the data, the retention period, how the data can be reused in other projects, and if any resulting anonymized data will be

shared in public repositories. It should be clear that participants have the right to withdraw their consent at any time without having to specify a reason.

- **Limit the retention period.** Personal data should only be retained for as long as it is necessary. A clear limit should be placed for its elimination or anonymization when the identification of the participants is no longer necessary.
- **Limit the access to direct identifiers.** Personal data can be masked through the technique of pseudonymization, referred in the following section, in which the identification of participants is no longer possible without further information.

In the EU, any research activity involving information about individuals is bound by the [General Data Protection Regulation \(GDPR\)](#). The GDPR defines more robust protections for two categories of personal data considered to be especially sensitive and high-risk:

- **Special categories of data** (GDPR Article 9):
 - Personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership.
 - Genetic data, biometric data, data concerning health or data concerning a natural person's sex life or sexual orientation.
- **Data relating to criminal convictions and offences** (GDPR Article 10):
 - Criminal convictions.
 - Criminal charges (past or ongoing).
 - Criminal offences (alleged or proven).
 - Related security measures.

The collection and processing of the special categories of personal data identified in Article 9 requires the unambiguous and explicit consent from the participants, e.g. through signed consent forms. The processing of data pertaining to criminal convictions and offences can only be allowed under the control of an official authority, such as the Prosecution Office, the courts or the competent police forces, and only if the EU or Member State law allows the processing of these data, provided that adequate safeguards are in place for ensuring the rights of data subjects, including clear limits on the purpose for data processing, defined retention periods and technical and organizational security measures.

Techniques for protecting sensitive data

Based on the principles outlined in the previous section, researchers are encouraged to implement one or more of the following techniques to ensure the safeguarding of collected and processed sensitive data.

Anonymization

Anonymization allows to greatly minimize the ethical and legal concerns of handling personal data. Through anonymization, data are processed to prevent it ever being associated with identifiable individuals, even by the research team. Under the GDPR, anonymized data is no longer considered to be personal or sensitive data. There are several anonymization techniques, including:

- **Suppression.** Removal of direct identifiers, such as names, addresses, emails and phone numbers. However, individuals could still be identifiable by indirect means, so this technique may not be sufficient to effectively anonymize a dataset.
- **Generalization.** Reducing the granularity of the data to make individuals less identifiable, e.g. replacing ages with age groups, exact dates with years or postcodes with regions. This is especially important when in studies involving small or unique groups or in very restricted geographical areas.
- **Data masking.** Replacing sensitive data with realistic-looking data through substitution, shuffling or character masking (e.g., Lxxx Cxxxxx). The resulting data can be useful for testing, analytics or training purposes.
- **Noise addition.** Adding small amounts of statistical noise to a dataset, e.g. based on a Laplace distribution, a Gaussian distribution or uniformly from a range of values. For example, the blood pressure values in a health dataset can be anonymized by adding noise to all values from a Gaussian distribution.
- **K-anonymity.** Data anonymization technique where every combination of values of quasi-identifiers (e.g. age, gender, postcode) can be indistinctly matched to at least k persons. For example, in a 5-anonymous dataset, at least 5 records share the same quasi-identifiers.

There are specific tools for data anonymization, most notably [Amnesia](#), which is provided by OpenAIRE. Amnesia allows for both pseudonymization and anonymization based on several of the above-mentioned techniques.

Pseudonymization

Pseudonymization is a de-identification technique in which direct identifiers, such as names, addresses, emails and phone numbers, are replaced with persistent codes whose corresponding values remain accessible by specific members of the research team. Therefore, a pseudonymized dataset still falls within the GDPR, since individuals can be identified through a table, a key or other means of corresponding the codes to the real identifiers. Common pseudonymization techniques include:

- **Tokenization.** Identifiers are replaced with random or sequential tokens generated by a tokenization system (e.g. emails in a dataset can be replaced with random tokens, such as T9F4-2A71).
- **Table-based pseudonyms.** Identifiers are replaced with codes whose corresponding identifiers are managed in a separate lookup table (e.g. a Patient ID table in which codes, such as P001, are matched to real names).
- **Hashing.** Identifiers are transformed through a cryptographic hash function into a fixed-length value, which is not easily reversible.

Encryption

Encryption is a protection technique in which data is converted into an unreadable ciphertext using a mathematical key. The original data can only be recovered by using the correct decryption key. Data encryption is essential for research activities involving sensitive data at several levels:

- **Protecting data at rest (i.e. during storage).** Encryption should always be used for protecting any sensitive data stored on researchers' laptops, external hard drives, institutional servers or cloud platforms. This includes both working data and any backups that were carried out. Common methods include full disk encryption (e.g. using BitLocker or FileVault), encrypted partitions or encrypted archives (e.g. using VeraCrypt).
- **Protecting data in transit (i.e. during transmission).** In larger research projects, data often needs to be shared between institutions or sites, often through cloud services. Sensitive data should only be transmitted if necessary and only if secure technologies are implemented, such as TLS/SSL (for encrypting data accessible through websites), SSH (for secure client-server communication), VPN (for secure internet access) or PGP (for secure email communication).

Implementing access control and restrictions

Access control measures should be implemented both in the active research phase, e.g. when processing sensitive data, and in the post-research phase, e.g. when archiving the resulting anonymized datasets in trusted repositories.

In the **active research phase**, the following should be considered:

- **Document access policies** in a data management plan as early as possible. This should clearly outline roles and responsibilities, which users will collect and manage sensitive data, the processing workflow and data retention and destruction procedures.

- Follow the **principle of least privilege** by only granting each team member the minimum access required for their tasks.
- **Manage access throughout the research lifecycle**, periodically reviewing and adjusting access when team members join or leave the project, and updating user permissions when the project moves from data collection to processing and analysis.
- In most cases, **role-based access control** should be followed in the processing workflow to ensure that only authorized members of the research team have access to sensitive information. Typically, a small group, such as the PI or a designated data manager, will have access to the sensitive data stored in an encrypted archive, while other research team members will only have access to pseudonymized or anonymized datasets for analysis.
- Require **strong authentication measures** for team members, such as institutional login with two-factor authentication (2FA) enabled.
- Use **secure platforms** for storing and sharing data, such as institutional secure servers or controlled-access shared drives. These platforms should log and monitor user access, so that the PI or data manager knows which files were last accessed and by which users. Sensitive data should never be stored on non-encrypted drives or on consumer cloud services that are not institutionally approved.

When **archiving datasets in repositories**, the following should be considered:

- **Many repositories do not allow to archive datasets containing personal or sensitive information.** This is the case of general-purpose repositories, such as Zenodo, Dryad or Figshare, but also of some disciplinary repositories, such as GBIF, which restricts the publishing of the exact locations of endangered species, or GenBank, which does not allow the inclusion of data revealing the personal identities. In these cases, the data needs to be de-identified and all sensitive information removed before archiving.
- Most repositories offer **access restriction options** when archiving datasets, including options such as public (i.e., open access), embargoed or restricted. Restricted datasets can only be accessed by authorized users. Some repositories have the functionality to request access to restricted datasets. An embargo period may be applied until after the publication of the results in a paper or through other means. Access can also be restricted for ethical reasons, e.g. to ensure the responsible use of anonymized datasets which result from the processing of personal data.
- **Secure repositories**, such as EGA, facilitate the distribution of personally identifiable information under controlled access. These repositories have the necessary infrastructure and governance for handling sensitive information,

including data access committees which are responsible for data release by request, based on participant consent and/or ethics review.

Sources and further information

CESSDA Training Team. (2020). *CESSDA Data Management Expert Guide*. CESSDA ERIC. <https://doi.org/10.5281/zenodo.3820472>

ELIXIR. (n.d.). *RDMkit: The ELIXIR Research Data Management toolkit for Life Sciences*. <https://rdmkit.elixir-europe.org/>

David, N., Cordeiro, M., Cipriano, G., Boavida, C. P., Figueiredo, J., Conceição, M. C., Ochôa, P., Gallagher, K., & Cunha, C. (2025). *Toolkit sobre Questões Jurídicas, Proteção de Dados e Licenças*. Zenodo. <https://doi.org/10.5281/zenodo.17632500>

OECD. (n.d.). *Roles and responsibilities of researchers*. Retrieved December 4, 2025, from <https://www.oecd.org/en/toolkits/access-to-research-data-from-public-funding-toolkit/data-governance-for-trust/roles-and-responsibilities-of-researchers.html>

OpenAIRE. (n.d.). *How to deal with sensitive data*. OpenAIRE. <https://www.openaire.eu/sensitive-data-guide>

Tóth-Czifra, E. (2019). *DARIAH Pathfinder to Data Management Best Practices in the Humanities*. Version 1.0.0. DARIAH Campus [Pathfinder]. <https://hdl.handle.net/21.11159/019595b2-cca1-70ad-b1e3-d088b4409de5>

UK Data Service. (2025). *Data Protection: Access control*. <https://ukdataservice.ac.uk/learning-hub/research-data-management/data-protection/access-control/>

University of York. (2025). *Restricting access*. <https://subjectguides.york.ac.uk/rdm/restricting>

Best practices for tracking data usage

Citation and attribution

The ultimate purpose of the FAIR principles is to increase the reusability of research data. This raises the question of how to monitor the usage of shared datasets over time. In principle, datasets created and shared following the FAIR principles will have a greater impact in terms of usage metrics, such as views, downloads and citations. The use of persistent identifiers and dataset citation are key practices that must be widely adopted in all research fields for promoting dataset reuse and usage tracking.

Persistent identifiers

A persistent identifier (PID) are long-lasting references to digital resources. There are several PID that are relevant for research data management, in particular for facilitating usage tracking. Trusted repositories will support one or more of the following PID systems for describing datasets and their authors, contributors and associated organizations:

- **Digital Object Identifier (DOI).** PID system for identifying academic, professional or governmental information items, such as publications or datasets. Example: <https://doi.org/10.4232/1.13024> (DOI of the dataset *European System of Social Indicators: Housing, 1980-2013* in the repository of GESIS – Leibniz Institute for the Social Sciences).
- **Handle.** PID system for assigning PID (referred to as “handles”) to information resources. Example: <https://hdl.handle.net/21.11129/0000-000B-D30B-B> (Handle of the *BioLexicon* dataset in the PORTULAN CLARIN repository).
- **Open Researcher and Contributor ID (ORCID).** PID system for identifying authors and contributors in research outputs. Example: <https://orcid.org/0000-0003-1279-3709> (ORCID iD of Tim Berners-Lee).
- **Research Organization Registry (ROR).** PID system for identifying research organizations, such as universities, schools or laboratories. Example: <https://ror.org/02xankh89> (ROR ID of Universidade NOVA de Lisboa).

Dataset citations and references

Beyond being a good academic practice, dataset citation is fundamental for allowing usage tracking using PID systems for identifying datasets and linking them to publications where they are referenced. The following should be considered:

- Incorporate **data citation principles** in your research practices. Data citations should be valued equally as citations of other research outputs, such as publications.

They are also fundamental for giving credit to the respective authors, to comply with legal attribution when required by the licensing terms, and to support claims in scholarly literature. Data citation should always rely on PID for identifying datasets, which should persist, along with its descriptive metadata, even beyond the lifespan of the data being described.

- Citations will be based on the **metadata elements** supplied for dataset description. The DataCite Metadata Schema, which is widely used for dataset description in repositories, includes mandatory elements that are essential for citing datasets, such as Identifier, Creator, Title, Publisher, Publication Year and Resource Type.
- Include **citations and references between datasets and publications** in the metadata descriptions. The DataCite Metadata Schema includes related identifiers (e.g. *isCitedBy*, *isReferencedBy*), which allow link datasets and publications in most repositories.
- Use tools such as [CiteAs](#) for generating dataset citations in widely used standards, including APA, Harvard Reference, Nature and Chicago Style. Many repositories will also include recommended citations in several styles in dataset records.
- Examples of dataset citations:
 - Noll, H.-H., & Weick, S. (2018). European System of Social Indicators: Housing, 1980-2013. (Version 1.0.0) [Dataset]. GESIS Data Archive. <http://doi.org/10.4232/1.13024> (APA 6th ed.).
 - Lopes-Marques, M. GBA3 Sequence alignment data for Phylogenetic and CodeML analysis. (2022). doi:[10.34636/DMPORTAL/DBDXKB](https://doi.org/10.34636/DMPORTAL/DBDXKB) (Nature citation style).

Data usage metrics

Common data usage metrics in repositories

Data usage metrics are standardized through initiatives such as the COUNTER Code of Practice for Research Data, put forward by Make Data Count. Repositories will include several usage metrics for each dataset record and other functionalities, such as usage reports. Common metrics include:

- **Views.** The total number of times that a user looked at any dataset-related page or metadata in the repository.
- **Downloads.** The total number of times a user retrieved dataset files or content.
- **Citations.** The total number of times a dataset was cited using a large database or citation registry, such as DataCite and Crossref.

Tools and initiatives for tracking data usage

Presently, the more relevant tools and initiatives for tracking data usage include:

- [Crossref](#). DOI registration agency providing an open digital infrastructure for connecting and tracking citations between publications and datasets. Crossref records are enriched through Event Data, which captures online citations and mentions, including sources such as Wikipedia, social media and blogs.
- [DataCite](#). DOI registration agency that also maintains a standardized metadata schema for repositories, along with a digital infrastructure for discovery and connectivity, including services such as the PID Graph available through its API.
- [Make Data Count](#). Initiative for promoting and standardizing open data metrics, collaborating with DataCite in the Data Citation Corpus, consisting of more than 10 million data citation records.
- [OpenAIRE Research Graph](#). Aggregates metadata from repositories, publishers and funders, tracking links between datasets, publications and software.

Sources and further information

CESSDA Training Team. (2020). *CESSDA Data Management Expert Guide*. CESSDA ERIC. <https://doi.org/10.5281/zenodo.3820472>

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing Citations and Usage Metrics Together to Make Data Count. *Data Science Journal*, 18. <https://doi.org/10.5334/dsj-2019-009>

Data Citation Synthesis Group. (2014). *Joint Declaration of Data Citation Principles* (M. Martone, Ed.). FORCE11. <https://doi.org/10.25490/a97f-egyk>

DataCite. (n.d.). *Citations and references*. DataCite. <https://support.datacite.org/docs/citations-and-references>

DataCite. (n.d.). *Views and downloads*. DataCite. <https://support.datacite.org/docs/views-and-downloads>

DataCite & Make Data Count. (2025). *Data Citation Corpus Data File* (Version 4.1) [Dataset]. Zenodo. <https://doi.org/10.5281/zenodo.16901115>

ELIXIR. (n.d.). *RDMkit: The ELIXIR Research Data Management toolkit for Life Sciences*. <https://rdmkit.elixir-europe.org/>

Make Data Count. (2024). *COUNTER Code of Practice for Research Data*. Make Data Count Project. <https://coprd.countermetrics.org/>

Puebla, I., & Chodacki, J. (2024). *Make Data Count: Driving metrics for the meaningful evaluation of data*. Zenodo. <https://doi.org/10.5281/zenodo.14261211>

The Dataverse Project. (2025). *Data citation*. <https://dataverse.org/best-practices/data-citation>

Tóth-Czifra, E. (2019). *DARIAH Pathfinder to Data Management Best Practices in the Humanities*. Version 1.0.0. DARIAH Campus [Pathfinder]. <https://hdl.handle.net/21.11159/019595b2-cca1-70ad-b1e3-d088b4409de5>