

PEASIIS Report

Research dataset

Programme for the Evaluation, Accreditation and Monitoring of Health Research Institutes

Sub-Directorate-General for Research Evaluation and Promotion

30 January 2026

Version 5.0

‘Research Dataset’ Working Group

Working group members

Dr Luis García Ortiz. Scientific Director of the IIS IBSAL. Coordinator.

Dr Francisco Tinahones Madueño. Scientific Director of the IIS IBIMA BIONAND Platform.

Dr Michela Bertero. Director of Strategy, IIS IDIBAPS.

Dr Pilar Rico Castro. Head of Department and Open Science Co-ordinator at FECYT.

Dr Remedios Melero Melero. Researcher at the CSIC.

Ms Concepción Campos Asensio. Librarian, Getafe University Hospital. Dr Cristina Lugones Sánchez. Head of the Data Management Unit at IBSAL.

Contact email:

Luis García Ortiz:direccioncientifica@ibsal.es

Francisco Tinahones:fjtinahones@uma.es

Michela Bertero:michela.bertero@idibaps.org

Pilar Rico: pilar.rico@fecyt.es

Remedios Melero:rmelero@iata.csic.es Concepción Campos:ccampos@salud.madrid.org Cristina Lugones: redcap@ibsal.es

Table of Contents

1. Executive summary 1

2. Introduction 1

3. Objective 2

4. Scope 2

5. Definitions 4

6. General protocol for data sets 5

1. Data responsibility 5

2. Data collection phases 6

3. Recommended tools for technical data management 6

4. Procedure for depositing data in open repositories 7

7. Structure and standardisation of the dataset 8

8. Indicators for evaluating the dataset 9

APPENDICES

APPENDIX I. Zenodo as an example of a multidisciplinary repository 12

APPENDIX II. File formats 13

APPENDIX III. Dictionary of variables. Examples 15

ANNEX IV. Controlled vocabularies by discipline 16

APPENDIX V. Metadata structure according to Dublin Core 17

APPENDIX VI. Data anonymisation 18

APPENDIX VII. Components of the README file 21

Bibliography 24

1. r executive summary

The Carlos III Health Institute, within the framework of the IIS Evaluation, Accreditation and Monitoring Programme (PEASIIS), has set up a working group with the aim of establishing a common framework for the collection, structuring, storage, documentation and deposit of research datasets generated in projects within accredited Health Research Institutes (IIS).

This working group was formed on 12 June 2025, when the main lines of work to be covered in the document were defined. Subsequently, the members have worked on drafting and reviewing successive versions of the document, culminating in the final version dated 10 November 2025 (see timeline).

This document proposes a general protocol for the management of research data in accordance with the FAIR principles (Findable, Accessible, Interoperable, Reusable). It sets out a minimum set of information that datasets must include in order to be incorporated into repositories, as well as indicators for evaluating the ISCIII's open access policies and the quality of the deposited datasets.

Timetable

12 June 2025	Establishment of the working group
29 August 2025	Version 1 of the document
24 September 2025	Version 2 of the document
20 October 2025	Version 3 of the document
7 November 2025	Version 4 of the document
10 November 2025	Final version of the document

2. Introduction

The Carlos III Health Institute, within the framework of the IIS Evaluation, Accreditation and Monitoring Programme (PEASIIS), has set up a working group with the aim of establishing a common framework for the collection, structuring, storage and documentation of research datasets generated in projects within accredited Health Research Institutes (IIS).

In the context of the Open Science movement, the ability to share and reuse datasets has become essential. In this regard, the FAIR principles (Findable, Accessible, Interoperable, Reusable) promote data that is locatable, accessible under clear conditions, compatible with other systems and reusable. By depositing data sets in specialised repositories, researchers not only increase the transparency and impact of their work, but also accelerate innovation, enabling others to explore new questions or verify existing results. Therefore, a research data set is not merely a by-product, but a valuable scientific asset in its own right. However, guidelines are needed to standardise the format and the information that must accompany these datasets in order to comply with these principles. Consequently, this document aims to establish a standard for the documentation and structure of the

datasets deposited in a repository. Its aim is to ensure the consistency, quality and traceability of the information, facilitating its understanding, reuse and maintenance over time. By defining the minimum elements that each dataset must include, it promotes more efficient and transparent data management, aligned with best practices in data organisation and governance.

3. Objective

The overall objective of this document is to establish a common framework for the collection, structuring, storage and documentation of research datasets generated in projects within accredited Health Research Institutes (IIS).

This procedure aims to ensure the quality, interoperability, traceability and reusability of the data, facilitating its integration into open-access institutional or thematic repositories wherever possible. It also promotes compliance with ethical and regulatory guidelines, as well as the FAIR principles (Findable, Accessible, Interoperable, Reusable). It will also contribute to transparency, efficiency and collaboration in scientific research, enhancing the reproducibility and impact of results.

Specifically, the following objectives are defined:

O1. To design a common general procedure or protocol for the storage and management of research data across all IISs, including:

- Collection phases: data source, frequency, storage, etc.
- Recommended tools for the technical management of data.
- Procedure(s) for open access deposit, wherever possible, in national or international repositories.

O2. Based on the common framework developed, to establish criteria for the standardisation and definition of one or more quality indicators for research datasets suitable for deposit in open repositories, applicable uniformly across all accredited IISs and other centres within the National Health System (SNS).

4. Scope

This procedure applies to all datasets generated, collected, transformed or used by IISs within the framework of scientific, technical or experimental development research activities, regardless of their field or the methodology used to obtain them. It covers all phases of the data lifecycle: planning, collection, validation, processing, analysis, documentation, curation, preservation, publication, deposit and re-use.

This includes primary and secondary data generated by researchers, technical staff and collaborators participating in research projects or units linked to the IIS. This procedure also applies to data derived from national or international collaborative projects in which the institution acts as a coordinating body, partner or participant. Data generated solely for administrative or internal management purposes not related to research is expressly excluded.

This protocol is aligned with the current regulatory framework on open science, data management and information protection, including:

- Law 14/2011 on Science, Technology and Innovation (amended by Law 17/2022), which promotes open access to public research results
- Regulation (EU) 2016/679 (GDPR) and Organic Law 3/2018 (LOPDGDD), concerning the protection of personal data.
- Decree 24/2021 transposing Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information.
- Regulation (EU) 2021/695 (Horizon Europe), which establishes the obligation to draw up a data management plan (DMP) and to ensure that data is as open as possible and as restricted as necessary.
- Annotated Grant Agreement for EU Funded Programmes 2021–27 (https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/aga_en.pdf)
- Organic Law 2/2023 on the University System, which sets out actions and procedures for the promotion of Open Science and Citizen Science.
- National Open Science Strategy 2023–2027 (ENCA), which establishes the management of research data in accordance with FAIR principles.
- State R&D&I Plan, which requires the submission of Data Management Plans (DMPs).

At institutional level, the ISCIII has developed:

- Technical guidelines for the preparation of DMPs, including tools for the IIS
- The Repisalud repository (https://repisalud.isciii.es/home), which implements open science policies for the publication of data and metadata in accordance with the FAIR principles (https://fairsharing.org/)
- A specific framework for sensitive and health-related data, with anonymisation protocols, access procedures, impact assessments and compatible licences

The standardisation of datasets addresses the need to ensure their quality, traceability, interoperability and reusability, in line with international standards and the requirements of funding bodies. The diversity of formats, structures and metadata schemas hinders the integration, access and reproducibility of scientific results, and compromises the visibility, impact and evaluation of institutional output. This procedure sets out

for the research community and Institutional Information Services (IIS) with the aim of reducing such heterogeneity, facilitating the implementation of open science policies and ensuring the efficient, ethical, responsible and sustainable management of research data. Furthermore, it enables a more comprehensive and objective assessment of compliance with open science requirements.

5. Definitions

- **Research data**: A collection of records, observations, measurements, experimental results, transcripts, codes, images or other digital or analogue elements that serve as the basis for validating, reproducing and drawing scientific conclusions. This includes both primary and secondary data.
- **Dataset**: A collection of information organised in a structured format that enables it to be understood, processed and analysed.
- **Metadata**: Descriptive (details of the data), structural (how the dataset is organised), technical (data type and method of collection), administrative (access to and management of the dataset) and contextual (additional relevant information) information that documents research data, facilitating its identification, interpretation, preservation and reuse.
- **Data Management Plan (DMP)**: A formal document describing how data will be managed during and after the research project, including aspects of storage, preservation, access and legal or ethical protection.
- **FAIR Principles**: A set of international guidelines aimed at ensuring that research data is Findable, Accessible, Interoperable and Reusable. These require storage in trusted repositories, the use of persistent identifiers, open interoperability standards, the application of clear access and reuse licences, and detailed metadata.
- **Sensitive data**: Data which, by its nature, may compromise the privacy, security or rights of individuals, groups or organisations. In health research, this includes clinical, genetic or health data, which are protected by the GDPR and the LOPDGDD.
- **Anonymisation**: A technical process involving the irreversible removal or transformation of personal data so that it is not possible to identify, either directly or indirectly, the data subject to whom it relates. This is a mandatory requirement for the public re-use of sensitive data under European legislation.
- **Pseudonymisation**: A technical process involving the processing of personal data so that it cannot be attributed to a specific data subject without the use of additional information. Such additional information must be stored separately and be subject to technical and organisational measures designed to ensure the identification of the natural person.
- **Repository**: A digital infrastructure, managed by an institution or scientific community, designed for the storage, preservation and open access to

research results, including datasets and their metadata. These repositories may be institutional, thematic or multidisciplinary.

- **Licence**: A legal instrument by which the holder of intellectual property rights permits third parties to make certain uses of the data. In the case of datasets, it is recommended to use open licences such as Creative Commons or Open Data Commons.

6. General protocol for data sets

1. Responsibilities regarding data

In the context of open science and the responsible management of research data, a division of responsibilities between research institutions and scientific teams is necessary. This ensures that data generated during the research process comply with legal and ethical requirements, and are structured, documented and preserved in a way that facilitates their access, interoperability and potential reuse.

Health Research Institute

The Health Research Institute, as the body responsible for the institutional research environment, must establish a regulatory and technical framework to ensure the proper management of research data, particularly data containing sensitive or personal information. Its responsibilities include developing data protection policies in line with current legislation, as well as standards for interoperability and digital preservation. The institution is responsible for providing secure and sustainable infrastructure for data storage and processing. It must also offer specialised training and technical resources to facilitate the proper implementation of data management plans (DMPs) in each project, as well as audit and control mechanisms to ensure the traceability and integrity of the data throughout the project lifecycle.

Research team

The principal investigator assumes direct responsibility for the generation, processing and documentation of data within the project, ensuring compliance with the FAIR principles from the design of the study through to the publication and preservation phases. This involves drawing up a Data Management Plan that addresses not only the ethical and legal aspects of data processing, but also its semantic structuring, its interoperability with international standards and its subsequent availability in appropriate repositories. The research team must ensure that the data are correctly described using standardised metadata, applying anonymisation techniques where appropriate, and facilitating their future reuse. It is also responsible for reporting incidents relating to

data security and coordinating with the institution to ensure its proper preservation or disposal upon completion of the project.

2. data collection phases

a. Data sources

- Internal sources: experimental data, clinical data, surveys, interviews, electronic records, biobanks.
- External sources: public databases, National Health Service (SNS) information systems, international cohorts, scientific literature.
- Classification by data type: structured (databases, tables), unstructured (images, audio, text), sensitive/personal data.

b. Frequency

- Continuous: monitoring systems, sensors, EHRs.
- Periodic: monthly, quarterly, according to the phases described in the project.
- One-off: single collection (e.g. surveys or biological samples).

c. Standardisation of data collection

- Use of standardised electronic forms (REDCap, OpenClinica).
- Automatic field validation and quality control.
- Assignment of unique codes and pseudonymisation at source.

d. Initial storage

- Secure storage on servers that comply with information security standards, personal data protection and digital preservation.
- Encryption of sensitive data in transit and at rest.
- Recording of access logs, audit trails of modifications and scheduled backups.

3. Recommended tools for the technical management of data

Functional area	**Suggested tools**	**Comments**
Data collection/entry	REDCap, LimeSurvey, OpenClinica	Compliant with regulations GDPR/LOPDGDD.
Database management	PostgreSQL, MySQL, MongoDB	Adaptable to different types of data and scalable.
Anonymisation/pseudonymisation	ARX, Amnesia, SDCMicro	Validated tools for clinical and biomedical data.

Functional area	Suggested tools	Comments
Version control and traceability	Git, DVC (Data Version Control)	Recommended for collaborative projects.
Metadata standardisation	Dataverse, CKAN, Invenio	Data repository software that supports schemas such as DataCite, Dublin Core, OpenAIRE and DCAT.
PGD documentation	DMPTool, Argos (OpenAIRE), OpenDMP, DMPonline	Recommended by agencies funding bodies (e.g. the European Commission).

4. Procedure for depositing in open repositories

a. Selecting the appropriate repository

- Data repository search engine: https://www.re3data.org/
- Institutional repositories. Examples: Repisalud, DIGITAL.CSIC.
- Thematic repositories. Examples: EGA (genomic data), Biomodels.
- Multidisciplinary repositories. Examples: Figshare, Zenodo (see Annex I).

b. Prerequisites for deposit

- Minimum data required for the reproducibility of the study, either anonymised or with consent for dissemination.
- Inclusion of standardised metadata (Dublin Core, OpenAIRE or DataCite).
- A clear and, preferably, open licence.
- Persistent persistent generated by the repository or externally assigned (DOI, Handle, ARK, URN, etc.).
- Internal review by an ethics committee or data unit (as appropriate).

c. Deposit process

Preparation of the final dataset:

- Clear structure; open formats are recommended (CSV, JSON, XML, etc.).
- Correct naming of files to be uploaded:
  - File names must be concise, informative and unique. They must include: the file date, project name, test (if there are several), type of data or analysis, identifier of the person responsible for its creation, file version and extension.
  - Avoid spaces and special characters, as these can cause errors. You may use patterns such as the following:
    CamelCase: 20250108HGPTest1PMartinezV3.xlsx

Snake_case: 20250108_HGP_ test1_PMartinez_v3.xlsx

- Associated technical documentation: Variable codes, methodology, analysis scripts (if applicable) and README file.

Upload to repository:

- □ Manual or automated upload (depending on the tool).
- □ Metadata review and validation.

Publication and citation:

- □ Generation of recommended citation.
- □ Notification to the project team and the research institute regarding the project's scientific report

d. Post-deposit maintenance

- □ Updates in the event of a new version of the dataset.
- □ Recording of impact and reuse (altmetrics, citations).
- □ Check the mechanisms for reporting errors or inconsistencies in the selected repository.

7. Structure and standardisation of the data set

Standardisation criteria

Based on the defined common protocol and in accordance with the FAIR principles, the following **mandatory minimum criteria** are established which a dataset must meet to be considered **standardised and suitable for deposit in open repositories**:

a. Technical criteria

- **Open format**: it is recommended that the dataset be in non-proprietary formats (e.g. .csv, .json, .tsv, .xml, .rdf, .txt, .tiff). (see Annex II)
- **Clear structure**: the data must be organised in a hierarchical or file-based structure, with documented relationships.
- **Consistency and validation**: the absence of inconsistent values, coding errors or duplicates must be ensured.
- **Version control**: a record of changes must be maintained.

b. Semantic criteria

- **Variable dictionary**: full definition of each field, unit of measurement, coding and variable codes. (see Annex III)
- **Use of controlled vocabularies**: adoption of standardised terminologies (e.g. ICD-10, MeSH). (see Annex IV)
- **Metadata structure**: minimum requirements as per Dublin Core, OpenAIRE or DataCite. (see Annex V)

c. Legal and ethical criteria

1. **Anonymisation or pseudonymisation** (see Annex VI)
2. **Informed consent,** where necessary**.**
3. **Clear licence terms**: preferably those that maximise re-use. Tools such as Creative Commons can help us select the most appropriate licence depending on the characteristics of the data.
4. **Ethical assessment**: Approval or report from the project's ethics committee, particularly when dealing with sensitive data.

d. Documentation criteria

- **A** completed, up-to-date and accessible **Data Management Plan (DMP)**.
- **README file**: an accompanying document describing the collection process, the structure of the dataset, the members of the working group, the person responsible for data management and the tools used for data processing, amongst other information (see Annex VII).

8. Indicators for the evaluation of the data set

It is proposed that these indicators be assessed once the project's scientific report has been submitted; this report must include information on the deposited datasets, structured as follows:

Field	**Description**
Code and associated project	
Project IP	
Principal Investigator's ORCID	
Data Management Plan (DMP) published	Yes / No
If yes, URL or identifier of the DMP	
Name of the deposited dataset	
Repository used	
Persistent identifier assigned to the dataset	
Date of creation	dd/mm/yyyy
Data type	Image, text, tabular, genomic, etc.
File format(s)	CSV, JSON, TIFF, etc.
Includes a README file and/or a dictionary of variables	Yes / No
Minimum metadata included *(see Annex V)*	Yes / No
Access rights	Public/Restricted/Embargo/Other

If there is more than one data repository for the project, each with a different persistent identifier, it is recommended that a separate form be completed for each one. Minimum metadata is required for registration in most data repositories (e.g. Dryad, Zotero, Figshare, OSF).

Indicators

Indicator	Description	Expected value	Comments
A1. Implementation	% of projects funded by ISCIII with at least one set of data published in a data repository	≥70%	Assessment in the annual report scientific report submitted upon completion of the project
A2. Existence of a PGD	% of ISCIII-funded projects with a data management plan deposited in open-access repositories	≥ 70%	Assessment in the scientific report submitted upon completion of the project
B1. Open file format	% of datasets in reusable open formats (CSV, JSON, etc.)	≥ 70%	Includes the formats recommended in Annex II.
B2. Presence of minimum metadata	% of datasets with metadata comprising, at a minimum, that required by the repository for deposit.	≥ 90%	Automatically verifiable
B3. Defined licence of use	% of datasets with an explicit licence	≥ 90 per cent	Preference for open licences
B4. Documentation of the dataset	% of datasets with variable codes and/or technical documentation	≥ 85 per cent	Attached to the repository
B5. Full anonymisation	% of datasets with minimal risk of re-identification	validation	Requires 100% by the researcher prior to deposit
B6. Identifier persistent identifier **assigned**	% of datasets with persistent identifier assigned	≥ 95%	Assessed in the repository destination

Clarifications:

Indicators A *assess compliance with the data policy; therefore, the denominator for the indicator comprises all ISCIII-funded IIS projects that have submitted their scientific report at the time of assessment.*

Indicators B *assess the quality of the datasets deposited in data repositories. Consequently, the denominator is the number of datasets originating from completed projects funded by the ISCIII that have submitted their scientific report at the time of assessment.*

Applicability and cross-sectional evaluation

- These indicators—all or some of them—should be gradually integrated into the IIS's **data quality and monitoring** system. It would also be necessary to provide parallel training to the IIS and researchers on the collection of information and the calculation of indicators.
- The evaluation of these indicators can be partially automated using tools such as **FAIR Evaluation Services** and **RO-Crate**, as well as through the metadata validators implemented by some repositories.

ANNEX I. Zenodo as an example of a multidisciplinary repository

Zenodo is a multidisciplinary open-access repository, funded by the European Commission through the OpenAIRE initiative and operated by CERN (European Organisation for Nuclear Research), which allows the deposit and preservation of materials resulting from research projects, without restricting their use based on the users' origin. It hosts a wide range of document types (data, publications, software, reports, theses, PGDs, etc.), most of which are open access, but it also allows the option to keep them behind a paywall where necessary, and to make only the metadata available.

Furthermore, each deposit on Zenodo is assigned a unique and persistent DOI (Digital Object Identifier). This ensures that the deposited research can be easily cited and remains accessible in the long term, even if the original location of the file changes.

Zenodo allows the deposit of a wide variety of files, including:

- Data sets: From experimental data to surveys and simulations.
- Scientific publications: Articles, preprints, technical reports and theses.
- Research software: Source code, scripts and applications.
- Presentation materials: Presentations at congresses and conferences.
- Reports: Internal documents, project summaries.
- Images and videos: Audiovisual material associated with research projects.

It also offers integration with GitHub, allowing files to be stored, assigned a DOI and ensuring their preservation. It also supports the use of standardised metadata, which improves the discoverability of repositories through content crawlers and aggregators.

Data communities on Zenodo

It enables the creation of scientific communities, facilitating the joint storage of data and information by groups of researchers, whether by project partners, research groups or centres. In the case of IISs, it allows the creation of an institute-specific community in which to compile all the data and documents that researchers deposit in the repository, making them easily accessible and identifiable for management by the IISs.

Storage and security

It provides secure and reliable storage for research data. Each user can deposit up to 50 GB per dataset (although it is possible to request more space for special requirements). The data is stored on CERN's infrastructure, ensuring high standards of security and redundancy.

ANNEX II: File formats

Type	Recommended formats	Formats not recommended, but generally accepted
Text documents	PDF/A (.pdf) ODT (.odt)	Microsoft Word (.doc) Office Open XML (.docx) Rich Text File (.rtf) PDF other than PDF/A (.pdf)
Plain text	Unicode text (.txt)	Non-Unicode text (.txt)
Compressed compressed	ZIP (.zip) 7-Zip (.7z) GNU Tar (tar.gz/.tgz) GNU Gzip (.gz)	WinRAR (.rar)
Markup language	XML (.xml) HTML (.html) Related files: .css, .xslt, .js, .es Markdown (.md)	SGML (.sgml)
Text-based programming	NetCDF Text-Fabric Python R	MATLAB
Spreadsheets	ODS (.ods) CSV (.csv)	Microsoft Excel (.xls) Office Open XML Workbook (.xlsx) PDF/A (.pdf)
Databases	SQL (.sql) SIARD (.siard) CSV (.csv)	Microsoft Access (.mdb, .accdb) dBase (.dbf) HDF5 (.hdf5, .he5, .h5)
Statistical data	General: .dat/ .txt JASP (.csv/.html) R (.RData/.Rds)	SPSS (.sav) STATA (.dta/.DO) SAS (.7dat; .sd2; .tpt) JASP (.jasp)
Images (bitmap)	JPEG (.jpg, .jpeg) TIFF (.tif, .tiff) PNG (.png) JPEG 2000 (.jp2) DICOM (.dcm)	
Vector images	SVG (.svg)	Adobe Illustrator (.ai) EPS (.eps) WMF/EMF (.wmf, .emf) CDR (.cdr)
Video	MXF (.mxf) Matroska (.mkv)	MPEG-4 (.mp4, .m4a, .m4v, …) MPEG-2 (.mpg, .mpeg, .m2v, .mpg2, …) AVI (.avi), QuickTime (.mov, .qt)
Audio	BWF (.bwf) MXF (.mxf) Matroska (.mka)	WAVE (.wav) MP3 (.mp3) AAC (.aac, .m4a)

	FLAC (.flac) OPUS	AIFF (.aif, .aiff) OGG (.ogg)
Computer-Aided Design (CAD)	AutoCAD DXF version R12 (ASCII) (.dxf) SVG (.svg)	AutoCAD DXF, versions other than R12 (ASCII) (.dxf) DWG (.dwg) DGN (.dgn)
Geographic Information Systems (GIS)	GML (.gml) MIF/MID (.mif/.mid) GeoJSON (.json) GeoPackage (.gpkk)	Esri Shapefiles (.shp & related files) MapInfo (.tab & related files) KML (.kml, .kmz) Esri Geodatabase (.gdb) Project files / Workspaces (.mxd, .wor, .qgs)
Georeferenced images	GeoTIFF (.tif, .tiff)	TIFF World File (.tfw & .tif, possibly with additional files) JPEG World File (.jgw & .jpg, possibly with additional files) ERDAS IMAGINE File Format (.img)
GIS Raster	ASCII GRID (.asc, .txt)	Esri GRID (.grd & related files) Surfer Grid (.grd; .srf) ERDAS IMAGINE File Format (.img)
3D	WaveFront Object (.obj) Polygon file format (.ply) X3D (.x3d) glTF 2.0 (.gltf; .glb) COLLADA (.dae) LASer (.las, .laz) IFC (.ifc)	Autodesk FBX (.bx) Blender (.blend) glTF 1.0 (.gltf; .glb) 3D PDF (.pdf) Google Draco (.drc) Artec (.a3d) Agisoft Metashape (.psx & .psz) STL (.stl) VRML (.wrl, .wrz, .vrml)
RDF	RDF/XML Trig (.trig) Turtle (.ttl) NTriples (.nt) JSON-LD	
Computer-Assisted Qualitative Data Analysis (CAQDAS)	REFI-QDA (Qualitative Data Analysis)	
Proteomics data	MSF (.msf)	
Flow cytometry standard flow cytometry	FCS (.fcs)	

Table compiled from the Data Archiving and Networked Services (DANS) format table and the Research Data Management guidelines.

ANNEX III: Dictionary of variables. Examples

Variable: occupation	
Code	**Label**
1	Student
2	Employee
3	Self-employed
4	Unemployed
5	Retired

Variable: date_infecc	
Format	**Description**
YYYY/MM/DD	Year, month and day on which the person first developed symptoms or was diagnosed with the infection.

Variable: BMI	
Format	**Description**
Two-digit decimal number (e.g. 23.45)	A value that relates weight to height. It is calculated as weight (kg) divided by height (m) squared.

ANNEX IV: Controlled vocabularies by discipline

Discipline	Name of the controlled	Example of accepted term	Example of an accepted URL
Multidisciplinary	Library of Congress Subject Headings	Psychology	http://id.loc.gov/authorities/subjects/sh85108459
Multidisciplinary	DBpedia	Artificial intelligence	https://dbpedia.org/page/Category:Artificial_intelligence
Food	FoodOn Food Ontology	maize flour	http://purl.obolibrary.org/obo/FOODON_03540073
Arts and Humanities	Art and Architecture Thesaurus	aqueducts	http://vocab.getty.edu/page/aat/300006165
Arts and Humanities	Art & Architecture Thesaurus	aqueduct	https://www.aatespanol.cl/terminos/300006165
Arts and Humanities	Getty Thesaurus of Geographic Names	Iberian Peninsula	http://vocab.getty.edu/page/tgn/7016676
Arts and Humanities	Thesauri and Dictionaries of Cultural Heritage of Spain	Amphora	http://tesauros.mecd.es/tesauros/ceramica/1010577
Arts and Humanities	Pleiades	Tarraconensis	https://pleiades.stoa.org/places/981551
Arts and Humanities	Period	Imperial Roman	http://n2t.net/ark:/99152/p0qhb66dt5q
Agricultural Sciences	AGROVOC Multilingual Thesaurus	feed additives	http://aims.fao.org/aos/agrovoc/c_2827
Life Sciences Life	Medical Subject Headings (MeSH)	Biotechnology	https://www.ncbi.nlm.nih.gov/mesh/68001709
Life Life	International Classification of Diseases (ICD)	Pneumonia	http://id.who.int/icd/entity/142052508
Life Life	Gene Ontology	Lysosome	https://amigo.geneontology.org/amigo/term/GO:0005764
Life Sciences	Descriptors in health sciences (bvsalud)	Oxygen Radical Scavenging Capacity Oxygen	https://decs.bvsalud.org/es/ths/resource/?id=56697&filter=ths_termall&q=Capacity%20for%20Absorption%20of%20O%20radicals%20x%C3%ADgeno x%C3%ADgeno
Life Life	The thesaurus on Prehistory	Anthracology	https://www.archeoindex.org/776

Table compiled from the table of controlled vocabularies of the Research Data Management system.

ANNEX V: Metadata structure according to the Dublin Core

Element	Description	Example
Title	Resource title	"Data on the prevalence of diabetes in adults – Spain 2023"
Creator	Author or organisation responsible for the creation	"X Institute for Health Research"
Subject	Subject or keywords (preferably using controlled vocabularies)	"Type 2 diabetes; epidemiology; public health"
Description	Brief description of the content of the dataset	"This dataset contains anonymised data from diabetes prevalence data collected between January and December 2023 in primary care centres ." "IIS"
Publisher	Organisation responsible for publication of the resource	"Department of Statistics, University Y" "15
Contributor	Other relevant contributors relevant	February 2024"
Date	Date associated with the resource (publication, creation, etc.)	
Type	Type of resource according to controlled vocabulary (DCMI Type Vocabulary)	"Data set"
Format	File format, preferably MIME type	"text/csv"
Identifier	Unique identifier for the resource (preferably DOI, Handle, URI)	"https://doi.org/10.5281/zenodo.9999999"
Source	Resource from which it is derived, if applicable	"2020 National Health Survey"
Language	Language of the (ISO 639-1)	"es"
Relation	Related resources (other datasets, articles, etc.)	"Related to: https://doi.org/10.1016/j.pubmed.2024.00123"
Coverage	Scope and/or time frame	"Spain, 2023"
Rights	Terms of use and licence	"CC BY 4.0"

ANNEX VI. Anonymisation of the data

The dataset must not include direct identifiers, i.e. information that is sufficient in itself to identify the patient (name, address, postcode, telephone number, photographs or videos). The inclusion of indirect identifiers—those which, when combined, could lead to the identification of the individual—is permitted, but only to the minimum extent possible; some authors recommend a maximum of three. Examples include affiliation, occupation, gender or anthropometric measurements.

Anonymisation generally involves the removal of direct identifiers and the modification of indirect identifiers. Pseudonymisation, or coding, refers to the replacement of identifying data with fictitious values that bear no relation to the originals. Pseudonyms may be irreversible when the original values are properly deleted and the pseudonymisation is carried out in a non-repeatable manner.

Any variable that is not required in the dataset should be removed to minimise the dataset. The appropriate selection of variables would be carried out as shown in the figure:

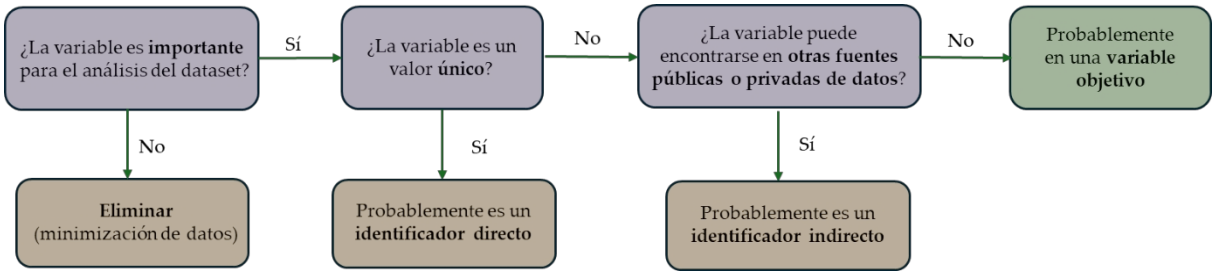


Figure based on the AEPD's basic guide to anonymisation

There are many anonymisation techniques that can be used to modify the variables in the dataset:

Technique	Description	Examples
Aggregation	Converting data into ranges	Use age ranges instead of date of or exact age (e.g. 10–20)
Combining	Merging two variables or more to create a new summary variable	Replace exact dates with time intervals between events (e.g. the duration of time spent on a survey)

Handling outliers	Grouping a continuous variable into broad ranges to avoid identifying outliers	A 72-year-old person is included in 'people aged 70' or 'older people'.
Generalisation	Adjust the precision of the data	Specific occupational role → occupation or area of specialisation City → region
Limiting the number of indirect identifiers to a maximum of three	Avoid using small subgroups (samples of fewer than 5).	No: Age + gender + occupation + number of children Yes: Age (aggregated) + gender + occupation
Remove/Delete	Include only the data necessary for the study to be reproducible	Remove identification numbers, email addresses, postal addresses, etc.

Table based on 'Best Practices: Sharing Human Subjects Data' and the AEPD's Basic Guide to Anonymisation.

k-a Anonymity

To ensure that the identification risk threshold has not been exceeded, following the application of the previous techniques, measures such as k-anonymity (or similar techniques such as l-diversity and t-closeness) may be used.

This is a data protection technique designed to ensure that an individual cannot be uniquely identified within an anonymised dataset. A dataset is k-anonymous if, for every combination of indirect identifiers, there are at least k-1 additional records that are identical in those attributes. This means that the probability of identifying a person from that data is at most 1/k, thereby reducing the risk of re-identification.

For example, if in a dataset every combination of age, postcode and gender values appears in at least five different records, it would be termed 5-anonymous, and would prevent the re-identification of a specific individual using that data.

Tools for anonymising the dataset

Amnesia: Used to the pseudo-anonymisation of sets of tabulated . https://amnesia.openaire.eu/

Pydeface and mri_deface (FreeSurfer): Used to remove or blur facial features in fMRI data https://pypi.org/project/pydeface/

ARX De-Identifier: https://arx.deidentifier.org/

Tool for creating unique identifiers for study participants: https://fitbir.nih.gov/content/global-unique-identifier

Tools for anonymisation anonymisation: https://blog.gramener.com/10-best-data-anonymization-tools-and-techniques-to-protect-sensitive-information/

Applications for de-identification in research (Johns Hopkins University): https://dataservices.library.jhu.edu/resources/applications-to-assist-in-de-identification-of-human-subjects-research-data/#image

ANNEX VII. Components of the README file

README files contain information that can identify and explain your dataset to users, regardless of whether it is accessed or not.

What should be included:

- A summary of the project, experiment or intervention on which this dataset is based.
- A description of the structure and content of the file.
- Definitions of all variables, abbreviations, missing data codes and units.
- Links to other publicly accessible locations of the data.
- Other sources, if any, from which the data were derived.
- Any other information that may influence the reuse or replication of the collected data.
- It is recommended that this be written in English, but it may also be included in other languages, such as Spanish, using a different file name (for example, README-es.txt).

File structure

GENERAL INFORMATION

1. Dataset Title
   Briefly summarise the content of the dataset, providing context for the procedures and results obtained.
2. Contact (Contact person)
   Contact researcher Name:
   Affiliation:
   Email:
   ORCID:
3. Organisation.
   Include the name and address of the IIS or centre
4. Project description
   Provide a brief summary of the project associated with the dataset, including its main objective and purpose.
5. Data description and file structure (Data description) *Describe how the data is structured and how another researcher might use it. Users of the dataset may not be familiar with field-specific terminology, so it is advisable to explain this briefly. Describe the relationships between the data files, the codes for missing data and any other abbreviations used.*
6. *Notes*
7. *Date the files were deposited (Publication Date): YYYY-mm-dd*

8. *Date the files were created (Created Date): YYYY-mm-dd*
9. *Language*

AUTHORSHIP

1. Authors (*list lead author(s) and other researchers*)

Principal Investigator:

Name:
Surname:
Telephone (optional):
Email:
ORCID:

Co-author 1:

First name:
Surname:
Telephone (optional):
Email:
ORCID:

METHODOLOGY

1. Methodology (Methods)

   Clearly describe how the data were collected, processed and organised. Indicate the techniques used (instruments, surveys, scripts, cleaning criteria) and any relevant decisions that affect the interpretation.

2. Software

   Describe any scripts, code or programmes (e.g. R, Python, Mathematica, MatLab), as well as the versions of the software used to run these files. If the relationship between the script and the attached files is not obvious, it is advisable to provide information on the workflow used to run them.

KEYWORDS

1. Keywords

   A list of keywords describing the dataset, indicating whether they come from a controlled or free vocabulary.

FUNDING INFORMATION AND PROJECT IDENTIFIERS

1. Project information (Grants)

   Include: Name, funding body, identification code, etc.

RELATED PUBLICATIONS

1. Related publication (Publication)

   Provide the full bibliographic reference, where applicable, including the DOI, URL or link to the published version

2. Related dataset (Links to other research outputs and datasets)

Links to publicly accessible locations of related datasets

FILES

1. Files

   Include a list of all files contained in the dataset. Use the following structure for each file included.

   File 1

   - *Name: [Full file name, e.g. sleep.csv]*
   - *Description: [Brief summary of the content, e.g. dataset with sleep variables and sociodemographic data]*
   - *SourceType: [the type of data collection; choose from: Observed, Experiment, Computational]*
   - *DataSource: [Explain how the data were obtained, e.g. Data collected from self-administered questionnaires]*
   - *FileFormat: [File format, e.g. CSV]*
   - *Variables: [variable, description, type; variable, description, type; e.g. sleephr, average hours slept; awakenings, average number of night-time awakenings]*
   - *Language: Spanish*

TERMS OF USE AND PRIVACY

1. Licences (Licence)

   Indicate the licences under which the dataset is distributed, specifying permissions, restrictions and conditions of use.

2. Data Processing

   Describe how personal data is protected, stating whether it has been anonymised, whether identifiers have been removed, or whether other privacy measures are in place.

OTHER

1. Data Codebook

   A detailed list of all variables present in the dataset. For each variable, state the name, data type, description, possible values and units, where applicable. If there is an attached file containing this information, state its name and how to access it.

2. How to Cite the Dataset

References

Broman KW, Woo KH. Data organisation in spreadsheets. The American Statistician. 2 January 2018;72(1):2–10.

Madroño Consortium. *Readme.txt template*. e-cienciaDatos. 2025. [cited 12 September 2025] Available from: https://edatos.consorciomadrono.es/readme.xhtml

Corcho O, Sánchez González L, Simperl E; European Data Portal; Publications Office of the European Union. Report on Data Homogenisation for High-value Datasets [Internet]. Europa: Publications Office of the European Union; 5 December 2023. Report No.: OA-09-23-557-EN-N. ISBN: 978-92-78-43830-2. DOI: 10.2830/446704. Available at: https://data.europa.eu/en/publications/reports/report-data-homogenisation-high-value-datasets

Cunha-Oliveira T, Ioannidis JP, Oliveira PJ. Best practices for data management and sharing in experimental biomedical research. Physiological Reviews. 1 July 2024;104(3):1387–408.

Dryad Digital Repository. Best practices for creating reusable data publications [Internet]. Dryad; [cited 29 August 2025]. Available at: https://datadryad.org/best_practices

DCMI Usage Board. Dublin Core™ Metadata Element Set, Version 1.1: Reference Description [Internet]. Dublin Core Metadata Initiative; 14 June 2012 [cited 21 October 2025]. Available from: https://www.dublincore.org/specifications/dublin-core/dces/

Extract from the European Council Conclusions (2014/C 240/01.0001.01.ENG). Official Journal of the European Union. 24 July 2014; C-240:13. [Internet]. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.C_.2014.240.01.0001.01.ENG

FAIRsharing. Available at: https://fairsharing.org/ [cited 15 October 2025].

Basic guide to anonymisation. Translated by the Spanish Data Protection Agency (AEPD) from the guide produced by the Personal Data Protection Commission (PDPC) of Singapore. Published October 2022. Madrid: AEPD; 2022. PDF [Internet]. Available at: https://www.aepd.es/documento/guia-basica-anonimizacion.pdf.

Carlos III Health Institute. Security aspects of handling sensitive data. Madrid: Carlos III Health Institute; 2023 [cited 15 October 2025]. Available at: https://impact.isciii.es/wp-content/uploads/2023/07/Aspectos-de-seguridad-en-el-manejo-de-datos-sensibles.pdf

Carlos III Health Institute. WG2 Report: Data Management Plan [Internet]. Ministry of Science and Innovation / ISCIII; 2023 [cited 15 October 2025]. Available at: https://www.isciii.es/documents/d/guest/informegt2_plangestiondatos_feb2023

Publications Office of the European Union. Metadata quality. In: *Data-Provider Manual*. [Internet]. [cited 29 August 2025]. Available from: https://dataeuropa.gitlab.io/data-provider-manual/metadata-quality/ GitLab

Publications Services, Stanford University Libraries. Data best practices and case studies [Internet]. Stanford (CA): Stanford University Libraries; [cited 29 August 2025]. Last updated 13 September 2023. Available at: https://guides.library.stanford.edu/data-best-practices