Project Number: 101186558
Project Acronym: CBeRa
Project Title: Strategic Integration of Complex Networks and Systems for Advancing Biomedical Research
Call: HORIZON-WIDERA-2023-TALENTS-01
Deliverable Number: D1
Nature: DMP
Dissemination level: Public
Due Date: 31 August 2025

D1.1 Data Management Plan

Plan describing the types of data generated and their management

Version	Date	Status
0.1	02.08.2025	Initial draft by Diogo Silva
0.2	06.08.2025	Reviewed by Felipe Xavier Costa
1.0	14.08.2025	Feedback session with Ethics Advisor Francis P. Crawley
1.1	17.08.2025	New draft by Diogo Silva
1.2	27.08.2025	Reviewed by Luís M. Rocha
2.0	28.08.2025	Final version by Diogo Silva

Table of Contents

Executive Summary

This Data Management Plan (DMP) outlines the principles, strategies, and initial measures adopted by the CBeRa project (Grant Agreement 101186558) for managing research and non-research data in accordance with the European Commission's requirements for Horizon Europe projects.

The project produces two main types of data:

- Scientific data, including multi-omics datasets, biomedical images, computational models, code, and social media interaction data. These datasets are crucial for achieving the scientific goals of CBeRa and will be handled according to community standards, FAIR principles, and Open Science mandates.
- Non-scientific data, including administrative, governance, training, dissemination, and outreach outputs. These datasets support accountability, institutional development, and impact beyond the scientific community.

All datasets will be managed following FAIR principles (Findable, Accessible, Interoperable, Reusable), GDPR requirements, and Open Research Data Pilot (ORDP) participation. A preliminary repository mapping has been established (as shown in Table 2) for both scientific and non-scientific datasets, with specific assignments and workflows to be finalized at M18 once the research team and Data Steward are in place.

The DMP is a living document and will be updated at M18 (D10.3) and M60 (D11.3). The Ethics Advisor plays a key role in ensuring compliance with ethical and legal standards, including adherence to UNESCO guidance on data management during crises.

This initial version (M6) provides a structured framework that demonstrates awareness of all key aspects of data management, while recognizing that additional details will be added in future updates as the project progresses.

1. Introduction

This Data Management Plan (DMP) was prepared for the ERA Chair project CBeRa, funded under the Horizon Europe WIDERA TALENTS program. It offers a clear roadmap for how scientific and non-scientific data will be collected, stored, shared, and preserved throughout the project. The DMP ensures compliance with FAIR principles, Open Science policies, and ethical standards, while also reflecting the commitments outlined in the grant agreement.

CBeRa aims to improve the excellence and competitiveness of the Católica Biomedical Research Centre (CBR) at Universidade Católica Portuguesa (UCP) by strategically integrating Complex Networks and Systems (CNS) science into its core research and education programs. The project seeks to position CBR as an international reference in holistic biomedical research and education, expand its global collaborations, and implement sustainable institutional reforms in line with European Research Area priorities and Open Science policies.

2. Dual-Layer Approach to Data

The CBeRa project will generate a wide range of data, which can be categorized into two main layers:

- Scientific data: Biomedical datasets, computational models, CNS simulations, and outputs from collaborative research. These datasets are essential to the project's scientific goals and require detailed management to ensure reproducibility, long-term preservation, and adherence to FAIR principles.
- Non-scientific data: Administrative, educational, and dissemination-related information such as reports, training materials, stakeholder lists, and communication outputs. These do not need the same level of technical detail as scientific data but still must be managed systematically to ensure accountability, transparency, and accessibility when necessary.

3. Scientific Data Management

3.1 Data Summary

3.1.1 Purpose of the data collection/generation and relation to project objectives

The CBeRa project aims to establish a leading research group in complex systems and network science applied to health and biomedicine. Data will be collected and created to support research on health-related behaviors, knowledge networks, medical informatics, and biological network modeling. Once formed, the research team is expected to conduct studies involving about 3–5 primary datasets annually, using both experimental and computational sources. These may include pilot

studies on network models of disease spread, longitudinal behavioral datasets, and high-resolution biological network mappings.

3.1.2 Types and formats of data

Data types include:

- Textual data from social media and forums (JSON, CSV): These datasets capture public discourse and health-related narratives, enabling computational analysis of behavioral trends, misinformation patterns, and patient support networks. Such data may be collected from open online platforms, respecting ethical and legal guidelines.
- Clinical and biomedical metadata (CSV, XML): Structured data on patient demographics, clinical procedures, and laboratory results. This metadata can support modelling of disease trajectories, treatment outcomes, and healthcare service utilization patterns.
- Multiomics datasets (CSV, HDF5): Comprehensive molecular-level data such as genomics, transcriptomics, proteomics, and metabolomics. These datasets are essential for understanding complex biological processes and integrating multiple layers of biological information.
- Scientific publication metadata (RIS, BibTeX, XML): Information describing published research, including authorship, citations, keywords, and abstracts. This data supports bibliometric analysis and the mapping of knowledge networks within biomedical research.
- Graph/network data (GraphML, edge lists): Data representing complex relationships between entities, such as molecular interactions, patient referral networks, or conceptual linkages in literature. This format is central to network science approaches within the project.
- Code and pipelines (Python, Jupyter Notebooks, GitHub repos): Reproducible computational workflows and analysis scripts used to process, analyze, and visualize project data. Sharing code promotes transparency and enables other researchers to replicate and build upon the work.

In addition to the listed data types, genomic variant datasets (VCF format) and imaging datasets (DICOM format) may be included, depending on collaborations established in the first project year.

3.1.3 Re-use of existing data

CBeRa will reuse data from public repositories (e.g., PubMed, ClinicalTrials.gov, DisGeNET) and previous NIH-funded datasets. Proper attribution and licensing conditions will be followed.

3.1.4 Origin of the data

Initial datasets come from NIH projects managed by ERA Chair Holder Prof. Luís M. Rocha, including social media mining, biomedical databases, and EHR pipelines. The CBeRa team will generate new data after its formation.

3.1.5 Expected size of the data

Based on similar NIH and EU biomedical projects, anticipated annual volumes are estimated as shown in Table 1.

Table 1. Projected annual data volumes for CBeRa datasets, presented by category and expressed in approximate storage requirements.

Dataset type	Projected annual volume
Social media text	5 – 20 GB
Clinical metadata	5 – 100 GB
Multiomics	50 – 500 GB
Networks	1 – 60 GB
Code	< 1 GB

These figures will be refined as the research team is established and workflows are defined. Based on similar NIH and EU projects, annual data generation could range from 50GB to 2TB, depending on whether large-scale omics or imaging datasets are included.

3.1.6 Data utility

These datasets will prove valuable to researchers in systems biology, digital epidemiology, health informatics, and computational social science. Aggregated and anonymized datasets will promote reproducibility and facilitate interdisciplinary use.

3.2 FAIR Data

3.2.1 Making data findable, including provisions for metadata

Data will be documented using standard metadata schemas (e.g., Dublin Core, schema.org). Persistent identifiers (DOIs) will be assigned via Zenodo or institutional repositories. Standard naming conventions and explicit versioning will be applied. For each dataset type, persistent identifiers will be assigned through appropriate repositories. Social media datasets will be archived in Zenodo or an institutional repository with a DOI; clinical and biomedical metadata will receive European Genome-phenome Archive accession numbers; multiomics datasets will

Funded by the European Union

be deposited with BioStudies or ArrayExpress identifiers; and all code repositories will be linked to Zenodo DOIs with Git commit hashes to ensure reproducibility. Metadata will follow recognized community standards per category: HL7 FHIR and ICD-10 codes for clinical data; MIAME and MINSEQE standards for omics datasets; Dublin Core and CrossRef schemas for publication metadata; and NDEx-compliant metadata schemas for network datasets. Metadata will likely follow Health Level Seven (HL7) FHIR standards for clinical data, and BioSchemas for life science datasets, once applicable workflows are confirmed.

3.2.2 Making data openly accessible

Open data will be deposited in repositories such as Zenodo, OpenAIRE, GitHub, or institutional repositories. Code will be published under MIT or GPL licenses. Sensitive data will be shared only in anonymized or aggregated form. Access and sharing policies will be tailored to each dataset type. Aggregated social media data will be openly shared, while raw data will only be accessible under appropriate licensing and ethical clearances. Clinical datasets will be available through access-controlled repositories and will require formal Data Transfer Agreements. Multiomics datasets will be publicly available via European Bioinformatics Institute repositories after any applicable embargo periods. Code and computational pipelines will be openly shared on GitHub and archived in Zenodo at the time of publication. For large datasets (>1TB), we may use the European Genome-phenome Archive (EGA) or ELIXIR node repositories, depending on data type and access restrictions.

3.2.3 Making data interoperable

Data formats will follow open standards: GraphML for networks, CSV/XML for tabular data, and JSON for APIs. Controlled vocabularies and ontologies (e.g., MeSH, ICD-10) will be used where relevant. Data quality will be maintained by following community standards, systematic documentation, and internal peer review before deposit.

3.2.4 Increase data re-use

All reusable data will be shared under clear licenses (e.g., CC BY 4.0). Embargoes may apply until publication or patent submission. Documentation will accompany each dataset, including README files, data dictionaries, and pipeline descriptions.

Where relevant, datasets will be prepared for integration into open knowledge graphs (e.g., OpenAIRE Research Graph) to maximize discoverability and reuse.

3.3 Allocation of resources

3.3.1 Costs and funding

Costs for data storage, management, and sharing will be covered by the CBeRa budget. Resources include institutional repositories, cloud services (if needed), and personnel effort.

3.3.2 Responsibilities

Until a group is fully formed, Prof. Luís M. Rocha and the project manager will oversee data stewardship. A dedicated data steward will be appointed once the ERA Chair holder assembles the team.

3.3.3 Data management expertise

Completed once Data Steward is appointed; will include qualifications and institutional support structures. The Data Steward is expected to have formal training in FAIR data management and biomedical informatics and will likely coordinate with UCP's IT and Ethics Committees.

3.4. Data Security

Data will be stored on secure, access-controlled servers at host institutions. Backups and disaster recovery protocols will be in place. Sensitive data will follow GDPR-compliant practices, including encryption and access logging.

We anticipate implementing two-factor authentication for all servers and encrypted cloud backups for critical datasets, with redundancy across at least two geographic locations.

3.5. Ethical Aspects

Sensitive data (e.g., EHRs, social media content) will be anonymized or obfuscated. All human subject data use will be governed by informed consent and ethics approvals. The CBeRa project will follow the host institution's ethics policy and EU GDPR.

GDPR-compliant data protection measures will be tailored to each sensitive dataset. Clinical data will undergo pseudonymisation, encryption at rest, and access logging, with identifiable information removed before analysis. Social media datasets will include only publicly available posts with usernames and identifying details stripped. Multiomics data will be stored in controlled-access repositories with no personal identifiers, and any sharing will be governed by signed agreements specifying permitted uses.

Consent procedures may include options for future use of anonymised data in other ethically approved research projects, aligning with GDPR and UNESCO Open Science recommendations.

3.6. Other Policies / Institutional Guidelines

CBeRa will align with EU and national policies, including Horizon Europe guidelines, and local institutional rules on data retention and access. Updates to this DMP will be made during project evaluations or whenever significant changes occur.

Long-term preservation will be ensured through institutional repositories with a minimum retention period of ten years, with mirrored copies stored in domain-specific repositories where available. Critical datasets will be backed up in at least two geographically distinct locations, and core metadata will remain publicly accessible indefinitely to support continued discovery and citation.

3.7. Data Management in Times of Crisis (UNESCO 2025 Guidance)

In alignment with UNESCO (2025) guidance on developing data policies for times of crisis facilitated by Open Science, the CBeRa project will adopt the following measures to ensure that data management practices remain robust, ethical, and responsive during emergencies such as pandemics, natural disasters, or urgent public health threats:

• Prioritize rapid, ethically compliant data sharing to aid response efforts, while safeguarding privacy and confidentiality.

• Implement fast-track anonymization and pre-publication release of non-sensitive data in trusted repositories.

• Ensure interoperability of data formats to facilitate immediate global data exchange with other research and health agencies.

• Maintain secure, redundant storage systems to protect data integrity in case of localized infrastructure failures.

• Apply proportionate and context-sensitive ethical safeguards for the use of sensitive and personal data.

• Engage with relevant authorities and ethics committees to expedite approvals required for data use in crises.

• Document all crisis-related data handling decisions for transparency and accountability.

This section was prepared following the suggestion of the project's Ethics Advisor, Francis P. Crawley, who emphasized the importance of aligning CBeRa's data management practices with UNESCO (2025) recommendations on Open Science in times of crisis.

4. Non-Scientific Data Management

Non-scientific data generated by CBeRa include administrative, educational, communication, and networking-related outputs. While these are not research datasets, they are essential to ensure accountability, transparency, capacity building, and impact. The following subsections mirror the structure applied to scientific data management to ensure consistency and completeness.

4.1 Data Summary

4.1.1 Purpose of the data collection/generation and relation to project objectives

The collection and generation of non-scientific data directly support the objectives of the CBeRa project. These datasets document governance processes, enable effective communication and dissemination, provide training materials for capacity building, and record collaboration activities. They ensure transparency and compliance with Horizon Europe's requirements.

4.1.2 Types and formats of data

Non-scientific data include:

- Administrative documents such as deliverables, sustainability plans, and internal reports (PDF, DOCX, XLSX);
- Educational and training materials such as syllabi, lecture slides, video recordings, and attendance records (DOCX, PPTX, MP4, CSV);
- Communication and dissemination outputs such as newsletters, website content, conference programs, and social media exports (HTML, PDF, PNG, CSV);
- Networking and collaboration data, including stakeholder lists, meeting minutes, and expert visit reports (CSV, DOCX, XLSX).
- Multimedia content (videos, infographics, recorded lectures) may contribute significantly to non-scientific data volume, ranging from 5–20 GB annually, depending on dissemination activities.

4.1.3 Re-use of existing data

Some non-scientific datasets will re-use existing templates, training modules, or communication materials from UCP or EU-funded projects. For example, prior CNS educational content and established dissemination practices may be adapted for CBeRa's purposes.

4.1.4 Origin of the data

Non-scientific data will be primarily generated by the CBeRa management team, particularly, administrative and governance documents, training and educational resources, and dissemination content. Networking records will arise from stakeholder engagement activities.

4.1.5 Expected size of the data

The overall size of non-scientific data is expected to be modest relative to scientific datasets, estimated between 10 GB and 50 GB across the project's lifetime, depending on the intensity of communication activities and the number of training events recorded.

4.1.6 Data utility

Non-scientific data will serve several essential functions. Administrative records will support accountability and auditability of project progress; training materials will strengthen institutional capacity and provide reusable resources for future courses; communication outputs will increase CBR's visibility; and networking data will inform long-term collaboration strategies.

4.2 FAIR Data

4.2.1 Making data findable, including provisions for metadata

Non-scientific datasets will be accompanied by metadata using Dublin Core or similar minimal standards. Deliverables will be deposited in CORDIS with persistent identifiers. Training materials, newsletters, and outreach resources will be archived in Zenodo or UCP's repository, ensuring discoverability. Public deliverables will also be available on the project's website.

4.2.2 Making data openly accessible

Public-facing datasets (deliverables, newsletters, training videos, website content) will be openly accessible under Creative Commons licenses where possible. Confidential datasets (e.g., stakeholder lists, internal governance documents) will remain restricted and accessible only to authorized personnel.

4.2.3 Making data interoperable

Non-scientific data will be stored in open or widely adopted formats (PDF/A, CSV, MP4, HTML). Consistent use of metadata and standardized file naming conventions will support interoperability and facilitate reuse.

4.2.4 Increase data re-use

Training resources, outreach materials, and public reports will be openly licensed to facilitate re-use by other institutions and stakeholders. Administrative records

may be used as models for future projects within UCP, but will not be openly shared due to confidentiality requirements.

4.3 Allocation of resources

4.3.1 Costs and funding

Costs associated with managing non-scientific data are covered within the CBeRa budget. These include institutional server space, website maintenance, and staff time dedicated to communications and data curation.

4.3.2 Responsibilities

The project manager is primarily responsible for non-scientific data management. They will work in coordination with the ERA Chair holder and Data Steward to ensure compliance with institutional policies and FAIR principles.

4.3.3 Data management expertise

Training will be provided to staff handling non-scientific data, focusing on FAIR data practices, long-term preservation of communication outputs, and GDPR compliance in the handling of stakeholder information.

4.4 Data Security

Non-scientific datasets will be stored securely on UCP servers with access rights defined by role. Confidential datasets such as stakeholder lists will be protected by encryption and access logging. Regular backups will be performed in accordance with institutional IT policies.

4.5 Ethical Aspects

The handling of non-scientific datasets, particularly stakeholder and participant data, will comply with GDPR and EU ethical standards. Informed consent will be obtained where necessary, and only the minimum personal information required will be collected.

4.6 Other Policies and Long-term Preservation

Non-scientific datasets, including deliverables, training materials, and outreach resources, will be preserved for at least ten years in institutional repositories. Where appropriate, mirrored copies will be stored in Zenodo or similar platforms to guarantee long-term accessibility and discoverability.

5. FAIR Principles and Open Access

This section consolidates the FAIR data management practices across both scientific and non-scientific datasets generated by the CBeRa project. Sections 3.2 and 4.2 already provide detailed narrative descriptions of how FAIR principles apply to each category of data. Here, a preliminary mapping of data sources has been

Funded by the European Union

summarized in Table 2. It offers a structured overview of data types, intended repositories, metadata standards, identifiers, access conditions, and licensing arrangements.

This mapping is provisional and will be updated in the M18 version of the DMP once the research team is in place and dataset generation has begun. In the meantime, it serves as a framework for how foreseen datasets will be handled such that they are findable, accessible, interoperable, and reusable (FAIR) to the greatest extent possible under ethical, legal, and technical constraints.

Table 2. Preliminary mapping of CBeRa datasets to repositories, metadata standards, identifiers, access conditions, and licensing, in line with FAIR principles

Data type	Intended repository	Metadata standard	Identifier	Access conditions	License
Multiomics datasets (genomics, proteomics, metabolomics)	ArrayExpress / EGA	MIAME, HL7 FHIR	Accession number	Controlled access (EGA) / Open (ArrayExpress, embargo possible)	CC-BY for open data
Biomedical imaging data	BioImage Archive	OME-TIFF, DICOM	Accession number	Restricted if sensitive; open otherwise	CC-BY (open)
CNS models & computational simulations	Zenodo / NDEx	SBML, GraphML	DOI	Open	CC-BY / MIT
Code & software	GitHub + Zenodo integration	CodeMeta / README	DOI via Zenodo	Open	MIT / GPL
Social media & web interaction datasets (aggregated, anonymised)	Zenodo	Dublin Core	DOI	Open (anonymised)	CC-BY
Training & educational resources	Zenodo / UCP institutional repository	Dublin Core	DOI	Open	CC-BY
Dissemination outputs (reports, newsletters, videos)	Zenodo / UCP repository	Dublin Core	DOI	Open	CC-BY
Stakeholder lists & internal governance documents	UCP secure repository (internal)	Minimal metadata (internal)	Internal identifier	Restricted (confidential)	Not applicable
Project deliverables (scientific & administrative)	CORDIS (EC portal) + Zenodo	EC metadata schema / Dublin Core	DOI (Zenodo) / Accession (CORDIS)	Public (CORDIS)	CC-BY

Where feasible, datasets will also be integrated with the European Open Science Cloud (EOSC), in line with the Grant Agreement requirements for repository federation. This will extend FAIR compliance beyond the project level, ensuring that CBeRa datasets are also discoverable and reusable within the broader European research ecosystem.

6. Roles and Responsibilities

The ERA Chair Holder and the Group Leader to be hired, supported by the Research Data Steward and project management team, have the primary responsibility for implementing the DMP. The Steering Committee will provide oversight and ensure alignment with project goals, while the Ethics Advisor will specifically monitor ethical and legal compliance in the management of sensitive data.

7. Conclusion

This preliminary DMP establishes the foundations for responsible, transparent, and FAIR data management across both scientific and non-scientific dimensions of the CBeRa project. It reflects the project's commitment to Open Science, GDPR compliance, and long-term preservation of data.

Key points include:

- A dual-layer approach covering both research data and project-supporting datasets.
- Repository mappings that demonstrate alignment with community standards and EC infrastructure (Zenodo, ArrayExpress, EGA, NDEx, CORDIS).
- Acknowledgement of restrictions where required, with justifications based on confidentiality, ethics, and data protection.
- Integration of ethics oversight and UNESCO guidance for data management in times of crisis.
- A clear update pathway at M18 and M60 to progressively increase detail as the research team, protocols, and Data Steward are established.

CBeRa recognizes that this DMP is preliminary and that specific details, such as precise dataset volumes, repository allocations, and budget allocations, will be finalized in subsequent updates. By clearly framing this deliverable as the first step in an iterative process, the project ensures compliance with Horizon Europe expectations while demonstrating readiness to evolve its data management practices as the project matures.