CERIC – ERIC
Scientific Data Policy

Contents

1. Introduction

The Central European Research Infrastructure Consortium, CERIC–ERIC, is a multidisciplinary Research Infrastructure for basic and applied research in all fields of Materials, Biomaterials and Nanotechnology. CERIC has been built by integrating leading national research facilities based in 8 countries (Austria, Croatia, Czech Republic, Hungary, Italy, Poland, Romania and Slovenia) into a unique European entity supporting the production of basic knowledge and technology transfer, and promoting the mobility of researchers in an international multicultural scientific environment. Based on this definition of CERIC – ERIC, the purpose of this document is to provide a common Scientific Data Policy for all its Partner Facilities in harmony with the FAIR principles[1], the PaNOSC and ExPANDS data policy frameworks and the National Data Policy of each country, if present.

By definition, to be FINDABLE, any data object should be uniquely and persistently identifiable; a data object is ACCESSIBLE by machines and humans under the conditions explained in this policy; data use a formal, accessible, shared, and broadly applicable language for knowledge representation in order to be INTEROPERABLE; the data object has a plurality of accurate and relevant attributes (usage license, provenance, community standards) to be REUSABLE.

The Scientific Data Policy is published by CERIC and accepted by the users during the submission of their proposals.

2. Purpose

The purpose of the CERIC-ERIC Scientific Data Policy is to define the time, the way and the place where data, metadata and raw data are or will be stored. CERIC has committed to providing FAIR data by the end of 2022. Therefore, this data policy should set the principles for managing data as the CERIC infrastructure for data management is being developed.

Having an open access data policy with data in well-defined formats has many benefits:

- It makes previously measured data available for further analysis without the necessity to repeat the experiment.
- It promotes data use, cross-disciplinary research and machine learning.
- Raw data becomes open to scrutiny by other researchers, which ensures scientific integrity and reproducibility of experiments.
- Scientists can mine data in previously unknown ways or reapply new methods to existing data.

This document describes the scientific data management at CERIC-ERIC Partners Facilities. The general process includes the generation of raw data from each experiment, which is then analysed by the research team. Whenever feasible and practical, the data format chosen by CERIC-ERIC for the raw data will be NEXUS/HDF5.

[1] https://www.go-fair.org/fair-principles/

3. Field of application

This CERIC-ERIC Scientific Data Policy applies to all CERIC-ERIC Instruments and Partners Facilities[2]. Having an identical approach to the management of scientific data will ease the life of scientists using more than one facility and add to the overall transparency of the scientific process.

[2] https://www.ceric-eric.eu/users/labs-and-instruments/

4. Definitions

Term	Definition
Partner Facilities (PFs)	Facilities ('Partner Facility') which provide access to users on behalf of CERIC members and which have the scientific and technical capability to contribute to the common strategic objectives, purposes and access capabilities
Metadata	The term metadata describes information referring to data collected from instruments, including (but not limited to) the context of the experiment, the experimental team, experimental conditions, electronic logbooks generated during the experiment and other logistical information.
Auxiliary Data	The term auxiliary data refers to data that provide contextual information regarding the experiment and its datasets but which are collected outside the context of the experiment conducted at the research facility, such as information about the sample images, provenance and preparation, data processing scripts, processing environment information such as software tools and versions used, etc.
Raw Data	The term Raw Data refers to data collected from measurements performed at CERIC – ERIC Partners Facilities. This includes data created automatically by software or added manually by staff expertise in order to facilitate the subsequent analysis of the experimental data. Detector data, administrative metadata, instrument metadata and scientific metadata are included in this definition
Beam Time	The term beam time refers to the period of time when the experimental team has access to the facility resources to conduct an experiment
Result	The term results pertains to data, and other outcomes arising from the analysis of raw data, e.g. algorithms, workflows etc. This does not include publications, which are handled by journals
Open Access	Open Access means belonging to the community at large, unprotected by copyright or patent and subject to the free of charge appropriation to anyone
Principal Investigator	The term principal investigator (PI) refers to the main proposer identified on the experiment proposal as the main person interacting with the research facility on behalf of the experimental team. For experiments outside of the facilities proposal system, the principal investigator can be considered to be the person initiating or performing the experiment
Experimental Team	The term experimental team includes the PI and any other person to whom the PI designates the right to access resultant raw data and associated metadata
Users	The term users refers to the members of experimental teams, which have obtained access to beam time

Term	Definition
Public Research	The term public research refers to publicly funded research which has been allocated access to the facility resources through a peer-review process and which is intended to lead to publication(s)
Proprietary Research	The term proprietary research refers to research done through purchased (commercial) access to the research facility
Embargo Period	An embargo is a period during which access to data or publications is reserved exclusively to the users
On-line Catalogue	The term on-line catalogue refers to a database of metadata containing links to raw data files, that can be accessed by a variety of methods, including (but not limited to) web-based browsers on desktop and mobile devices
Data Management Plan (DMP)	The term Data Management Plan (DMP) pertains to a document which is a defined strategy that covers the data produced, volumes, metadata requirements, data retention periods, data disposal, processing and analysis requirements and tools. The DMP shall enable the clarification of all aspects of data management between the facility and the users before the experiment takes place.
Processed Data	The term processed data pertains to the data obtained by processing raw data in an automated manner usually done at the facility
Data Object	The term data object is a sequence of bytes with a persistent identifier that refers to the collection of metadata, data, files, and (possibly) software describing a data collection. In the context of this document, the data collection is the output from one or multiple experimental sessions
Proposal	The term is referred to a detailed description of the planned experiment requesting to use one or more of the CERIC-ERIC instruments
Custodian	The term custodian refers to the Institute storing and providing access to raw data, metadata and results
Long Term storage	Storage for a period longer than the embargo. Usually, the speed of access to these data is slower.

Table 1: definitions

5. General principles

5.1 This data policy governs the curation of and access to scientific data and metadata collected and/or stored at the facility. This includes raw, processed and auxiliary data.

5.2 Acceptance of this policy is a condition for the award of beam time.

5.3 Users shall not attempt to access, exploit or distribute raw data or metadata unless they are entitled to do so under the terms of this policy.

5.4 Deliberate infringements of the policy may lead to denial of access to raw data or metadata and/or denial of future beam time requests at the facility.

5.5 Users shall ensure raw data and processed data are collected with accurate metadata such that raw and processed data fulfil the FAIR principles. The facility will define a minimum subset of metadata as an appendix to this policy.

5.6 Users shall endeavor to include auxiliary data to augment the experimental data.

5.7 Users are required to follow any recommendation provided by the facility on what constitutes good data management and any guidelines for completing DMPs.

5.8 GDPR compliance of data and metadata at the facility is ensured by that facility.

5.9 The facility will at its own discretion apply all reasonable efforts to ensure an accurate storing and curation of data as well as an uninterrupted access to data. However, failures caused by technical or human mistakes cannot be ruled out. The facility cannot warrant an absolutely accurate storing and curating. Access to data might be temporarily limited or impossible, especially due to necessary maintenance, service updates or failure of third-party service providers.

5.10 The facility cannot be made liable in case of unavailability or loss of data or data analysis software.

5.11 Access to raw data, facility processed data, auxiliary data, results (if uploaded), and the associated metadata is restricted to the experimental team during the embargo period. Thereafter, they will become openly accessible.

5.12 The embargo period begins at the end of the experiment session.

5.13 Raw data, facility processed data, auxiliary data, and results (if uploaded) will be stored by the facility for a minimum duration of 10 years. Metadata will be stored forever.

5.14 The PI can request an extension of the embargo period by following the facility defined procedure.

5.15 Data can always be made openly accessible earlier on request of the PI.

5.16 Access to raw data, processed data, auxiliary data, results (if uploaded), and the associated metadata in the facility is via a remotely searchable and indexed on-line catalogue using an open protocol.

5.17 Facility support staff (e.g. instrument scientists, computing staff) have access to all data or metadata curated by the facility in order to provide support to users. The facility reserves the right to use data still under embargo to improve facility processes and performance.

5.18 The research facility will release open data under an appropriate license.

6. Persistent Identifiers

6.1 Persistent identifiers, for example DOIs, shall be generated for raw data and metadata.
6.2 Persistent identifiers shall be generated for processed data that is generated by facility-maintained automated systems.
6.3 The experimental team shall be able to create a DOI for one or more specific datasets to be cited in a publication.
6.4 Users shall cite the persistent identifier in any publication that refers to the data (or to a subset of the data).

7. Raw data and associated metadata

7.1 All raw data and the associated metadata obtained as a result of publicly-funded access to the research facility will be open access after the embargo period, with the CERIC-ERIC Partner Facility in which the experiment took place acting as the custodian.
7.2 All raw data and the associated metadata obtained as a result of proprietary research will be accessible exclusively by the experimental team. Proprietary research users must agree with the Industrial Liaison office how they wish their raw data and metadata to be managed before the start of any experiment.
7.3 It is the responsibility of the PI to ensure that the metadata collected meets the minimum requirements by the facility and domain standards.
7.4 All raw data will be curated in well-defined formats, for which the means of reading the data will be made available by the facility.
7.5 Metadata that are automatically captured by instruments will be curated in a catalogue or similar repository which links the metadata to the raw data they are describing.
7.6 It is recommended to add additional rich metadata, as relevant for the domain, specifically considering functionality such as data discovery.
7.7 Data taken on or metadata from user supplied equipment must be provided to the facility for curation.
7.8 Raw data and metadata will be read-only for the duration of its lifetime.
7.9 Data will be migrated or copied to archival facilities for curation.
7.10 Open data are machine downloadable via an open protocol from the remote data catalogue.
7.11 Each dataset will have a unique identifier. Anybody publishing results based on open access data must quote the unique identifier (and related publications if available and appropriate).
7.12 The on-line catalogue will enable the linking of experimental data to experimental proposals and reports. Access to proposals will only be provided to the experimental team, reviewers and appropriate facility staff, unless otherwise authorized by the PI. Access to experiment reports is open to all.
7.13 The PI has the right to transfer or grant part or all of their rights to another registered person at any time.
7.14 The PI has the right to create and distribute copies of the raw data at any time.

8. Processed data and associated metadata

8.1 All processed data and metadata that are generated by facility-maintained automated systems during publicly-funded experiments will be made open access after the embargo period with CERIC-ERIC Partner Facility in which the experiment took place as custodian.

8.2 All processed data and metadata generated by facility-maintained automated systems during proprietary research will be accessible exclusively by the client who obtained the access. Proprietary users must agree with the Industrial liaison office how they wish their processed data and metadata to be managed before the start of any experiment.

8.3 Processed data generated by facility-maintained systems shall be curated in well-defined formats.

8.4 The facility does not guarantee readability for user-generated processed data in the case the processed data are stored in a non-standard format.

8.5 Users must include appropriate metadata describing the provenance of the processing carried out.

8.6 The metadata for processed data should be interpretable across domains and communities.

9. Auxiliary data

9.1 Auxiliary data stored together with raw data for publicly-funded experiments will be made open access after the embargo period CERIC-ERIC Partner Facility in which the experiment took place acting as custodian.

9.2 Auxiliary data shall be curated in the original format.

9.3 The facility does not guarantee the readability of auxiliary data.

10. Results

10.1 The intellectual property rights for results derived from the analysis of the raw data are determined by the contractual obligations of the person(s) performing the analysis.

10.2 The facility will provide a means for users to upload results and associated metadata to the facility and enable them to associate these results with raw data collected from the facility.

10.3 Users must include the appropriate metadata, which describe the provenance of the results.

10.4 These results might be stored long-term by the research facility. It will not be the responsibility of the facility to curate these data, e.g. to ensure that software to read / process these data is available

10.5 Access to the results of analyses performed on raw data and metadata is restricted to the person or persons performing the analyses during the embargo period, unless otherwise decided by the PI.

11. Good practices

11.1 The experimental team is encouraged to ensure that experiments' metadata are as complete as possible, as this will enhance the possibilities for the experimental data to search for, retrieve and interpret their own data in the future.

11.2 The experimental team is strongly encouraged to provide a complete log of the protocol carried out and what happened during the experiment. The logs must be entered in the electronic logbook if the facility provides one. In the absence of a facility electronic logbook, the experimental team must use other means (electronic if possible) and link the logbook to the data.

11.3 The facility undertakes to provide means for the capture of such metadata items that are not automatically captured by an instrument, in order to facilitate recording the fullest possible description of the raw data.

11.4 Researchers who aim to carry out analyses of open data and metadata are encouraged to contact the original PI to inform them and suggest a collaboration, if appropriate. Researchers must acknowledge the source of the data and cite its unique identifier and any publications linked to the same raw data.

11.5 PIs and researchers who carry out analyses of raw data and metadata are encouraged to link the software used to obtain the results of these analyses with the raw data / metadata using the mechanisms provided by the on-line catalogue. Furthermore, they are encouraged to make such software and results openly accessible.

11.6 Researchers are strongly encouraged to follow best practices adopted by many journals concerning citing the software used and developed for the data analysis.

11.7 For each publication using facility data, authors are strongly encouraged to make available the analysis procedure description, scripts, software and software environments that completely describe the process of data analysis from the raw and metadata to the published results, and which allow others to reproduce that analysis.

11.8 Authors are encouraged to deposit these files at the facility as auxiliary data associated with the dataset at the time of the submission of the manuscript, and to make them available as open access after the publication date.

11.9 Where a software tool cannot be made available, for example for licensing reasons, the analysis procedure description should explain which tool and version has been used, and how the analysis could be repeated if that tool was available.

12. Termination of custodianship

If the facility decides to stop acting as a custodian and/or to maintainer and provider of the metadata catalogue, the facility will inform the PIs concerned in a timely manner allowing them to make a copy of the data, metadata, and results that were generated by their proposal(s), provided the facility is aware of the e-mail address of the PI.