DARIAH Data Policy

Version 1 - December 2023

Contributors: Françoise Gouzi, Laure Barbot, Matej Ďurčo, Sally Chambers, Toma Tasovac
Contact: info@dariah.eu

Table of Content

1. Introduction

We live and work in the age of digital abundance. The massive quantities of available digital content do not, however, automatically translate into easy-to-use research workflows. Digitally enabled, data-driven research in arts and humanities has the potential to transform scientific enquiry and translate scientific advances into comprehensive solutions, policy actions and socio-economic impact, but arts and humanities researchers still face a number of challenges finding, accessing, producing and reusing digital resources: filtering out the noise, processing metadata of different quality and granularity, clarifying reuse rights for digital objects, choosing the right standard to model data, identifying the best tools to process a myriad of data formats, etc.

The DARIAH Data Policy outlines the basic principles and recommendations for managing and sharing research data within the DARIAH infrastructure and beyond.[1]

The term 'research data' is highly domain- and context-specific, and has been widely discussed in the Arts and Humanities[2]. We refer to research data as any facts, figures or information created, collected or generated for the purpose of scholarly analysis. Research data includes structured, semi- or unstructured raw data, containing text, images (2D or 3D), audio-visual recordings, geospatial information (maps and GIS) etc.

1.1 Target Audiences

The DARIAH Data Policy is primarily aimed at two target audiences: data producers and data providers. In this context, we define these two groups of actors as follows.

A **Data Producer** is an entity (a person or an organisation) that creates, generates or collects data for research purposes (see section three).

A **Data Provider** is an entity (a person or an organisation) that supplies or makes data available to others. A data provider may or may not be the original producer of the data. Professionals involved in making data available at data-providing institutions include **data stewards**, **data librarians** or **research data managers** (see section four).

---

[1] This is the first version of the DARIAH Data Policy. It will be shared with various DARIAH Bodies for consultation and updated accordingly. We would like to thank Dr. Erzsébet Tóth-Czifra for her work on an earlier version of this document.

[2] Edmond, J. , Horsley, N. , Lehmann, J. , & Priddy, M. (2022). The Trouble With Big Data: How Datafication Displaces Cultural Practices. London, Bloomsbury Academic. Retrieved December 20, 2023, from http://dx.doi.org/10.5040/9781350239654

1.2 DARIAH Resource Types

A large majority of DARIAH resources are produced at DARIAH Partner Institutions in member countries and contributed by their respective national consortia. In addition, there are resources produced in the context of various DARIAH activities, e.g. by DARIAH Working Groups, funded by the DARIAH Theme calls or by DARIAH and affiliated entities within EU projects.

The term **DARIAH resource** does not indicate direct ownership or control by the DARIAH consortium. Instead, it refers to the **context of production and (re)use**: a DARIAH resource is a resource created or used within the DARIAH ecosystem. The same principle of **contextual affiliation** applies to **DARIAH data:** DARIAH does not claim ownership of this data.

When we speak of DARIAH data, we do so in order to foster a common approach to making data produced and/or provided by our members discoverable both within the infrastructure and in the wider context of the Social Sciences and Humanities Open Marketplace and the European Open Science Cloud (EOSC).

DARIAH currently recognises 6 main resource types: *publications, datasets, training materials, software, services, data sources*. In the context of the DARIAH Data Policy, we will primarily focus on two specific resource types: *datasets* and *data sources,* which are defined in the table below.

Resource Type	**Definition**[3]
Dataset	A dataset is an organised collection of data. It is generally associated with a unique body of work, typically covers one topic at a time and is treated as a single unit by a computer.
Data source	A Data Source is a specific Service that exposes (meta)data about different types of Digital Objects. Examples of data sources are repositories, scientific databases, catalogues, etc.

Section two of this document places DARIAH's data policy in the wider context of DARIAH's involvement in Open Science. Section three addresses data producers, whereas section four addresses data providers.

[3] DARIAH categorises its resources in alignment with EOSC Exchange and the SSH Open Marketplace. Definitions are coming from the SSH Open Marketplace items types, the EOSC Exchange glossary and the OpenAIRE types.

2. DARIAH and Open Science

From its foundation, DARIAH has fostered **best practices in data management and reuse** for arts and humanities researchers, as part of its commitment to Open Science. See, for instance, the *DARIAH Impact Case Study: An Open Science Voice for the Humanities – A Humanities Voice for Open Science*.

Recent changes in research policy and research funding recognise data sharing as a necessary condition for scientific innovation and fostering changes of practice on a systemic level. Open data mandates (and data management plans) are increasingly becoming conditions of research funding both on the European and on national, institutional levels, and impacts the working conditions and practices of all researchers within the broader arts and humanities domain and not only those who identify as digital humanists. Furthermore, datasets are increasingly being described and contextualised through articles in data journals[4] and as a result being recognised as an innovative form of research output.[5]

DARIAH is a champion of **domain-specific approaches to implementing Open Science**: we aim to help understand and apply generic OS frameworks such as **FAIR** (Findable, Accessible, Interoperable, Reusable) and **CARE** (Collective Benefits, Authority to Control, Responsability, Ethics), to the concrete situations and practices of the arts and humanities researchers. The FAIR principles[6] describe how research outputs should be organised so they can be more easily accessed, understood, exchanged and reused. Whereas the CARE principles describe how data should be treated to ensure that Indigenous governance over the data and its use are respected. The CARE principles reflect the crucial role of data in advancing Indigenous innovation and self-determination.[7] For a DARIAH training resource on how the CARE principles complement the FAIR principles, see for instance *Thinking about the CARE Principles in the Digital Humanities* on DARIAH-Campus.

A DARIAH Working Group dedicated to Research Data Management in Arts and Humanities aims to support DARIAH to fulfil its mission of facilitating access to data, and supporting the work of the research communities with best practices and methods. It

[4]See, for instance: Iain Hrynaszkiewicz, Natasha Simons, Azhar Hussain, Rebecca Grant, Simon Goudie. Developing a Research Data Policy Framework for All Journals and Publishers. Data Science Journal, 19 (1). 2020. https://doi.org/10.5334/dsj-2020-005; and Jihyun Kim. An Analysis of Data Papers Templates and Guidelines : Types of Contextual Information Described by Data Journals. Science Editing 7 (1) : 16-23. 2020. https://doi.org/10.6087/kcse.185

[5] Toma Tasovac, Laurent Romary, Erzsébet Tóth-Czifra, Rahel C. Ackermann, Daniel Alves, et al.. The Role of Research Infrastructures in the Research Assessment Reform: A DARIAH Position Paper. 2023. ⟨hal-04136772⟩

[6] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. 2016. https://doi.org/10.1038/sdata.2016.18

[7] Australian Research Data Commons and Global Indigenous Data Alliance

organises workshops to raise awareness but also produces guidelines **translating generic statements on RDM and FAIRness to the specific needs of the A&H communities**. The expert members of this group recently released an interactive online publication *Research Data Management for Arts and Humanities : Integrating Voices of the Community* which provides an exhaustive overview of the challenges faced by the Arts & Humanities disciplines with RDM.

3. Data Producers: how to share your data

Proper data management practices ensure the accuracy and reliability of research findings. They allow for the preservation of raw data, enabling others to verify and reproduce the results, which is fundamental to the scientific method. Researchers should be able to replicate analyses using the same datasets to validate and build upon existing findings. Efficient data management promotes data sharing and collaboration within the research community. Furthermore, properly managed data ensures its long-term accessibility and usability. This is crucial for preserving research outputs for future generations and avoiding loss of valuable information.

Research Data Management (RDM) is a broad term encompassing all practices and actions to ensure that research data is: secure, sustainable, easy to find, understandable, and (re)usable. RDM includes activities such as planning and funding, collecting and analysing, preserving (publishing) and sharing, as well as discovering and reusing. It is about taking proper care of data, not only during a research project, but also before and in the longer term.[8]

Similarly, the **Research Data Lifecycle** is a key concept within RDM which can be understood as the different stages that research data go through before, during, and after a research project. Each stage of the research data lifecycle entails various data management activities, and the choices made in one phase influence the next one.

A **Data Management Plan (DMP)** is a core tool for Research Data Management (RDM) to assist researchers in using RDM best practices. A DMP is a document specifying how research data will be handled both during and after a research project. It identifies key actions and strategies to ensure that research data are of a high-quality, secure, sustainable, and – to the extent possible – accessible and reusable[9]. Online tools such as DMP Online, can help researchers create, review and share Data Management Plans.

[8] https://www.ugent.be/en/research/datamanagement/why/rdm-explained.htm
[9] See for example:
https://www.ugent.be/en/research/datamanagement/before-research/datamanagementplan.htm

3.1 Project planning and design phase

Creating a research data management plan in the early stages of project planning is essential for ensuring the success, standards compliance, and long-term impact of the research. It contributes to efficient and ethical data handling, enhances collaboration, and facilitates the reproducibility of research findings. Developing a data management plan helps researchers define the scope and objectives of their project. It forces them to clarify what data will be collected, how it will be managed, and what tools and methodologies will be used. This clarity is essential for maintaining focus and ensuring that the project stays on track.

For instance, if you work with human subjects, to collect appropriate consent from your participants, you already have to be aware of how and where exactly you will publish their input, how anonymisation will be carried out etc. You can use a DMP to identify and flag these issues and arrive at an agreement with your project partners. Both third-party data providers (such as Cultural Heritage Institutions) and data repositories should be involved from the planning phase when the project is still flexible enough to accommodate and sufficiently address preconditions of FAIR and open sharing.

Recommendations

- ❖ Allocate 5% of the project budget for data management throughout the project
- ❖ Use the Heritage Data Reuse Charter template (Annex 1) between CHIs and researchers to guide mutual reuse agreements in the project planning phase

Useful resources

- Planning to meet the costs of managing research data to be FAIR
- DARIAH Working Group Ethics and Legality in the Digital Arts and Humanities (ELDAH) provides GDPR compliant consent form templates for DH research purposes. This tool enables researchers to correctly approach the management of personal data in compliance with the General Data Protection Regulation (GDPR).
- Zotero library Data management best practices in the Humanities (DARIAH RDM Working Group)
- DARIAH Pathfinder to Data Management Best Practices in the Humanities
- "Open Data for Humanists, A Pragmatic Guide"

- A collection of DARIAH-relevant DMPs are available in the following Zotero bibliography
- OpenAIRE tool: How to identify and assess Research Data Management (RDM) costs?

3.2 Data modelling, standardisation and documentation

When data is shared, it becomes decoupled from the context of its creation. Therefore, to enable other researchers to understand and reuse existing data, this context of creation has to be captured in the form of rich and preferably standardised documentation that is accessible to scholars from different backgrounds. This can be a README file or any other supplementary document where you describe provenance information, contributors, how the data have been curated, how to cite it and what are its limitations and sources of incompleteness (especially in cases where you can share only a subset of your resources/outcomes due to limitations in terms of effort or legal, ethical restrictions). However, it is even better if you make most of this information available in standardised formats using (meta)data models, vocabularies and ontologies that are widely shared in your disciplines (if any).

Recommendations

- ❖ Use open file formats[10] to publish your data. Avoid using proprietary file formats; if confronted with proprietary file formats (e.g. as legacy data) convert these to open file formats. However, if the conversion causes information loss, keep the original datasets too (there may be better conversion techniques in the future).
- ❖ Describe the datasets with rich metadata using widely adopted standards such as, for instance, TEI, MEI, Epidoc, EAD, CIDOC CRM etc.
- ❖ Use existing controlled vocabularies for enriching the data and creating metadata wherever possible.

Useful resources

- Data model and standards (DARIAH RDM Working Group)
- CLARIN Standards Information System, especially the recommendations for standards and specifications and the data formats

[10] https://en.wikipedia.org/wiki/List_of_open_file_formats

- The LexSeal self-assessment framework translates both the FAIR principles and the 6 core principles of the Heritage Data Reuse Charter to actual data curation practice and provides an easily understandable self-assessment framework for data quality. Having been co-developed by DARIAH and ELEXIS, it had been designed for lexicographical data sets but curators of other data types can also find relevant know-how in it.
- The SSH Open Marketplace Workflows have been designed to **support researchers in selecting and using the appropriate standards for their particular disciplines and workflows.** It will help you to think of data curation as a process that is well-aligned with your specific research lifecycle as well as with your disciplinary contexts.
- DARIAH provides a suite of Vocabs services that allow for collaborative creation, maintenance and publication of vocabularies and taxonomies of any kind.

3.3 Publication

Making data available in a trustworthy data repository is a core requirement of open and FAIR data mandates. By doing so, you make sure that humans and machines alike can find your data, and that the data will be preserved over the long term, and therefore remain available for future research or verification.

Trusted repositories typically have established procedures and standards to ensure the integrity of the data they host. They may implement measures such as checksums and data validation checks to prevent corruption or tampering. They also require documentation and metadata for deposited data. This includes information about how the data was collected, processed, and any relevant contextual details. Proper documentation contributes to better understanding and usability of the data, increasing its quality. Furthermore, reputable data repositories often implement version control mechanisms, allowing users to track changes and updates to datasets over time. This helps maintain a clear historical record of the data, ensuring that users are aware of any modifications or improvements. Finally, trusted repositories will assign persistent identifiers to your datasets, which will facilitate their discoverability and citability.

Dynamic data-provision approaches such as Wikidata or a federated set of APIs for programmable corpora (e.g. Dracor[11]) are very valuable tools to work with data, but should, nonetheless, be accompanied by regular snapshots to trustworthy repositories (which can be referenced in a stable manner).

[11] Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In Proceedings of DH2019: "Complexities", Utrecht University, https://doi.org/10.5281/zenodo.4284002

Recommendations

- ❖ Clarify all questions regarding intellectual property rights and privacy before making your datasets public.
- ❖ Deposit datasets in a reliable, trustworthy and preferably certified[12] repository.
- ❖ Use an institutional repository, if available and if required by your institutional data sharing policies, or, alternatively, use one of the DARIAH affiliated repositories: DARIAH HAL collection or DARIAH Zenodo Community.
- ❖ Publish datasets under CC-BY 4.0 licence by default. Only choose a more restrictive licence if necessary.

Useful resources

- DARIAH affiliated repositories can be found browsing the DARIAH service catalogue: https://www.dariah.eu/tools-services/tools-and-services/
- DARIAH's Open Access Guidelines (HAL ID: halshs-02106332)
- OpenAIRE: How to find a trustworthy repository for your data?
- Registry of Research Data Repositories (re3data)
- FAIRsharing gives access to a curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies (see section "Certifications and community badges" to find if the repository is certified)
- Creative Commons licences
- This resource developed within the framework of the CENDARI project outlines key steps of sustaining outputs of research projects

[12] https://www.coretrustseal.org/why-certification/requirements/

4. Data Providers: how to connect your repository to the DARIAH ecosystem

While the previous section focused on DARIAH recommendations to data producers (e.g. researchers and research institutions), this section concentrates on data providers. It guides the people and institutions responsible for data services along some of the recommended paths should they wish to be better connected to the DARIAH ecosystem.

4.1 DARIAH Data Sources

Some of DARIAH Services - see DARIAH Services Policy - can be considered Data Sources. Because of the complex institutional landscape of European Research Infrastructures, it does not necessarily mean that all the datasets in these data sources are automatically considered DARIAH data. For example, an institutional and discipline agnostic repository (such as the French Open Archive HAL) is onboarded as a DARIAH service, but not all of the scientific papers or other objects in this repository (i.e in Physics or Life sciences) should automatically be considered DARIAH data.

Most of the DARIAH data sources are services declared by DARIAH National Coordinators as part of their national reporting. Most known data sources are repositories, but scientific databases, journal and publisher archives, research information systems (CRIS) and aggregators of Data Sources (e.g. OpenAire Research Graph or re3data) are also considered as data sources. In the context of this data policy, we focus exclusively on only repositories.

Recommendations

- Register the repository as a data source in the Marketplace and tag it properly (see Guidelines for Contributing to the SSH Open Marketplace).
- Register your repository with OpenAIRE by making the repository metadata compliant with the OpenAire Guidelines.

Metadata that is compliant with the OpenAIRE Guidelines will be included in the OpenAire Research Graph, and, if it satisfies the inclusion criteria for EOSC Providers of data sources[13], it will also be integrated into the EOSC Catalogue.

4.2 Trustworthiness and interoperability

Over the last 10 years, many universities and research centres have developed data repositories allowing permanent access to datasets in a trustworthy environment. However, it is not always easy to evaluate the quality of these repositories[14]. To this end, some certification instruments[15] for digital repositories have been developed, a prominent one being the CoreTrustSeal (CTS). At the same time, there are also data repositories that have no certification but have earned trustworthiness through many years of reliable operation by trustworthy providers and large user bases, like Zenodo.

Furthermore, because researchers in the arts and humanities rely on thick descriptions[16] of their research objects and because aggregation systems tend to favour reductionist metadata schemas for the sake of interoperability, DARIAH data sources should pay special attention to reflecting the context of creation and reuse of DARIAH data reflected in the metadata schemas they use.

Recommendations

- Study and follow the requirements of the Core Trust Seal.
- Implement OAI-PMH protocol as a default mechanism to facilitate harvesting of the metadata by third party aggregators, in order to promote dissemination and thus findability of the resources.
- Expose metadata adhering to the DataCite Metadata Schema from OpenAIRE as a minimal schema.
- Expose metadata in multiple formats according to the requirements of relevant aggregators, to optimally convey/propagate existing information.

---

[13] https://eosc-portal.eu/eosc-providers-hub/how-become-eosc-provider/how-become-eosc-provider-a-general-overview#Inclusion_criteria_for_EOSC_Providers_and_different_kinds_of_EOSC_Resources

[14] Disappearing repositories -- taking an infrastructure perspective on the long-term availability of research data https://doi.org/10.48550/arXiv.2310.06712

[15] CoreTrustSeal-Requirements-2023-2025_v01.00, https://doi.org/10.5281/zenodo.7051095

[16] Tóth-Czifra, E. 2020. 10. The Risk of Losing the Thick Description : Data Management Challenges Faced by the Arts and Humanities in the Evolving FAIRData Ecosystem. In Edmond, J. (Ed.), Digital Technology and the Practices of Humanities Research. Open Book Publishers. Tiré de http://books.openedition.org.inshs.bib.cnrs.fr/obp/12064

- ❖ Use specialised schemas tailored to domain-specific needs internally in order to capture as much as possible relevant/available information (e.g. regarding the context, provenance).

- ❖ Use persistent identifiers (PIDs) to **identify datasets in the repository**[17] following one of the established PID systems: e.g. DOI, Handle, ARK, URN.

- ❖ Honour **Data Citation Principles**[18] (Importance, Credit and Attribution, Evidence, Unique Identification, Access, Persistence, Specificity and Verifiability, Interoperability and Flexibility) enabling data reuse.

- ❖ If providing resources with restricted access, adopt Federated Identity Management using technologies such as Security Assertion Markup Language (SAML) and OpenID Connect (OIDC)[19] and integrate with relevant Authentication and Authorisation Infrastructures (AAI) most notably eduGain and/or DARIAH-AAI (as part of EOSC-AAI Federation).

Useful resources

- Tibor Kálman (2022). Introduction to Persistent Identifiers. Version 1.0.0. DARIAH-Campus. [Webinar recording]. https://campus.dariah.eu/id/4uEjJqTVxczKYmJyc3w9n
- CoreTrustSeal Requirements 2023-2025 (V01.00). Zenodo. https://doi.org/10.5281/zenodo.7051012
- Mari Kleemola, & René van Horik. (2020, April 23).. SSHOC Webinar: How to improve the quality of your repository - SSHOC and certification of repositories. https://doi.org/10.5281/zenodo.3774396
- ePIC Persistent Identifiers for eResearch http://www.pidconsortium.net/
- Inclusion Criteria to OpenDOAR: https://v2.sherpa.ac.uk/opendoar/about.html
- Suggest a repository to re3data: https://www.re3data.org/suggest

[17] https://www.pidwijzer.nl/en

[18] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk and DARIAH Winter School 2016: 'Open Data Citation'. Version 1.0.0. DARIAH-Campus. [Video]. https://campus.dariah.eu/id/v3NdCcSyB--q9OzWeluGI

[19] https://wiki.oasis-open.org/security/FrontPage and https://openid.net/developers/specs/