DARIAH-EU
Digital Research Infrastructure
for the Arts and Humanities

Data management resources for arts and humanities scholars. At its present state it is a work in progress. I am planning to add further modules e.g. about IPR and software sustainability - you are very welcome to suggest relevant resources to those topics as well. The document is licensed under a CC-BY license.

Please feel free to leave comments or contact me directly at erzsebet.toth-czifra@dariah.eu.

Thanks for all your feedback in advance.
Best,
Erzsébet Tóth-Czifra]



DARIAH-EU
Digital Research Infrastructure
for the Arts and Humanities

1. Why research data management?

Systematically planning how you will collect, document, organize, manage, share and preserve your data has many benefits. It helps to build a common framework of understanding with your collaborators and other stakeholders such as data archivists or professionals of GLAM institutions. But you can also think of your future self as your primary collaborator, imagining yourself looking for that one missing reference or document buried somewhere in your (physical or virtual) files. Proper data management allows you to prevent loss of information, to virtually travel back and forth in time and ask questions from your future/past self and also to make your data future proof for others to find, understand, use and cite. In addition, by documenting your data and recommending appropriate ways to cite it, you can be sure that you will receive credit for your data products and their use.

This resource list brings together tools, videos, short articles, and other training materials that might be relevant when reflecting on your data management processes both in the immediate context of your research and in their broader disciplinary context. Its aim is to equip you with tools and practical advice, but more importantly, with conceptual twists that will help you to establish ethically committed, optimal and as open as possible research and data management workflows. Each section consists of a brief introduction, describing the context; resources in sharing best practices for maximizing the potential of arts and humanities research data; and a list of tools you can make an integral part of your data management routines. Think of it as an inventory: you can browse through the resource list and select the items or topic areas the most relevant for your work. Let's get started with a healthy boost of motivation:

- Data Champions (Why research data management 4/4)

https://www.youtube.com/watch?v=0uW1tvcHy4w#action=share

- Primer on Data Management: What you always wanted to know*

https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

- An introduction to the basics of Research Data

https://www.youtube.com/watch?v=q2aiDJzJPuw

- The Open Research Data and Materials chapter of the FOSTER Open Science Handbook will make you familiar with the core concepts and practices of (open) data management.

https://book.fosteropenscience.eu/en/02OpenScienceBasics/02OpenResearchDataAndMaterials.html

- ...and a dose of cautiousness: How NOT to share your data: avoiding data horror stories https://zenodo.org/record/400982#.XMB8hqRS-Uk

- Can you recall when data is first considered when starting a new project?
- When interacting with data, we can be recognised as data producers, but also as consumers, curators and evaluators of data. Which of these roles apply to you?

Have you ever lost information or (re)sources that had been relevant to your work? How it could have been avoided?

2. Data in the Humanities

This section discusses research data management practices in a domain-specific context. Depending on what discipline or sector you work in, you may not even realize you are working with data. In the arts and humanities, this term covers a great variety of resources: primary sources, secondary sources and outputs (such as laser scanner data, musical notations, voice recordings, geospatial data, survey data, corpora, annotations, text encodings, bibliographies, dictionaries, critical editions etc.) that can be both digital and analogue, machine created or human created, highly structured or utterly unstructured. In many cases, data are 'born' or become qualified as data in the course of the research process. Think about, for instance, publications or social media posts. The ways in which we interact with them (learning from them vs. collecting, structuring and enriching them for analysis) define whether they are considered as data in specific research scenarios. The resources below provide an overview of the nature of humanities data.

2.1. For a comprehensive overview: **The PARTHENOS training module Manage, Improve and Open Up Your Research and Data** is a full, independent training module that will guide through some of the fundamental issues in research data management in the humanities such as: **what is metadata; who are the stakeholders in a data management in a humanities context; how to manage cultural heritage assets; data quality assessment; ethics and research; data management planning; or Research Infrastructures and data policy**.

2.2. The Open Data for Humanists: a Pragmatic Guide is a great starting point for researchers who are eager to share their data but don't know where to begin. It gives **practical advice on a range of questions and issues, for example:** what questions to ask your archivist during your first visit to the archive; how to **manage source data; processural phase data and**

publications as communication of data; and how small changes across these three points in your research workflow can make big differences.

2.3. Introduction to Humanities Research Data Management: this presentation by Ulrike Wuttke gently introduces humanities and cultural heritage researchers to activities and issues around **planning, organizing, storing, and sharing data and other research results** and products. **It will guide you to identify the main characteristics, potentials, and challenges in your own research data. It will encourage you to think through all phases of data collection; data documentation and metadata; ethics and legal compliance; storage and backup; selection and preservation; data sharing and responsibilities in data sharing; and resources for data management in the context of your own project.**

Are you working in a digital or in an analogue environment? What can be considered as research data in your field and what is not? What are the scholarly workflows that turn source material into data? How do you develop a shared understanding about them with your collaborators? Who are the different stakeholder groups that are involved in your data management workflows? Have you ever encountered any difficulties in accessing cultural heritage resources? How can cultural heritage professionals (archivists, librarians, museologists) support your work? How do you discover data that is relevant for your research and which factors help you to assess its quality and trustworthiness?

3. The devil is in the context: a processural view on data curation

In the scholarly discourse about data management, there has been much discussion of the desirable format and structure of the final outputs of a research project, which we call research data. However, managing your research data is by far not a final output-oriented mechanism: proper data management takes place in all phases of the research lifecycle. By planning how the data will be analyzed well ahead, one can choose how best the data should be collected.

DARIAH-EU
Digital Research Infrastructure
for the Arts and Humanities

The formats, standards or data model you work with determine whether, and how, other research communities will be able to make sense of your data. You will need to answer questions related to documentation, storage, quality assurance and ownership for each stage of the data lifecycle so that your research can be viewed, managed, accessed and ultimately assessed in terms of the integrity of *processes*, rather than only as products. The more people can follow how you reached a conclusion based on your analysis procedures (such as additional modifications to the data, the model used, the code and codebook used to run the analysis, and hardware and software specifications and other supplementary materials that help the interpretation of your data), the more confidently they can reuse your data. The use of data management plans (see section 5) help structure your thinking in this respect.

3.1. The **Parthenos Standardization Survival Kit (SSK)** is 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 life-cycle as well as with your disciplinary contexts. It collects r**eal life researcher-oriented use cases** from literature to heritage science, including history, social sciences, linguistics, etc. **This format allows domain experts to share their guidance and advocacy about the use of standards in a meaningful way, that is, embedded in specific research scenarios.** Those scenarios can be seen as a living memory of what should be the best research practices in a given community, made accessible and reusable for other researchers wishing to carry out a similar project but unfamiliar with the recommended tools, formats, or methods to use.

3.2. The Top 10 FAIR Data & Software Things, Humanities: Historical research resource shows how the fairly abstract FAIR principles can be translated into real community (best) practices. It will provide you with practical advice on **how to structure, organize and model your data,** what are the benefits of using **controlled vocabularies and ontologies**, **how to license** and how to **cite** your data properly.

Open formats

Choosing an appropriate, open format for your data significantly increases the chances of being readable well into the future by anyone on any software program. The Wikipedia article on open format will help to check which open formats work best for your data.

DARIAH-EU
Digital Research Infrastructure
for the Arts and Humanities

Can you recall how your source data has been ‘cooked’ in your recent projects? How your working environment (sources, tools available etc.) affected these procedures?

How can you facilitate mutual understanding of each other’s data within your discipline? Do you have shared vocabularies, ontologies or metadata standards available?

Do you have experiences with interpreting data from other scholarly domains?

4. Sharing your data

Making your data available in a trustworthy data repository is an essential step in your research data management workflow. By doing so,

- You can make sure that others (humans and machines alike) can find your work, beyond accidentally stumbling across it on your project website.
- You can make sure that your data will be preserved over a long term, and therefore remain available for future research or verification.
- You make your data citable. Assigning persistent identifiers to your data will point to their exact location , even if that changes over time.
- Sharing your data in a repository inevitably brings quality improvement. By depositing them, you will clean them, document them to make them understandable for third parties and add the metadata necessary for their deposit.

Depending on the scope of your project and the volume of your data, depositing can be realized in a complex, iterative process with a strong emphasis on precise data quality assessment, but it can also be simple and easy. As a first step, check which kinds of support structures are available in your institution. If your institution has solid data repository facilities and/or data stewards who can help your work, you are lucky! It is well-worth involving them from the project planning phase onwards and discussing the best solutions, which work well for your research but are also compliant with the standards, specifications and protocols along which the repository operates. Repository staff can also assist you with understanding any specific data management requirements and associated costs. If your funder requires a data management plan, this information should be included.

In deciding where to store your data, you may have a number of choices about who will look after it and how its findability and potential can be maximized. You can use generic repositories such as Zenodo, which will take almost any datasets, but you can also use more specific ones such as the DARIAH-DE Publikator that are geared towards certain disciplines or research data types. If your institution has data sharing policies, it is advisable to use your own institutional repository.

- The OpenAire repository guide helps you specify your expectations towards repositories. Its second, more detailed version has been optimized for Horizon 2020 projects.

- You can find suitable research data repositories that best match the technical or legal requirements of your research data in the re3data.org database. The re3data database has a special application for arts and humanities, the demonstrator instance of **Data Deposit Recommendation Service (DDRS).**
- A comprehensive overview of repositories and knowledge-bases, alongside the standards they implement and the policies that endorse their use, can be found at https://www.fairsharing.org, which is a manually curated, community driven registry of standards (models, formats, reporting guidelines, terminologies, identifier schema, metrics), databases and journal and funder data policies.

4.1. Cite to be cited!

One of the main findings of the Barriers and Pathways to Community Engagement report conducted by the DARIAH Community Engagement Working Group is that data citation is not yet an established scholarly practice in the humanities domain. You can easily change this for the better and contribute to a research culture where data sharing is better rewarded in terms of citations. Once a Persistent identifier is linked to your data, it makes easier both for humans and machines to properly cite it. You can further increase your citation chances by recommending data citation formats for your data sets.

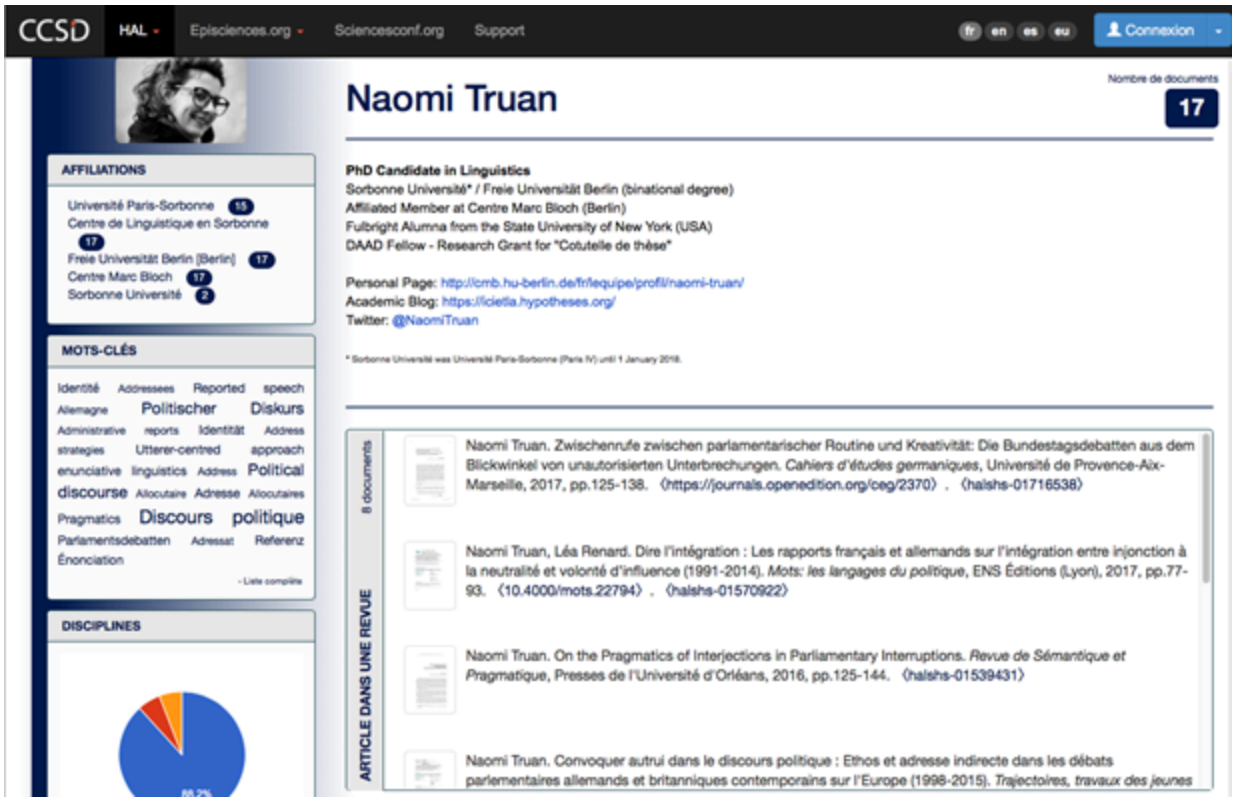
- According to the DataCite standards, the minimum metadata that is necessary to identify and locate the research products should include: the name of the creator, the year of publication, the title, the publisher and the identifier. If relevant, you can also add version and resource type information. For more detailed guidelines you can check the Force11 Joint Declaration on Data Citation Principles, a common, cross-disciplinary reference document.
- The **CiteAs** tool helps you to get the correct citation forhttp://it.dariah.eu/dariah-winter-school-open-data-citation-for-social-science-and-humanities/m for diverse research products, from software to preprints.
- The Open Data Citation in the Social Sciences and Humanities module of the DARIAH Winter School puts data citation best practices in a domain-specific context.

4.2. A case study: different levels of being an open scholar

DARIAH-EU
Digital Research Infrastructure for the Arts and Humanities

Going beyond the scope of the research paper and sharing a broader scope of your research outputs allows you to address different communities of audiences to share different parts of the research process and results. For instance,

- sharing publications (Green) Open Access, so they're accessible to all researchers and to the general public as well,
- sharing research ideas, musings, or results via blogs and social media (Twitter, Facebook etc.) with your peers and the broader public, and
- sharing data (annotations in this case)) with your peers and especially for use by Digital Humanities researchers.



You can learn more about Naomi's lessons here: https://www.fosteropenscience.eu/node/2605

What are the support structures available for you in terms of data management? What are or what would be the most important criteria for you when depositing your data sets?

Is data citation a prevalent practice in your field?

Which elements do you recognise as compulsory in the unambiguous identification of a dataset?

5. A recipe for your research project: the Data Management Plan

Research communities are increasingly supported, incentivised, and in some cases even required, to establish a Data Management Plan at the outset of a project to consider the approach for creating, managing and sharing all research outputs. The Data Management Plan (DMP) is a standard support document that helps to formalise your thoughts and materialise your choices of how you manage your data throughout the life-cycle of your project. Creating a data management plan for your research project is similar to writing a recipe for a meal, which enables others to cook it too (nb: counter-incentives of sharing may also show a degree of similarity). It allows you to create an internal roadmap for your project and to explain why you have chosen certain solutions and formats; how the data you are working with is relevant to the research questions; where and what the value is in your data; what compromises you had to make e.g. due to legal restrictions; what the access conditions are as well as the estimated costs of data curation and deposit; and who is responsible for each data management activity.

Writing a DMP does not necessarily require the investment of a lot of time and effort. The training resources referenced above in section 2. will give you a good introduction on how to get started. More specifically:

- You can find a chapter dedicated to Data Management Planning in the PARTHENOS training module Manage, Improve and Open Up Your Research and Data Training module.
- The Introduction to Humanities Research Data Management resource also gives you a range of good advice and practical examples on how to create your own executable research data management plan and how to utilize it for your own good.

DMP Online

Although the real benefits of having a Data Management Plan go far beyond merely achieving compliance with funder requirements, there are tools you can use to ensure that you meet these specific stipulations. One of the most widely used of them is the Digital Curation Centre's **DMPonline**. It provides tailored guidance and examples to help researchers write data management plans. The tool includes templates from a number of funding organisations

(e.g. within the UK, USA, and the Netherlands) as well as for Horizon2020 and the European Research Council. You can explore the tool yourself at http://dmponline.dcc.ac.uk

Further resources:

- Horizon 2020 Data management Guidelines: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- The European Research Council's guide for Open Research Data and Data Management Plans: https://erc.europa.eu/document-category/guides
- Arts and Humanities Research Council Data Management Plan: https://zenodo.org/record/1745533#.XMFsHaRS-uU
- National Endowment for the Humanities - grant examples: https://www.neh.gov/grants
- RDMO - Research Data Management Organiser (available in German, a tool to support the systematic planning, organisation and implementation of the data management throughout the course of a research project. RDMO is funded by the Deutsche Forschungsgemeinschaft (DFG). https://rdmorganiser.github.io/

Do you use to think through your project from a data point of view in its planning phase?

Where can you explore the already possible requirements and solutions for data sharing?

How do you decide what to keep and what to destroy?

How do you share the responsibilities for each data management activity among your collaborators?

6. Data in publications and data as publications

Another challenge in pushing forward the data sharing culture in the humanities is that the traditional paradigm of publishing papers still serves as the highest value currency of academia. Compared to this,the rewards of, and incentives for, data sharing is negligible. However, sincedata sharing is increasingly encouraged by funders policies, we have more and more

opportunities to bring these two types of scholarly publications, data sharing and article or book publishing, closer to each other.

6.1. The networked publication: interlinking the underlying data with your papers

By allowing the audience of your papers to look 'behind the scenes', and connecting them with the underlying data, you support the validation of your findings–and take a huge step forward in research transparency. Many publishers support their authors to cite data in their publications. Some are even issuing policies that require researchers to make the data underlying their published results available within a data repository.

- Do you know journals or book publishers in your discipline with data policies? Here you can find some examples:

Humanities (MDPI): https://www.mdpi.com/journal/humanities/instructions
Journal of Contemporary Central and Eastern Europe (& all Taylor and Francis journals): https://www.tandfonline.com/action/authorSubmission?journalCode=cdeb20&page=instructions#dsp

6.2. Data journals in humanities

In addition, or in parallel, to making your data available in a repository, you can maximize the credit that you gain for them by publishing them in a data journal. Data journals, usually built on the top of data repository services, are designed to introduce, describe and contextualize data sets to facilitate their online exploration. One of the important benefits of publishing your data in terms of research data management is that it undergoes a peer-review process, which means that experts validate its quality, provide templates for its proper description and offer guidance on where to deposit it.

- Data journals in the humanities:

Research Data Journal for the Humanities and Social Sciences

Journal of Open Humanities Data Dataverse

Journal of Cultural Analytics

Journal of Open Archaeology Data

DARIAH-EU
Digital Research Infrastructure
for the Arts and Humanities

Journal of the Text Encoding Initiative

(Video) Journal of Embodied Research

Book or journal peer review is an established institution for the quality assessment of research papers but how and from whom can you receive feedback for your data?

What strategies are available for you to increase the transparency of your research findings?

Which components of your work can you make available via publication? What remains hidden in traditional journal or book publications?

