What is data in the humanities and how you can make the most from your hard-earned research data?

Erzsébet Tóth-Czifra
Open Science Officer

DESIR Winter School, Lisbon, 10.12.2019

With special thanks to: Laurent Romary, Jennifer Edmond, Toma Tasovac, Ulrike Wuttke and Paola Masuzzo.
We no longer produce only scholarly outputs that can be placed on a bookshelf.

“Research data are first-class citizens in science and scholarship.” (Paola Masuzzo)
Data in the arts and humanities: still a dirty word?

Miriam Posner @miriamkp
7:50 PM - Oct 31, 2018 from Los Angeles, CA - Twitter Web Client

Humanists out there, specifically non-digital humanists: If someone were to call the sources you use “data,” what would your reaction be? If you don’t consider your sources data, what makes them different?

54 Retweets 170 Likes

Danica Savonick @DanicaSavonick - Nov 1, 2018
I call it data in the second sentence of this post! More when teaching close reading to students than discussing with colleagues. For undergrads just beginning close literary analysis, it works well.

Adam Arenson @adamarenson - Oct 31, 2018
Replying to @miriamkp
I think that part of what I do as a historian is data creation, alongside knowledge creation: I find meaning in historical artifacts that has been overlooked or ignored, and my original research argue for its importance/relevance.

Miriam Posner @miriamkp - Oct 31, 2018
Yes, I can see that.

Matthew DeForest @mmdeforest - Nov 1, 2018
Replying to @miriamkp
Data feels like it consists of discrete units (e.g. numbers in a table). Most of what I’ve done resists being rendered as discrete units. Even when counting the number of times an author uses a word, those words exist in sentences that resist meaningful data scrubbing.

Miriam Posner @miriamkp - Nov 1, 2018
I agree, I’ve had that same feeling. Occasionally I’ve been told that whatever it is that’s between the words could, in theory, be counted, but that seems unlikely.

Jenny Sayers @jenterysayers - Nov 1, 2018
Replying to @miriamkp
When I hear “data,” I think “record” (not taken or given, but produced) as well as “index” or “trace.” The media I study may be records and traces of history, but they do not always point to actuality or specific events. They are processes, not objects, of knowledge.

https://twitter.com/miriamkp/status/1057706465866133504
1. Introduction to Research Data Management
   Identifying humanities research data
**Exercise:** What are your data?
   Why do we care, why should we care?
   The FAIR principles

2. Data management good practices in the research workflow
   1. Data reuse and data collection
      **Exercise:** How do we find data for reuse?
      **Exercise:** Data citation
      2. Data processing and analyzing
      3. Data sharing, storing and publication
      **Exercise:** The networked publication
      **Exercise:** How to find a suitable repository for your research?
   Discussion: Pick a statement.
What do you see?
What do you see?

What all this tells us about the nature of research data in the humanities?

• Multilingualism.

• Humanities research lives from enrichment of data (layers of interpretation) data curation happens in a continuum: the way cultural heritage resources are made available form a continuum with layers of analysis based on them.

• Problematic to distinguish between primary data (raw data) and secondary data.

• Access to machine readable artifacts and digital collections is crucial.

• Shared ownership between data creators, data curators and the human subjects, researchers, Cultural Heritage Institutions and publishers.

• Layers of analysis are separated by institutional and infrastructural silos and only in the rarest cases can they stay connected with each other.

• Humanities are a very broad research discipline, many specific research contexts, but also increasingly interdisciplinary research.
What are we talking about when talking about humanities data?
From manuscripts to innovative and unexpected ways to access history

- Philosopher, social and legal reformer
- Digitizing his 40,000 untranscribed folios (in 8 years!)
- https://blogs.ucl.ac.uk/transcribe-bentham/

WikiMedia instance to deliver images to volunteers to transcribe the texts in a machine-readable format (TEI-XML)

Huge success: more than 22,000 manuscripts transcribed, 96% quality checked.
From manuscripts to new ways to access history

- Automatic handwritten text recognition and transcription
- The more people using it the more useful it gets
- Starting out from a digitization project creating a mechanisms by which others can auto transcribe their texts on the large changing how people do history, how people can access history, the questions we might ask from these resources etc.

https://transkribus.eu/Transkribus/
Novel ways in which cultural resources are made available
However.
We should not forget about the cultural knowledge iceberg sunken into an analogue world.

Image source: Harry Verwayen (EUROPEANA), visualization based on the results of the ENUMERATE Survey Report on Digitisation in European Cultural Heritage Institutions.
Gaining access to Cultural Heritage resources can get pretty complex though...

“At noon, we arrived in Speyer, where the chapter had already allowed us access to the archives to compare our copies of documents with the originals. We had also been promised that we would receive further material. However, it took us eight days to find out what we wanted to know. Because here it is like everywhere else, nothing happens without a multitude of difficulties. […] The chairman forced us to dine with him every evening. Only once did we have our peace. We also had to spend every evening with the archivist, who was awarded with a gold medal worth 25 ducats.”

Andreas Lamey, 1769, quoted after Voss 2002: Schöpflin, p. 604.

Checklist to keep in your pocket during your first visit to the (digital) archive

The Cultural Heritage Data Reuse Carter

- A tool to allow allow both Cultural Heritage Institutions, infrastructure providers and researchers and to clarify their goals at the beginning and the project, to specify their exchange protocols, citation and attribution standards, hosting responsibilities.

- To help start the right data conversations

Learn more at: https://www.dariah.eu/activities/open-science/data-re-use/
And my data?
Exercise 1: What are your research data?

- In your discipline?
- In your current project?
- Think of everything that helps the interpretation of your data and your research process!

❖ Form groups
❖ Discuss and note results on sticky note
❖ Bring sticky notes to front

Image source: Men sitting around a table discussing the contents of the piece of paper one of them is holding up. Process print. Credit: Wellcome Collection. CC BY

Your contribution is just as important!
A model of research workflow

- Never as linear as one would expect
- Data sharing should be kept in mind from the beginning
- “Your primary collaborator is yourself from 6 months now and your past self doesn’t answer emails” (Rachel Ainsworth)
Easy to say so...

**Will I be plagiarized?**

What exactly can/should I put online?

How to enable others to follow exactly what I did?

All my mistakes and uncertainties will be visible?

How to find a safe home to my data?

Is it good for my career or am I just giving away my resources?
Leveraging on the open, on the digital

...and making it work for her career advancement!

Naomi Truan
Wissenschaftliche Mitarbeiterin at Uni Leipzig

Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI

All publications available online in HAL and ORTOLANG:
https://cv.archives-ouvertes.fr/naomi-truan

Blogging research experience under:
https://icietla.hypotheses.org
Naomi’s lessons

• Remain **anchored** in your field
  Respects the methods and publication practices of her field
• Get all the **benefits** of being digital and open
  **Sharing is not giving away!**
  Astonished to be cited at LREC
  Encoding practices taken up by the Dutch Language Institute
• And she does even not know what EOSC, FAIR, Plan S and DMP mean...
  But she knows about **source documentation** (AKA meta-data), TEI and CC-BY
• The seed of an ambassadors’ network
Your funder might also have a word or two about data sharing...
Your funder might also have a word or two about data sharing...

Explanation of the FAIR data principles
Wilkinson et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3, doi:10.1038/sdata.2016.18

<table>
<thead>
<tr>
<th>Principle</th>
<th>In other words</th>
<th>Researcher’s responsibility</th>
<th>Requirements to be fulfilled by the repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1. metadata are assigned a globally unique and persistent identifier (PID), for example a DOI, AUK, PID... These identifiers allow to find, cite and track metadata.</td>
<td>Each data set is assigned a globally unique and persistent identifier. Certain repositories automatically assign identifiers to data sets as a service. If not, researchers must obtain a PID via a PID registration service.</td>
<td>Ensure that each data set is assigned a globally unique and persistent identifier. Certain repositories automatically assign identifiers to data sets as a service. If not, researchers must obtain a PID via a PID registration service.</td>
<td>A repository needs to have a predictable way to assign a PID to each component of a dataset (e.g., each file or manuscript), in order to be able to include those identifiers into the corresponding metadata before the submission.</td>
</tr>
<tr>
<td>P2. Data are described with rich metadata (defined by R1 below)</td>
<td>Each data set is thoroughly (see below, in R1) described; these metadata document how the data were generated, under what terms (license) and how it can be reused, and provide the necessary context for proper interpretation. This information needs to be machine-readable.</td>
<td>Fully document each data set in the metadata, which may include descriptive information about the content, quality and condition, or characteristics of the data. Another researcher in any field, or their computer, should be able to properly understand the nature of your dataset. Be as generous as possible with your metadata (see R1).</td>
<td>Allow researchers to upload metadata for each data set.</td>
</tr>
<tr>
<td>P3. Metadata are not stored in the same file as the data they describe.</td>
<td>The metadata and the data set they describe are separate files. The association between a metadata file and the data set is obvious thanks to the mention of the data set’s PID in the metadata.</td>
<td>Make sure that the metadata contains the data set’s PID.</td>
<td>Allow researchers to upload metadata for each data set.</td>
</tr>
<tr>
<td>P4. Metadata are registered or indexed in a searchable resource</td>
<td>Metadata are used to build easily searchable indexes of data sets. These resources will allow to search for existing data sets similarly to searching for a book in a library.</td>
<td>Provide detailed and complete metadata for each data set (see P2).</td>
<td>Request and store part of the metadata in a structured way, for example by providing a form with specific fields to be completed or by providing an XML schema to be used by the researchers. For example the storing of PIDs, author names, disciplines, etc. will facilitate the creation of indexes. However, it must remain possible to provide arbitrary metadata in addition.</td>
</tr>
</tbody>
</table>

Have your heard about the FAIR principles?

http://www.snf.ch/SiteCollectio nDocuments/FAIR_principles_translation_SNSF_logo.pdf
1. **Findable**: easy to find by both humans and computer systems and based on mandatory description of the metadata + a Persistent Identifier that allow the discovery of interesting datasets;

2. **Accessible**: stored for long term such that they can be easily accessed and/or downloaded with well-defined licence and access conditions (Open Access *when possible*), whether at the level of metadata, or at the level of the actual data content;

3. **Interoperable**: ready to be combined with other datasets by humans as well as computer systems 🏞️ standard metadata schemas, vocabularies, ontologies if applicable.

4. **Re-usable**: ready to be used for future research and to be processed further using computational methods 🔄 proper licensing.
DANS's checklist to evaluate FAIRness of datasets

- Lightweight approach
- https://docs.google.com/forms/d/e/1FAIpQLSf7t1Z9I0Boj5GgWqik8KnhtH3B819Ch6lD5KuAz7yn0l0Opw/viewform

Checklist to evaluate FAIRness of data(sets)

You would like to deposit one or several dataset(s) at a digital repository but you are not sure whether the information you provide is sufficient and in line with the principles of FAIR (Findable, Accessible, Interoperable, Reusable). This checklist helps you assess the quality (FAIRness) of your dataset(s) and the trustworthiness of the repository that you have chosen.

The assessment will cover four levels:
1. The data repository you are planning to use
2. The metadata with which you describe your dataset
3. The dataset itself
4. The data files of which your dataset consists

This checklist, furthermore, draws upon two core concepts: that of the trustworthy repository and that of FAIR data. The CoreTrustSeal (CTS) Data Repository Certification (https://www.coretrustseal.org/) is taken as an example for certified trustworthy repositories. Repositories with such a certification are to a large degree already compliant with the FAIR principles. A list of CTS-certified repositories can be found here: https://www.coretrustseal.org/cts-certification-certified-repositories/

More information about FAIR and the principles per character is provided on the website of the Go-FAIR initiative: https://www.go-fair.org/fair-principles/

The checklist consists of 7 sections including a feedback section at the end. The structure of the questions per letter will be as follows:
However, the problem is: by the time you get to this checklist, it’s too late.

Reconstructing FAIR relevant documentation of finished data sets is virtually impossible.
Data management good practices in the research workflow
How to make the whole workflow as transparent and open as possible?
Why Research Data Management 3/4: Analyzing data

https://rdmpromotion.rbind.io/blog/2019/01/24/video-pre-release/
1. Data collection and reuse
How do we find data for reuse?

Discussion:

- When was the last time you used data collected, curated or generated by someone else?
- Where/how did you find it and which factors helped its discoverability?
- How could you access the material in question?
- How did you trust in the creators?
- How the limitations of the data set (incompleteness, uncertainties) were indicated?
- How your research methods affected the collection of your data?
- How the collection of your data affected your research methods?
A data (re)use case study that highlights many of the challenges DH research is facing:


"During vacation, I had an idea for a future project for which data from a database is needed. Some key features and search functions in the text sources of the database are available online for free, but it’s not possible (and maybe illegal, too) to parse it with a script. That is why I made something silly: I told my idea to a responsible person and asked for raw data, plain texts, and license policies. (I will describe the project idea in another post if everything works as expected). Within some days, I received an answer: they will not confer about IF they grant me access, but about HOW they can do it! Let’s dig up the treasure.”
We should not forget about the cultural knowledge iceberg sunken into an analogue world

„The collections we hold, and the subset we can digitise and make available for re-use are only a tiny proportion of what once existed. [...] Some items can’t be digitised because they’re too big, small or fragile for scanning or photography; others can’t be shared because of copyright, data protection or cultural sensitivities. We need to be careful in how we label datasets so that the absences are evident.”

(Mia Ridge)
2. Data processing and analyzing
How will I know how the dataset I’m interested in had been cooked’?

• A major twist in FAIR research culture: the separation of data from its context of creation.

• Explaining how the data had been ‘cooked’: rich provenance metadata (incl. the description of the software environment) is of crucial importance for both cultural heritage professionals.

• Standards: bridges between repositories, enable to bring together isolated data and to give them a richer context, improving their readability.
Data documentation and metadata

• How can you minimize the hassle for other people to find the materials you used and created?

• Your documentation should indicate finding aids and other resources used

• Not everything has to be kept!

• Once you have developed a suitable data model, you are also advised to develop a data dictionary which documents the model.

• This document may contain the following information:
  - a list of all the column names used in the data spreadsheet
  - description of the purpose and the contents of these different columns, explaining abbreviations etc.
How others can make sense of your data?

An example: interview data
• The audio file of the interview
• The interview transcript in the form of a digital text file
• The discussion guide or questionnaire which explains the methodological approach and is necessary for the comprehensibility of the results of the study.
• The project explanation as well as the declaration of consent of the interviewee, which documents compliance with the legal provisions of the GDPR.
• The codebook which e.g. documents the development categories and variables used
• The documentation of the procedure for anonymization and pseudonymization
• The indexing information (metadata), which guarantees the citability of the interview and its findability.

File naming conventions

Looks familiar?

Image source: Stanford Library
File naming conventions

The specifics usually matter less than just having some.

Common elements (UK Data Service):
- Version number
- Date of creation (date format should be YYYY-MM-DD);
- Name of creator;
- Description of content;
- Name of research team/department associated with the data;
- Publication date;
- Project number.
Can I run your data on my tools?

I sent you the data.

Thanks!

...This is a Word document containing an embedded photo you took of your screen with the spreadsheet open.

Yeah? Does your computer not support .odm files? Maybe you need to update.

Since everyone sends stuff this way anyway, we should just formalize it as a standard.

Image source: Stanford Library
Open formats

Formats preferred by the DANS repository. See the full list here: https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/before-depositing/file-formats
5 star development scheme for Open Data

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data. Here, we give examples for each step of the stars and explain costs and benefits that come along with it.

★ make your stuff available on the Web (whatever format) under an open license

★★ make it available as structured data (e.g., Excel instead of image scan of a table)

★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)

★★★★ use URIs to denote things, so that people can point at your stuff

★★★★★ link your data to other data to provide context

Source: https://5stardata.info/en/
 Ontologies, vocabularies

• Put structure on your messy data so that it opens up to others (people, machine, applications, related databases).
• Check whether some of the general topics and terms (persons, locations, concepts) that you focus on have already been assigned persistent identifiers or URIs in one of the ontologies that are relevant for your field.
• Ontologies are one of the ways in which we can make datasets interoperable,

The Basel Register of Thesauri, Ontologies & Classifications (BARTOC)
The CLARIN Concept Registry and the DARIAH/ACDH collection of vocabularies.
‘An Ontologist and a Data Scientist walk into a bar: Data in Research Projects’

https://www.youtube.com/watch?v=WNG1iLB4KtA&index=1&list=PLKq1g7snsFGc7f1_Aidypmz62d7i6Uh4x
Metadata standards in the Arts and Humanities (teaser)

• **TEI** (Text Encoding Initiative): www.tei-c.org
• **CEI** (Charter Encoding Initiative): http://www.cei.lmu.de/index.php
• **MEI** (Music Encoding Initiative): https://music-encoding.org/
• **CMDI** (Language Resources, CLARIN):
• **IIIF** (International Image Interoperability Framework): https://iiif.io/
• **EAD** (Encoded Archival Description, for finding aids): https://www.loc.gov/ead/
• **Dublin Core** (description of digital documents): http://dublincore.org/

The choice of appropriate standards is more of a community issue than a technical one.

An overlay platform dedicated to promoting a wider use of standards within the Arts and Humanities:

- Documenting existing standards by providing reference materials.
- Fostering the adoption of standards.
- 18 scenarios: Heritage science scenarios + “traditional” DH ones → Living memory of best practices
- Developed within the framework of the EU project PARTHENOS:

http://ssk.huma-num.fr/#/
Linking datasets with publications

Exercise:

Work in groups.
Consider the following three articles.
• To what extent can the data sets that are mentioned in the articles be accessed?
• Are the data sets also in preferred formats?
• Which kinds of additional documentation would further increase their accessibility and reusability for other disciplinary communities?

Give a passport with your data – prepare a readme file

- A readme file provides information about a data file and is intended to help ensure that the data can be correctly interpreted, by yourself at a later date or by others when sharing or publishing data.
- Standards-based metadata is generally preferable, but where no appropriate standard exists, for internal use, writing “readme” style metadata is an appropriate strategy.

Exercise:

1. Go to https://cornell.app.box.com/v/ReadmeTemplate where you will find a readme template

2. What are the main components of the document?

3. Are these well-aligned with your research processes and data? How would you adapt it for your own research?

Source: Cornell University, Research Data Management Service Group. https://data.research.cornell.edu/content/readme

3. Data sharing, storing and publication
Data sharing and storing

With collaborators while research is active

- Likely to be on a networked filestore or
- Central institutional file share
- Easy to change or delete

(Open) data sharing

- Institutional, disciplinary or generic repository

Data are mutable

Data are stable, searchable, citable, clearly licensed

Adopted from: Marjan Grootweld: ‘What do you need to know to prepare a good Data Management Plan?’ CC-BY 4.0
Exercise

• The format of the metadata is often prescribed by the data repository which will manage the data set.
• Compare the metadata fields that need to be completed at a Zenodo upload with the discipline-specific requirements of DANS EASY.

OR: https://tinyurl.com/DANSmetadata

How to select a repository that best fits your research?

What are your governing criteria when selecting a repository?
How to select a repository that best fits your research?

1. Use an external data archive or repository already established for your research domain to preserve the data according to recognised standards in your discipline.

2. If available, use an institutional research data repository, or your research group’s established data management facilities.

3. Use a cost-free data repository such as Zenodo.

4. Search for other data repositories here: re3data.org.

Source: https://www.openaire.eu/opendatapilot-repository-guide

Image adapted from Marjan Grotteweld’s presentation But can I trust your data?
The subsequent use of data requires more knowledge than the collection of new data.
I often feel unsure about the reuse conditions of Cultural heritage data that are relevant for me.
It is not easy to apply standards to my work.
The publication of research data does not contribute to building a reputation.
The management and publication of research data causes costs, which I can’t carry.
If I publish my research data, somebody might scoop me and publish findings based on my data.
When I publish my research data, my research becomes completely transparent and even the smallest errors become apparent.
My research belongs to me!

Ready, set, explore, reuse!


DARIAH Open blog
https://openmethods.dariah.eu/

Parthenos Training Suite
https://training.parthenos-project.eu/

OpenMethods platform
https://openmethods.dariah.eu/


The Standardization Survival Kit
http://ssk.huma-num.fr/#/
Further useful resources
What is a Data Management Plan?

• A data management plan is designed to encapsulate & articulate details about data from collection to curation to preservation to dissemination to destruction.

• A data management plan should be an ongoing process rather than a level of grant requirement for a funding agency program solicitation.

• Who are the involved parties and what are the responsibilities of each of them?
• What kind of support can you seek in your institution?
An example:

This rubric is designed as a checklist or marking aid for those reviewing data management plans for submission to the Arts and Humanities Research Council (AHRC).

<table>
<thead>
<tr>
<th>Performance Criteria</th>
<th>Detailed</th>
<th>Performance Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Section 3</strong></td>
<td>The data storage for the project is clearly described, covers all the data to be stored and is suitable, so far as it is possible to judge. The plan may also reference institutional storage policies or pages.</td>
<td>There is some description of the data storage solution the project will use but it is not clearly described or does not cover all of the data being produced.</td>
</tr>
<tr>
<td><strong>Section 3a</strong></td>
<td>The back-up process for active data storage is clearly described or referenced, and is appropriate for the data to be collected.</td>
<td>Data backup is mentioned, but no detail, or link to institutional policy is provided. Back-up process described might be inadequate for the data being collected and stored.</td>
</tr>
<tr>
<td><strong>Section 4a</strong></td>
<td>The long-term storage plan for the data is described. This might be a repository or other appropriate solution. The solution(s) identified cover all the data to be retained.</td>
<td>A long-term storage plan is mentioned, but detail may be lacking or the solution(s) identified may not cover all the data to be retained.</td>
</tr>
<tr>
<td><strong>Section 4b</strong></td>
<td>The long-term retention schedule is described for all data. The retention period is appropriate to the data and in keeping with any consent from participants.</td>
<td>The long-term retention schedule is mentioned, but may not cover all data or may be inadequate or inappropriate.</td>
</tr>
<tr>
<td><strong>Section 4c</strong></td>
<td>The costs for the long-term storage are clearly described. Alternatively, it is stated that the data will be stored for the long-term in a repository with no ingest costs.</td>
<td>No long-term retention schedule is mentioned.</td>
</tr>
<tr>
<td><strong>Section 5a</strong></td>
<td>The value to all relevant disciplinary areas have been clearly outlined for each data type. Consideration has been given to the different types of value data can provide and these are described appropriately. It is</td>
<td>The value of the data to the disciplinary area is mentioned but it is not clear or may be poorly explained. There may be missing details about which disciplines may benefit from this</td>
</tr>
<tr>
<td><strong>Section 5b</strong></td>
<td></td>
<td>The value of the data to the disciplinary community is not mentioned.</td>
</tr>
</tbody>
</table>
Use **Tools for Data Management Planning**

e.g. DCC DMPOnline

[https://dmponline.dcc.ac.uk/](https://dmponline.dcc.ac.uk/)
Research Data Management Organiser

https://rdmorganiser.github.io/en/
How do we find data for reuse?

Discussion:

• Which kinds of entities/digital objects can have PIDs?
• How many PID types and providers can you name?
• How many of them you use?
• Is there a shortage of PIDs in the humanities domain?
How do we find data for reuse?

- **Persistent Identifiers (PIDs)** ensure that online references to publications, research data, and persons remain stable and available in the future *even if their location changes*.

- A PID is a specific type of a [Uniform Resource Identifier (URI)](https://en.wikipedia.org/wiki/Uniform_Resource_Identifier), which is managed by an organisation that links a persistent identification code with the most recent Uniform Resource Locator (URL).

- **Many functions**: disambiguation, citability (humans, machines), linking, instrumental in the creation of knowledge graphs etc.
How do we find data for reuse?

- PIDs and granularity:

  ark:/12148/btv1b8449691v

  ark:/12148/btv1b8449691v/f29

  urn:cts:greekLit:tlg0012.tlg001.perseus-grc1.1.1–1.10

  Q7245 ▷ go to: https://evelin.ifi.uni-heidelberg.de/

  731081 ▷ go to: https://openknowledgemaps.org/viper/

  10.14293/S2199-1006.1.SOR-UNCAT.CL49QV3.v1

  hal-01836189