

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

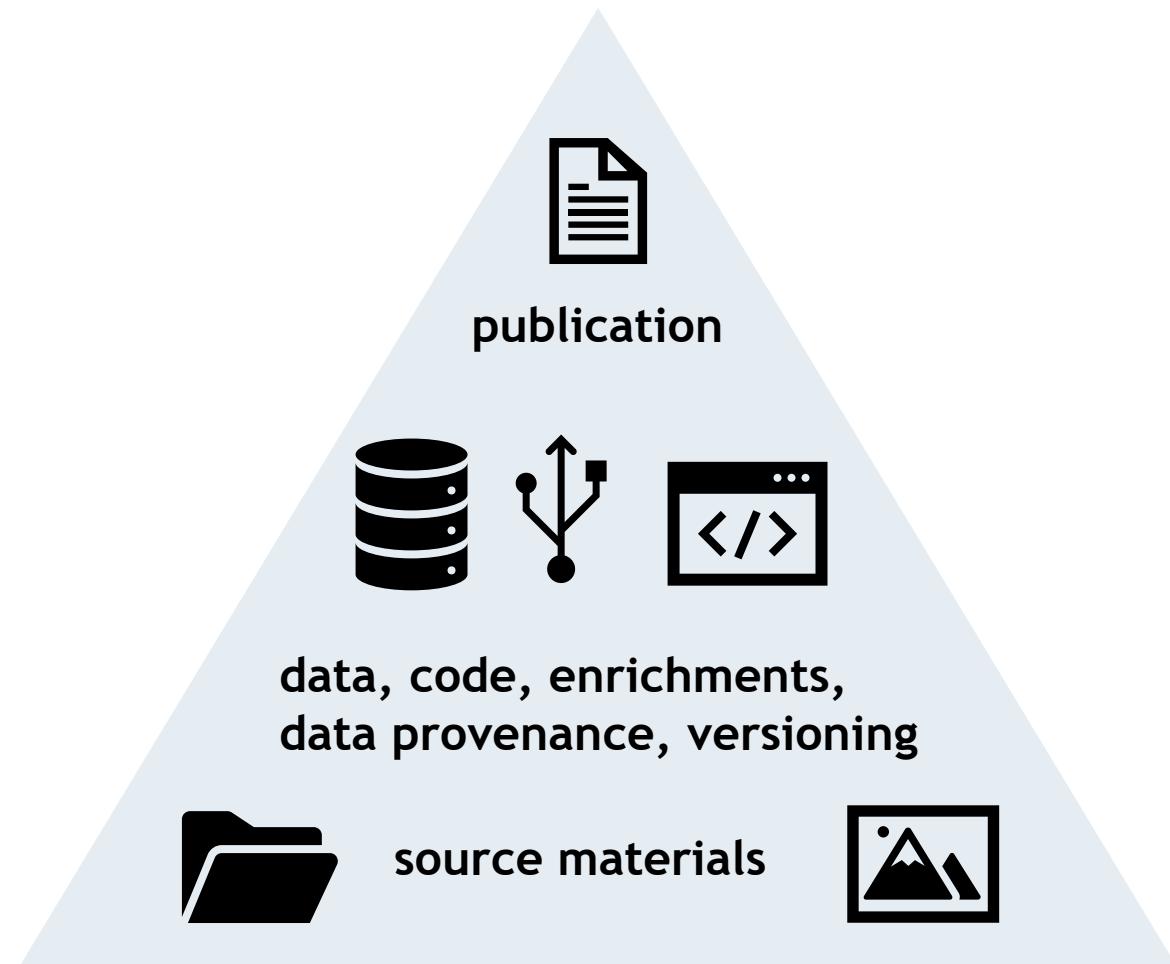


DESIR Winter School, Lisbon, 10.12.2019



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)



Miriam Posner
@miriamkp

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?

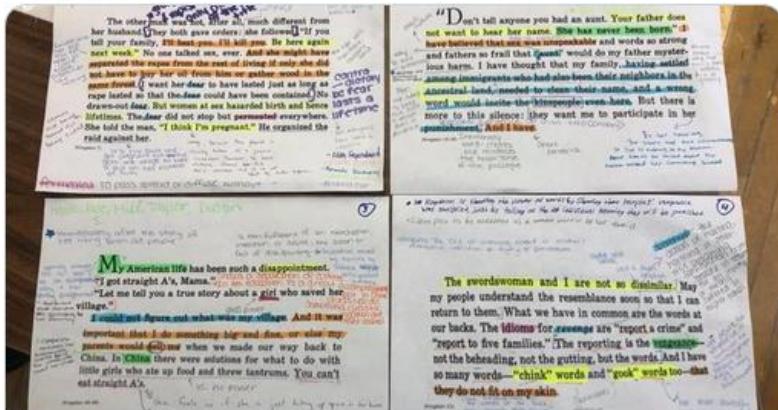
7:50 PM · Oct 31, 2018 from Los Angeles, CA · Twitter Web Client

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.



Collaborative Close Reading

Close reading - observing the stylistic details of a text in order to analyze an author's use of language - is a skill taught in almost all college

Data in the arts and humanities: still a dirty word?

 **Adam Arenson** @adamarenson · Oct 31, 2018
Replies 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 argues for its importance/relevance.

1 6 1

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

1 1 1

 **Matthew DeForrest** @mmdeforest · Nov 1, 2018
Replies 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.

1 4 1

 **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.

1 1 1

 **Jentery Sayers** @jenterysayers · Nov 1, 2018
Replies 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.

1 1 10

<https://twitter.com/miriamkp/status/1057706465866133504>

Running order

1. Introduction to Research Data Management

Identifying humanities research data

Excercise: 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

Excercise: How do we find data for reuse?

Excercise: Data citation

2. Data processing and analyzing

3. Data sharing, storing and publication

Excercise: The networked publication

Excercise: How to find a suitable repository for your research?

Discussion: Pick a statement.



Image source: [Pixabay](#), CC0.

What do you see?

• • * « • •
г .-,;....'
ун ь в«п
Singen, ше
1фе ben 1 bienft angelen*
л.-,, , ... '/
4 * • • ' . | >*
ГО ЮксствНг, и КЕЦИ^Х, гаже Кж'Г/л
СЛЭЖЕЫ КАСАУТСА.
| •
&ott, гÓТТТU Brü, ®
ie ©oft^eif. Бжккто, Божккво. •;
/ ^ „,-
ди Готт^литя.. Нч ©
oft beiíSoafec, Bors \
f(iÄ Флтгря, КГА GN'Ä ,
вес €>офт. БОГА вы НZ.
Тотч A'f* GO'MI. Вгх Д\х стыи« ©
оти вес) fällige ©eifi,
ГОТТА Д\р А ХАЙЛИП
ГАЙСТА.;, -, GT(ÍATfl}A,
ie ſetitge SDrdfaíigf ¿ti, ТРОИЦА. '
Г\ДИЛИГ4 * --уj--'
ТИГКАЙХ.

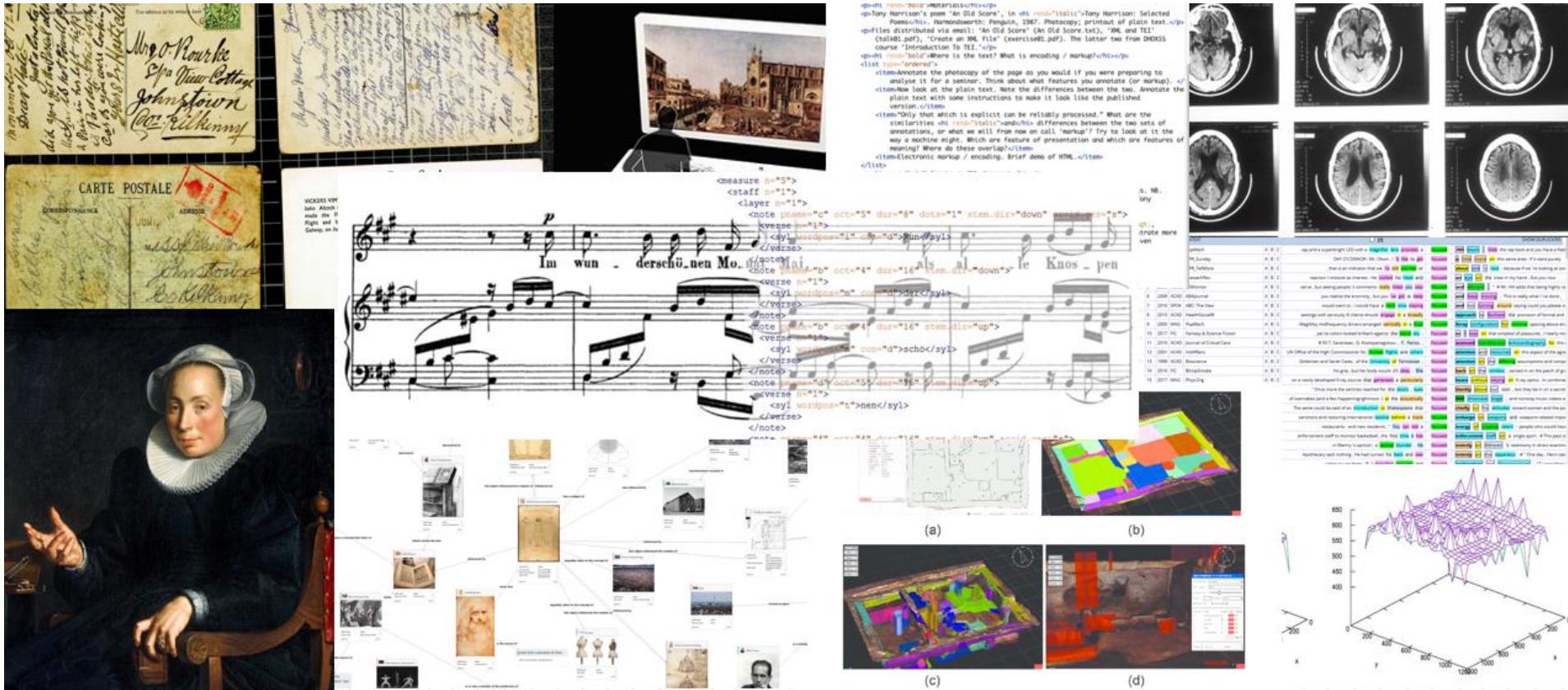
What do you see?

<p>Von der Gottheit, und den Dingen, welche den Gottes- dienst angehen.</p> <p>О Бжествѣ, и вешчѣ, таже Бжїл слѹжбы касаютса.</p> <p>Gott, Гóттъ. Die Gottheit. дн Гóттхайнцъ. Gott der Vater, Гóттъ дѣрѣа Фатеръ. Gott der Sohn. Гóттъ дѣрѣа Сонъ. Gott der heilige Geist, Гóттъ дѣрѣа Хайлігт. Гóттъ дѣрѣа Трайфал- тигкайтъ.</p> <p>Бѓз, Бѓгъ. Бжество, Бжество. Бѓз Сўз, Бѓгъ Стіцъ. Бѓз Сынъ, Бѓгъ Сынъ. Бѓз Ахъ сватый. Бѓгъ Трайца, святѣа</p>	<p>... * « * ... и нь в«п Singen, ше 1фе ben 1 bienft angelen* Л.-,, , ... / 4 * * * . >* ГО ЮжествНг, и КЕЦИ^Х, гаже ЮжГ/л СЛЭЖЕЫ КАСАУТСА. I . &ott, rÓTTTU Brü, ® ie ©oft^eif. Бжкто, Божкто. ; / ^ „ - ди Готт^литя.. Нгч © oft beißSoafec, Bors \ f(iÄ Флтгря. КГА GN'Ä , bec €>офф. БОГА вы НZ. Тотч A'f* GO'MI. Вrix Дх стыи« © oti bec) fällige ©eiffi, ГОТТА Д р А ХАЙЛИП ГАЙСТА.; - , GT(ІATfl)A, ie Šetitge SDrdfaüigf չtí, ТРОИЦА. ІДИЛИГ4 * --у--'- ТИГКАЙТХ.</p>
---	--

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 opublishers.
- 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

[UCL Home](#) » [Transcribe Bentham](#) »



Transcribe Bentham

Welcome to Transcribe Bentham!

By Louise Seaward, on 6 December 2017



'Many hands make light work. Many hands together make merry work', wrote the philosopher and reformer, Jeremy Bentham (1748 – 1832) in 1793.

In this spirit, we cordially welcome you to *Transcribe Bentham*, a double award-winning collaborative initiative which is crowdsourcing the transcription of Bentham's previously unpublished manuscripts.

- [Home](#)
- [Transcription Desk](#)
- [Blog](#)
- [About Us](#)
- [People](#)

Anyone can start transcribing at our [Transcription Desk](#). Your transcripts will contribute to the production of Bentham's [Collected Works](#) and preserve Bentham's writings into the future.

Recent Posts

- [Transcription Update – 22 November 2019](#)
- [Transcription Update – 31 October 2019](#)
- [Transcription Update – 30 September 2019](#)
- [Preliminary download of Jeremy Bentham, Writings on Political Economy, Volume IV: Circulating Annuities and other writings on National Debt](#)
- [Transcription Update – 30 August 2019](#)
- [Transcription Update – 31 July 2019](#)

Enter your email address

- Philosopher, social and legal reformer
- Digitizing his 40.000 untranscribed folios (in 8 years!)
- <https://blogs.ucl.ac.uk/transcribe-bentham/>



Transcribe Bentham
A Participatory Initiative

[UCL Home](#) » [Transcribe Bentham](#) » [Transcription Desk](#)

[Create Account](#) [Log In](#)

★ Keep up to date with the latest news - subscribe to the [Transcribe Bentham newsletter](#); Find a new page to transcribe in our list of [Untranscribed Manuscripts](#)

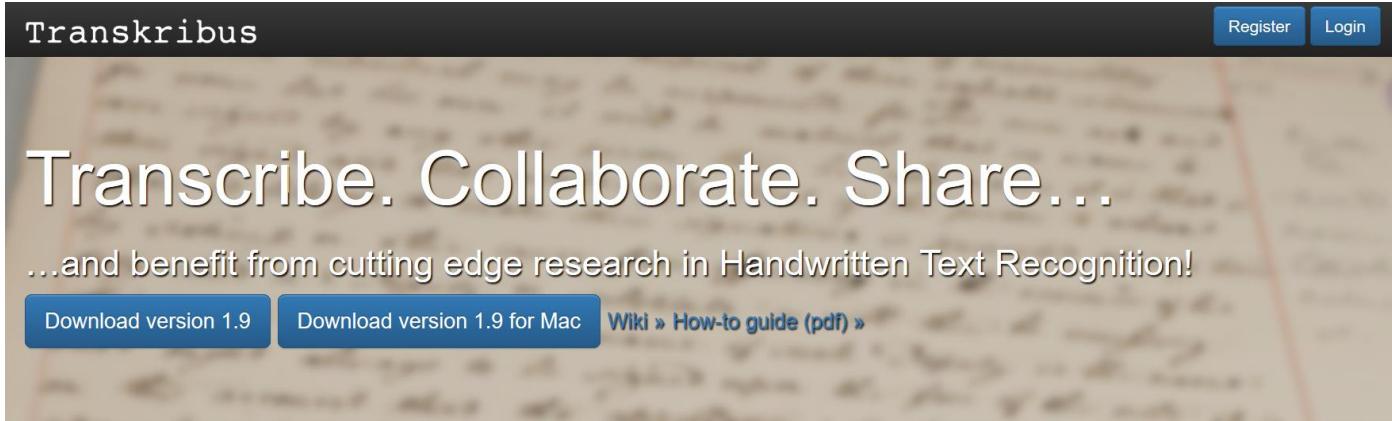
JB/035/064/002

[Click Here To Edit](#)

be given? If the Judge has been for it, if the cause happens to attract his attention, if it be early in the day before he is broken with fatigue — if he is that sort of character, by no means equally abundant with incorruption in English judicature in whom zeal for justice preponderates over personal prudence. Observe that responsibility between these two stools responsibility falls to the ground. The Judge is not responsible for the verdict, for it is not his verdict. The Jury are not responsible for it, for it is of their opinion to be altogether irresponsible. When the Judge has summed up the evidence, that is led say read to the Jury what has been said to them already which yet is no he has done every thing that he is obliged to do,

- 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



The screenshot shows the Transkribus homepage. At the top, there's a dark header bar with the word "Transkribus" on the left and "Register" and "Login" buttons on the right. Below the header is a large image of a historical manuscript page. Overlaid on the image is the text "Transcribe. Collaborate. Share...". Underneath this, it says "...and benefit from cutting edge research in Handwritten Text Recognition!". At the bottom of the main section are four buttons: "Download version 1.9", "Download version 1.9 for Mac", "Wiki", and "How-to guide (pdf)".

Scholars

Are you transcribing historical documents? Handwritten or printed, from the middle ages or from the 20th century? Would you like to do this in a highly standardized, flexible and reliable way? And do you appreciate to get support from automated tools such as Handwritten Text Analysis?

Archives

Are you responsible for large collections of handwritten and printed documents? Do you believe that digitisation paves the way to realise new opportunities to access, enrich and explore archival material? And are you open to involve humanities scholars and volunteers so that they can work with these documents?

Volunteers

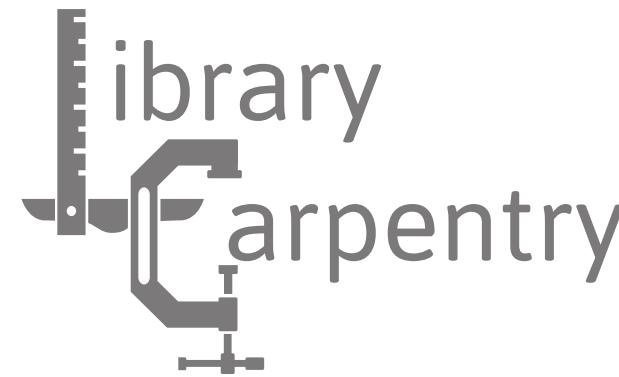
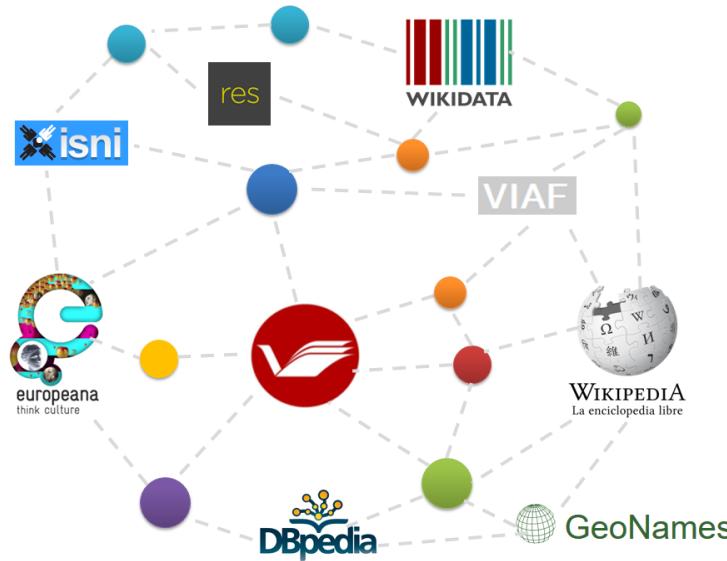
Are historical letters, postcards, manuscripts or medieval documents fascinating for you? Do you enjoy deciphering handwriting – this wonderful feeling when you can read something which may be hidden to most other people? And do you believe that everyone can make a valuable contribution to

Scientists

Are you a computer scientist and working in the fields of computer vision, document analysis, pattern recognition, natural language processing or a related field? You are seeking interesting documents from 1000 years of handwriting, printing and publishing? And you would really enjoy to get some reference data in a

- 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.

Novel ways in which cultural resources are made available



However.

We should not forget about the cultural knowledge iceberg sunken into an analogue world

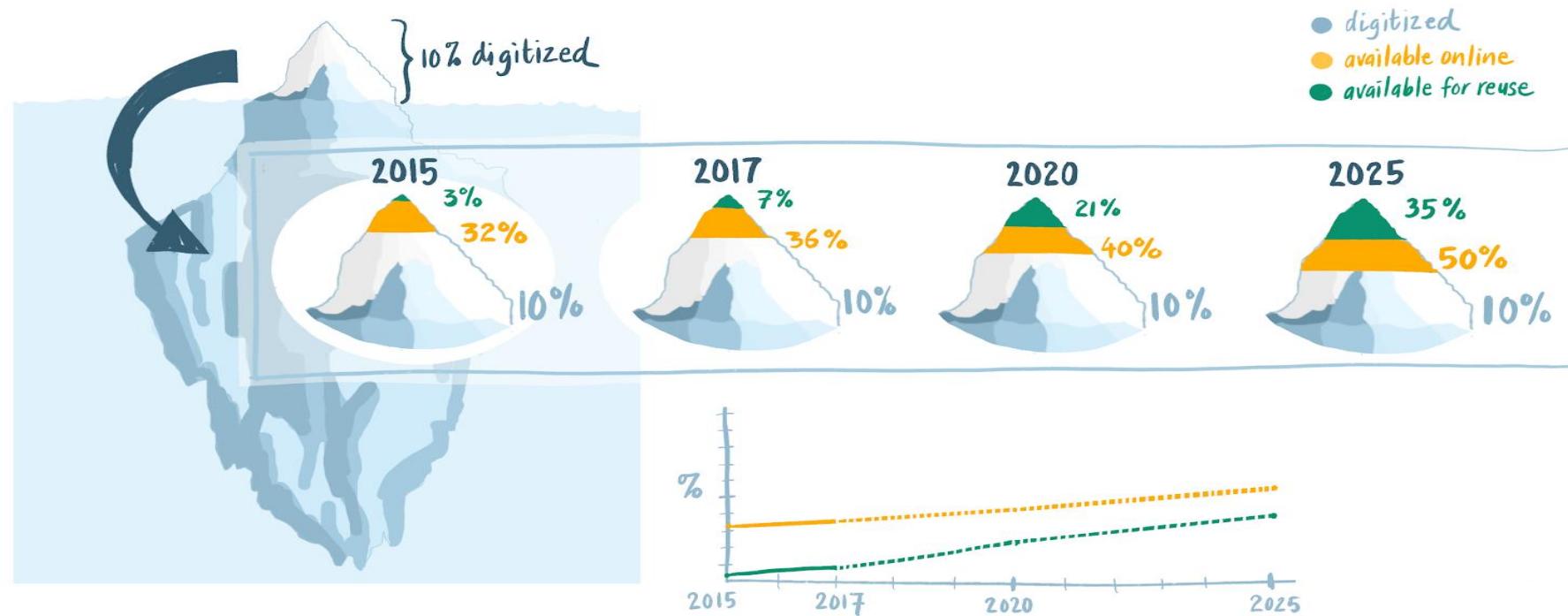


Image source: Harry Verwayen (EUROPEANA), visualization based on the results of the [ENUMERATE Survey Report on Digitisation in European Cultural Heritage Institutions](#).

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.



Source: Henrike Hoffmann, Moritz Mähr and Daniela Zetti: How the search engine came to the historians? Presentation slide from the sharing the experience: [Workflows for the Digital Humanities workshop](#), Neuchatel, 06.12.2019.

December 10, 2018

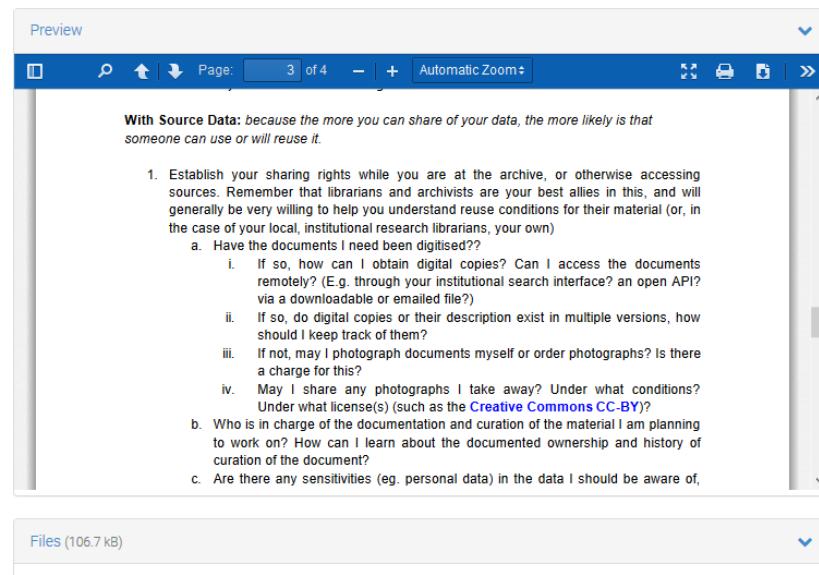
Other Open Access

Open Data for Humanists, A Pragmatic Guide

Edmond, Jennifer; Tóth-Czifra, Erzsébet

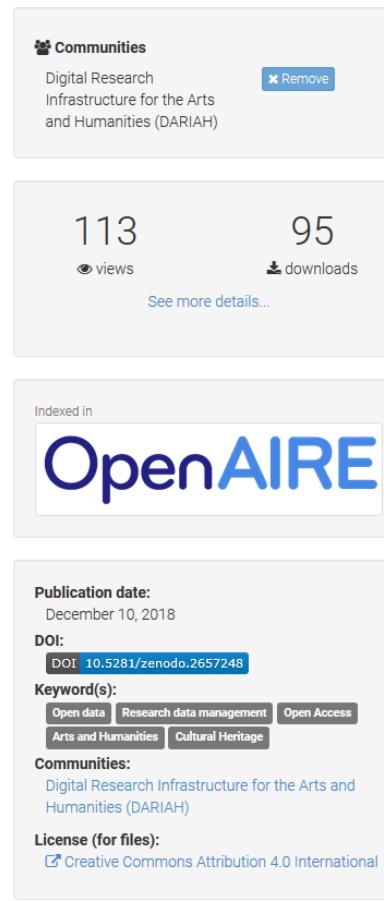
In the arts and humanities, digital data production is still expensive, challenging and time-consuming. We all know this, and yet the results of these processes often in the end can't be reused by other researchers, meaning that we reinvent (or redigitise) the wheel far too often. This resource is aimed at giving practical advice for arts and humanities scholars who are willing to take their first steps in research data management but don't know where to begin. Our approach to data management views it as a reflective process that exposes and tweaks existing behaviours, rather than one that introduces specific tools. It is intended to encourage awareness of one's own processes and mindfulness about how they could be more open and how and how small changes across three points in your research workflow can make big differences.

Preview



With Source Data: because the more you can share of your data, the more likely is that someone can use or will reuse it.

1. Establish your sharing rights while you are at the archive, or otherwise accessing sources. Remember that librarians and archivists are your best allies in this, and will generally be very willing to help you understand reuse conditions for their material (or, in the case of your local, institutional research librarians, your own)
 - a. Have the documents I need been digitised?
 - i. If so, how can I obtain digital copies? Can I access the documents remotely? (E.g. through your institutional search interface? an open API? via a downloadable or emailed file?)
 - ii. If so, do digital copies or their description exist in multiple versions, how should I keep track of them?
 - iii. If not, may I photograph documents myself or order photographs? Is there a charge for this?
 - iv. May I share any photographs I take away? Under what conditions? Under what license(s) (such as the Creative Commons CC-BY)?
 - b. Who is in charge of the documentation and curation of the material I am planning to work on? How can I learn about the documented ownership and history of curation of the document?
 - c. Are there any sensitivities (e.g. personal data) in the data I should be aware of,



Communities

Digital Research Infrastructure for the Arts and Humanities (DARIAH) ✖ Remove

113 views 95 downloads

See more details...

Indexed in

OpenAIRE

Publication date: December 10, 2018

DOI: DOI 10.5281/zenodo.2657248

Keyword(s): Open data, Research data management, Open Access, Arts and Humanities, Cultural Heritage

Communities:

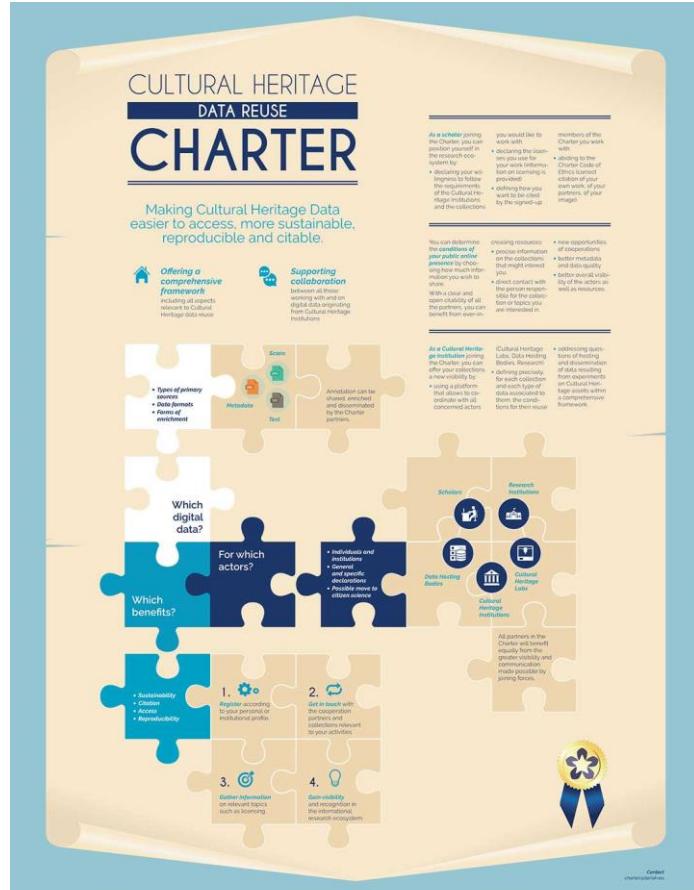
Digital Research Infrastructure for the Arts and Humanities (DARIAH)

License (for files): Creative Commons Attribution 4.0 International

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

Open Data for Humanists, A Pragmatic Guide

Edmond, Jennifer, & Tóth-Czifra, Erzsébet. (2018, December 10) <http://doi.org/10.5281/zenodo.2657248>



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?

Excercise 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!

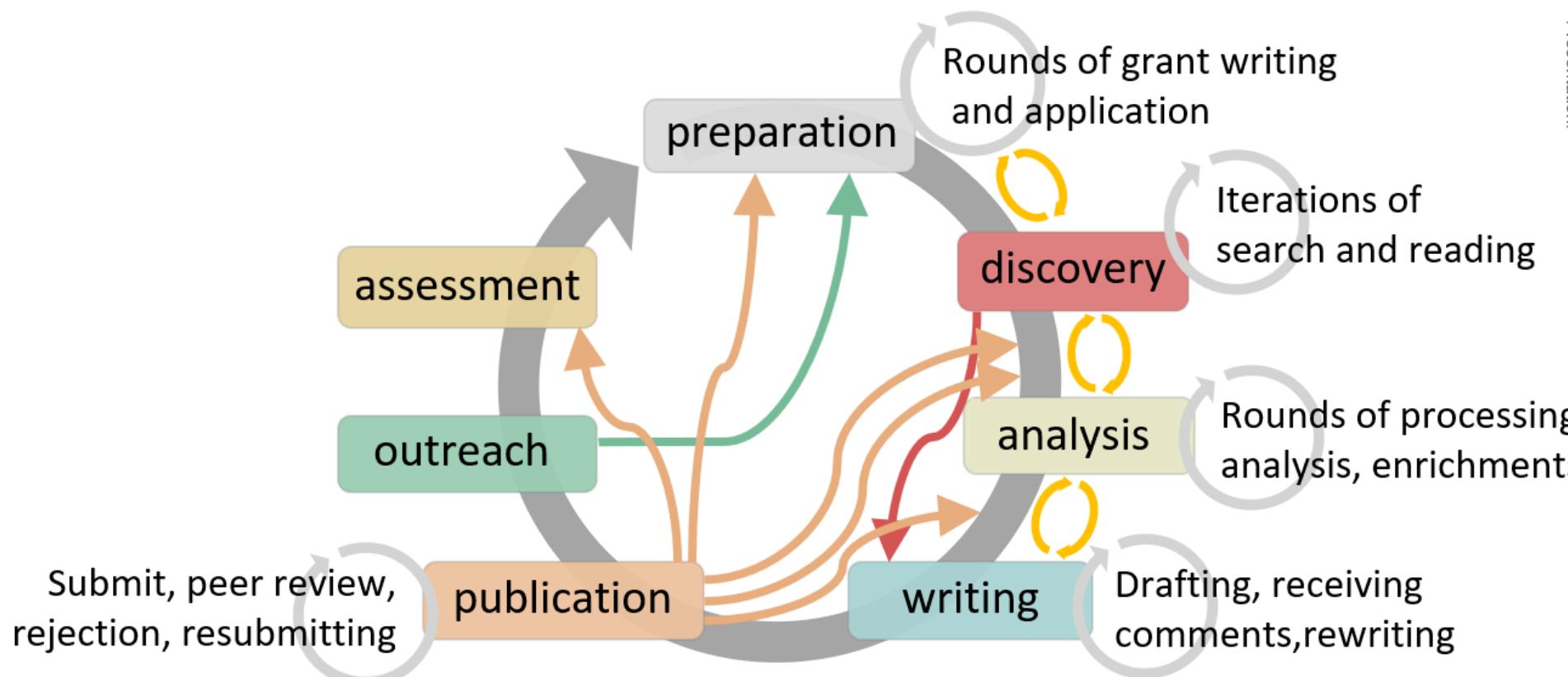
- ❖ Form groups
- ❖ Discuss and note results on sticky note
- ❖ Bring sticky notes to front

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?

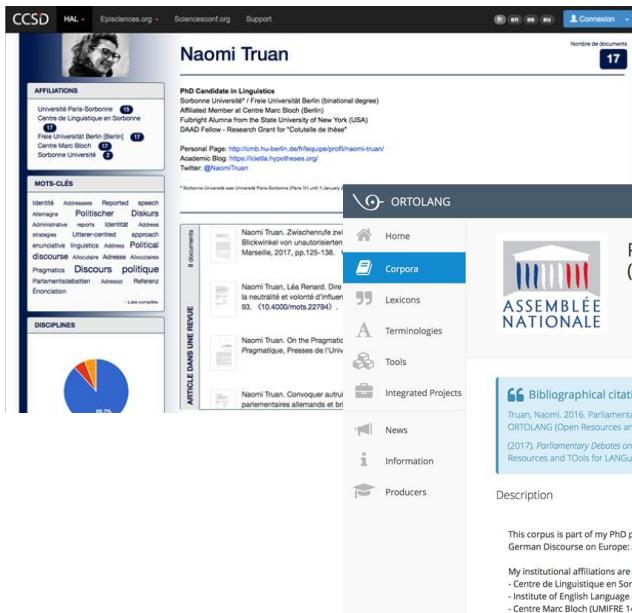
What exactly can/should I put online?

All my mistakes and uncertainties will be visible?

How to find a safe home to my data?

How to enable others to follow exactly what I did?

Is it good for my career or am I just giving away my resources?



CCSD HAL · Sciencesconf.org · Support · Connection · Nombre de documents 17

Naomi Truan

AFFILIATIONS
Université Paris-Sorbonne (13)
Centre de Linguistique en Sorbonne (11)
Free University Berlin (Berlin) (11)
Gesellschaft für Sprache und Kultur (GSK) (11)
Sorbonne Université (9)

MOTS-CLÉS
Identité · Addresses · Reported speech ·
Allemagne · Politischer Diskurs ·
Administrative reports · Identität · Address
discourses · Utterance-centered approach ·
document analysis · Discourse analysis ·
discourse · Address · Annotates ·
Pragmatics · Discours politique ·
Parlementariedaten · Adressat · Reference
Envoiement · Article dans une revue ·
Language

DISCIPLINES
Linguistics (1)

ARTICLE DANS UNE REVUE
Naomi Truan, Convoyer autres parlementaires allemands et français (2017)

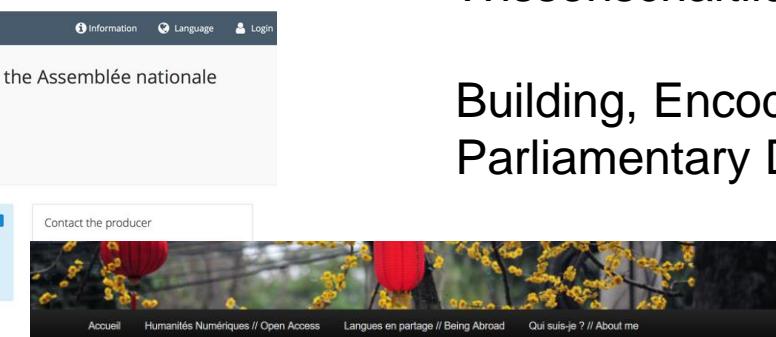
PDF Candidate in Linguistics
Sorbonne Université / Freie Universität Berlin (bilingual degree)
Affiliate Member at Centre Marc Bloch (Berlin)
Fulbright Alumni from the State University of New York (USA)
DAAD Research Grant for "Cvoleute de thèse"
Personal Page: <http://icetla.hu-berlin.de/francaise/prof/naomi-truan/>
Academic Blog: <https://icetla.hypotheses.org/>
Twitter: @NaomiTruan

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

Blogging research experience under:
<https://icetla.hypotheses.org>

Leveraging on the open, on the digital

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



ORTOLANG · Information · Language · Login

ASSEMBLÉE NATIONALE

Parliamentary Debates on Europe at the Assemblée nationale (2002-2012)

Bibliographical citation
Truan, Naomi. 2016. Parliamentary Debates on Europe at the Assemblée nationale (2002-2012) [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage). <https://hdl.handle.net/11403/fr-parl>. (2017). Parliamentary Debates on Europe at the Assemblée nationale (2002-2012) [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr. <https://hdl.handle.net/11403/fr-parl/v1>.

Description

This corpus is part of my PhD project entitled: "Representations of the Other in the British, French and German Discourse on Europe: A Corpus-Based Contrastive Discursive Analysis".

My institutional affiliations are the following:
- Centre de Linguistique en Sorbonne (CeLSO) - EA 7332 (Université Paris-Sorbonne)
- Institute of English Language and Literature (Freie Universität Berlin)
- Centre Marc Bloch (UMIFRE 14 CNRS-MAEE; USR 3130)

Contact the producer

Accueil · Humanités Numériques // Open Access · Langues en partage // Being Abroad · Qui suis-je ? // About me

Publié le 03/07/2017 · Précédent · Suivant ·

**Je te donne, tu me donnes, nous
nous donnons... nos données**

J'ai eu l'occasion, le 1^{er} juin 2017, de présenter en une minute mes réflexions au cours de la table ronde « **Publier à l'ère numérique** » co-organisée par Marc Bloch, DARIAH-EU, le bureau de la coopération universitaire de l'Université d'Allemagne à Berlin et l'Université franco-allemande (UFA) – merci Laurent Romary pour l'invitation ! Bien que nos interventions aient voulus ici approfondir les trois grands points que j'ai développés de souligner, une fois encore, à quel point le mouvement de l'accès (open access) représente une avancée certaine pour l'état de nos connaissances.

Naomi Truan @BerLinguistin · Jul 25

For all the corpus linguists interested in sharing their data (annotated corpora and metadata information):

Check the “Open Resources and TOols for LANGuage” (ORTOLANG): [ortolang.fr](http://www.ortolang.fr)

The infrastructure is supported by @CLARINERIC and @Huma_Num. #CLconf2019



1



5



11



Show this thread

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...



H2020 Programme

Guidelines on
FAIR Data Management in Horizon 2020



**European
Research
Council**



European Research Council
Scientific Council

Established by the European Commission

[Open Research Data and Data Management Plans](#)

Information for ERC grantees

by the ERC Scientific Council

Your funder might also have a word or two about data sharing...

www.snf.ch
Wildhainweg 3, P.O. Box 8232, CH-3001 Berne

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](https://doi.org/10.1038/sdata.2016.18)

Principle	In other words	Researcher's responsibility	Requirements to be fulfilled by the repository
To be findable: Data and metadata should be easy to find by both, humans and computer interesting data sets and services. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.	F1. (meta)data are assigned a globally unique and persistent identifier (PID), for example a DOI , ARK , RRID ... These identifiers allow to find, cite and track (meta)data.	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.	A repository needs to have a predictable way to assign a PID to each component of a dataset (e.g. each file or nanopublication), in order to be able to include these identifiers into the corresponding metadata before the submission.
	F2. data are described with rich metadata (defined by R1 below)	Fully document each data set in the metadata, which may include descriptive information about the context, 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).	Allow researchers to upload metadata for each data set.
	F3. metadata clearly and explicitly include the identifier of the data it describes	Make sure that the metadata contains the data set's PID.	Allow researchers to upload metadata for each data set.
	F4. (meta)data are registered or indexed in a searchable resource	Provide detailed and complete metadata for each data set (see F2).	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 PID's, author names, disciplines, etc. will facilitate the creation of indexes. However, it must remain possible to provide arbitrary metadata in addition.

Have you heard about the FAIR principles?

[http://www.snf.ch/SiteCollection
Documents/FAIR_principles_translation_SNSF_logo.pdf](http://www.snf.ch/SiteCollection/Documents/FAIR_principles_translation_SNSF_logo.pdf)

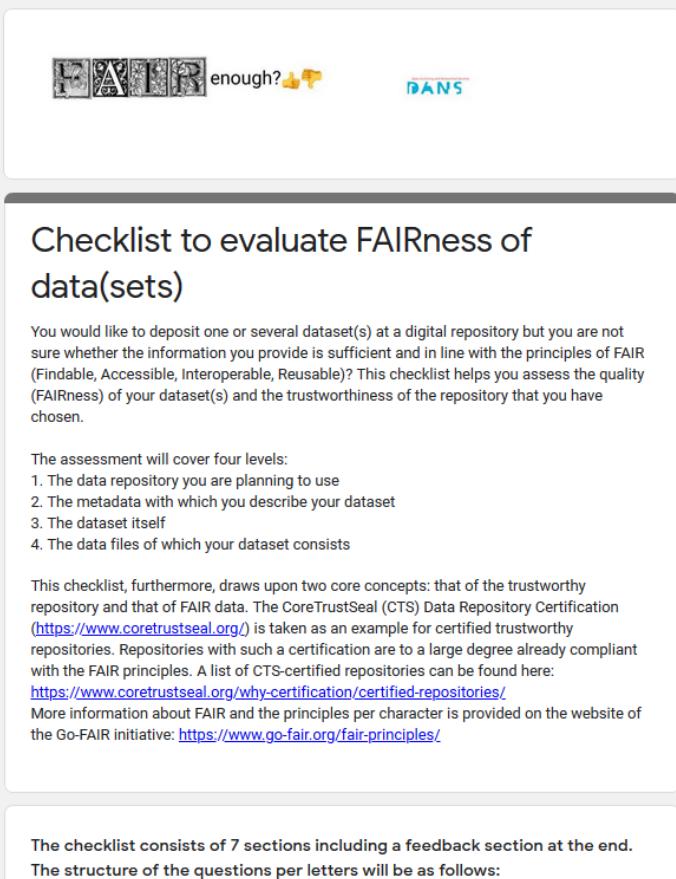
FAIR principles in a nutshell...

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;
1. **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.



Image source: [TIB blog](#), CC-BY 4.0

DANS's checklist to evaluate FAIRness of datasets



The screenshot shows a Google Form titled "Checklist to evaluate FAIRness of data(sets)". The form includes sections for dataset description, repository assessment, and a feedback section. It features the DANS logo and a "FAIR enough? 🙌" header.

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/why-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 letters will be as follows:

- Light-weight approach
- <https://docs.google.com/forms/d/e/1FAIpQLSf7t1Z9lOBoj5GgWqik8KnhtH3B819Ch6ID5KuAz7yn0l0Opw/viewform>

However, the problem is: by the time you get to this checklist, it's too late.

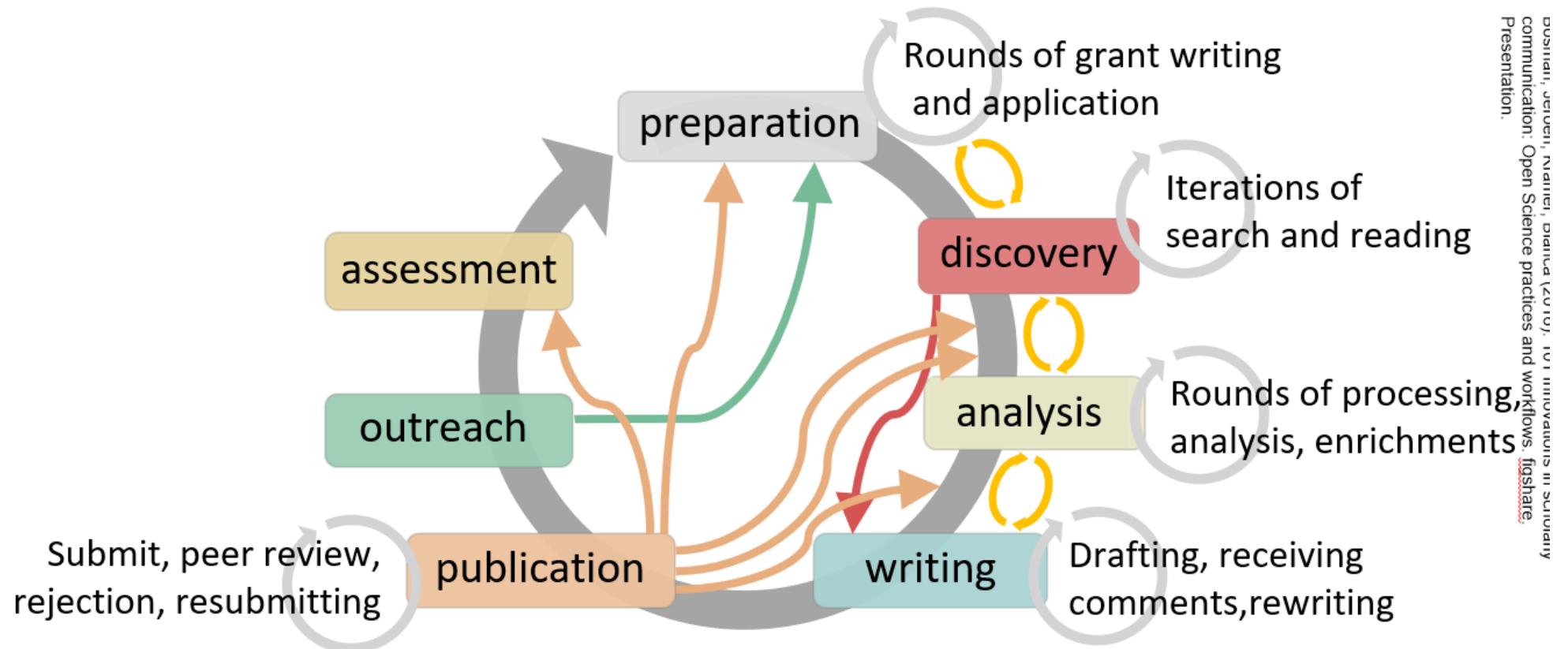
Reconstructing FAIR relevant documentation of finished data sets is virtually impossible.



Source: [Imageflip](#), CC0

Data management good practices in the research workflow

How to make the whole workflow as transparent and open as possible?





Digital Research Infrastructure
for the Arts and Humanities



Share

Download (133.91 MB)

<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?

Data Basics With Databases – The Wonders of Data Material

BY STEFAN KARCHER · PUBLISHED 2018-10-05 · UPDATED 2018-12-06

Vacation time is over and – which is a bit surprising for PhD students – I had much time to think about my dissertation and my academic future. I did some researches and picked up many new interesting fields someone should deal with. However, from time to time, I got angry about an issue which came across a lot. Have you ever noticed how stupid some subject-related databases are? I mean, having them helped us find sources and texts, which enriched our own work, was a wonderful opportunity 10 years ago; but today, 10 years later the relation between the whole bunch of data sets became the focus of some researchers' attention.

The need (or wish) to work with the full material of a database leads to the wide discussion about Open Access in science. I don't want to deepen the aspect here in this post, but to make my opinion clear: Open Access is a great convenience for an open and free science. I support this position and I think my contribution to the scientific society is to share research results and raw data in an easy, accessible way. Yet, I understand the need of legal restrictions, payment, and license policies. Publishers, universities, and scholars invest so much time and money to develop their systems; and, of course, charges and copyright restrictions are necessary for financial profitability. I'm happy to pay for access if that's the price to support scientific progress. One can't be so naive to think open science is for free. Someone must pay for it and even my time/work as a PhD costs resources and money. To make a long story short: If a database is hidden behind payment or license restrictions, there should be good reasons for it. And I hope the reason is not profit, but the necessity to keep the system running.

In this context, it is very important to differ between two types of access. The



A data (re)use case study that highlights many of the challenges DH research is facing:

Stefan Karcher, "Data Basics With Databases – The Wonders of Data Material," in *INFODITEX -BLOG*, 2018-10-05, <https://infoditex.hypotheses.org/245>.

"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.”

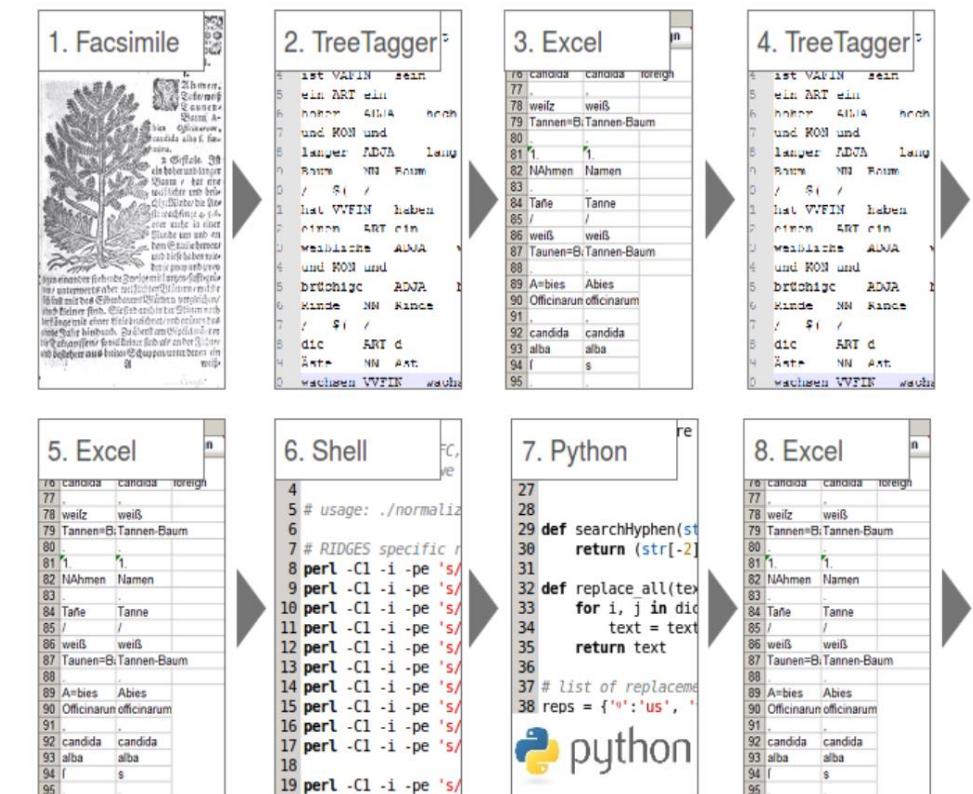
(Mia Ridge)

<http://www.openobjects.org.uk/2017/03/piles-material-patchwork-embed-production-usable-collections-data-library-work/>

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.



Image source: [Pixabay](#), CC0.



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.

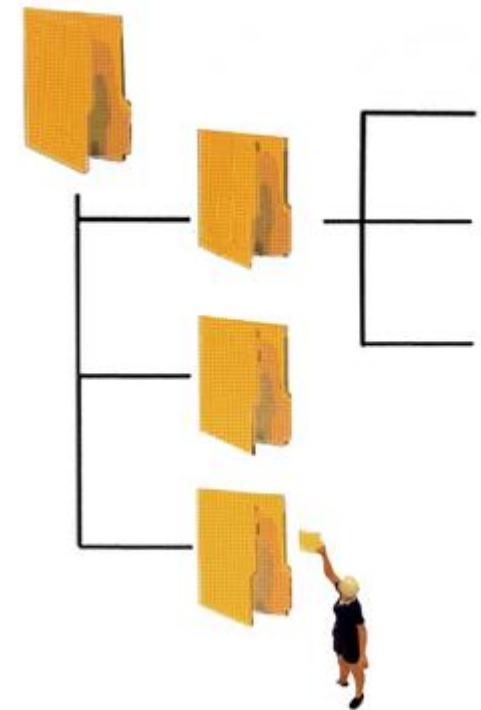
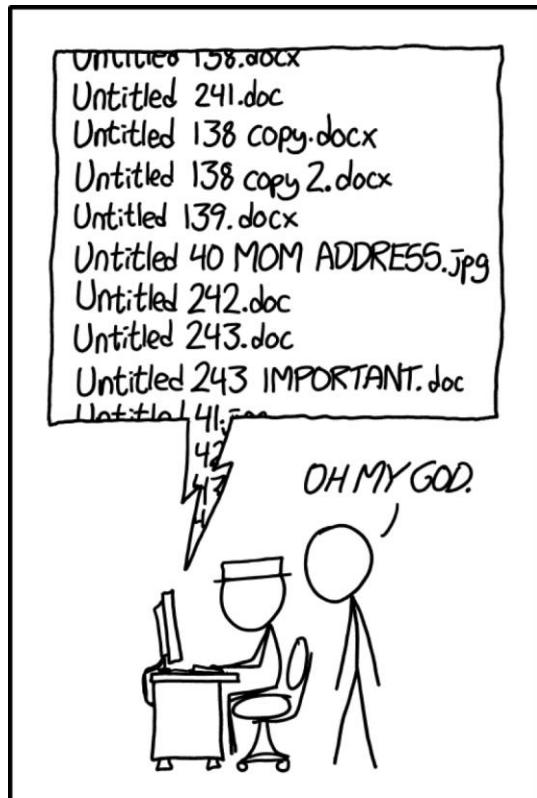


Image source: [CESSDA](#), CC-BY 4.0

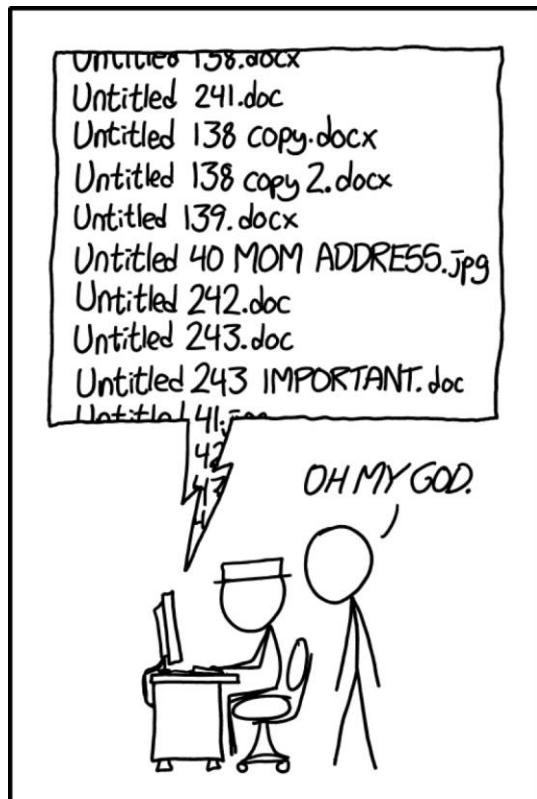
File naming conventions



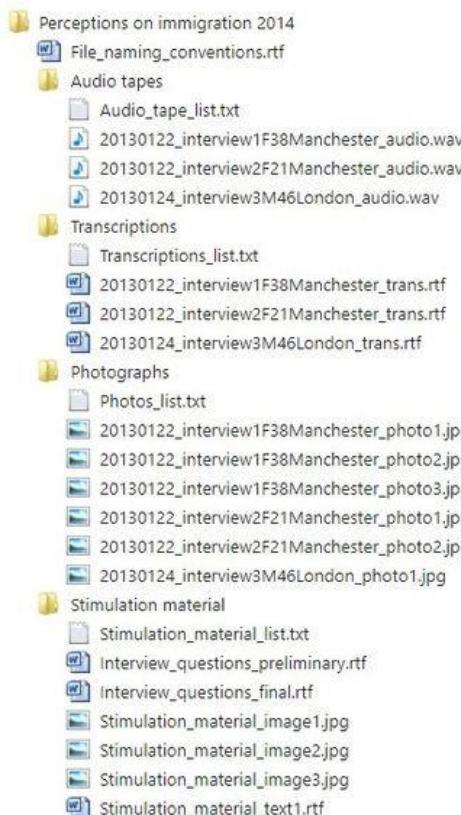
PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

Looks familiar?

File naming conventions



VS.

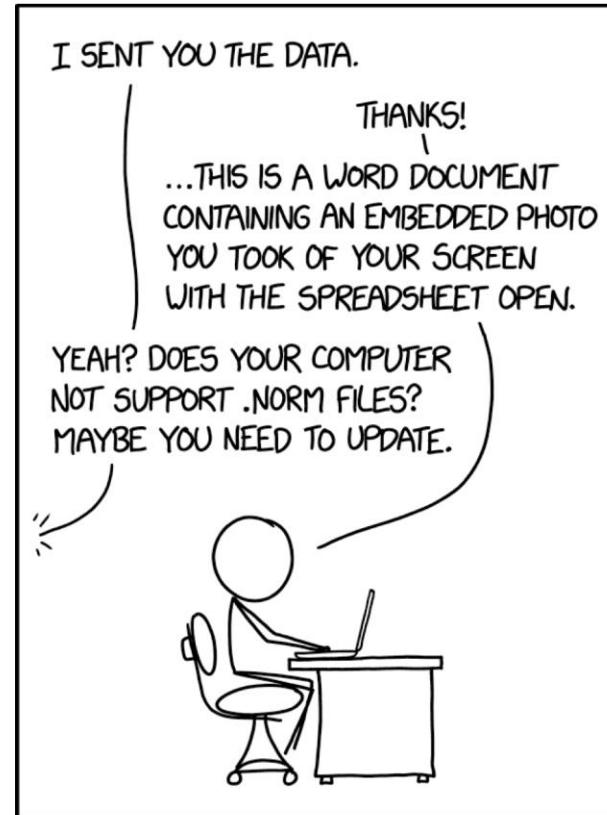


The specifics usually matter less than just having some.

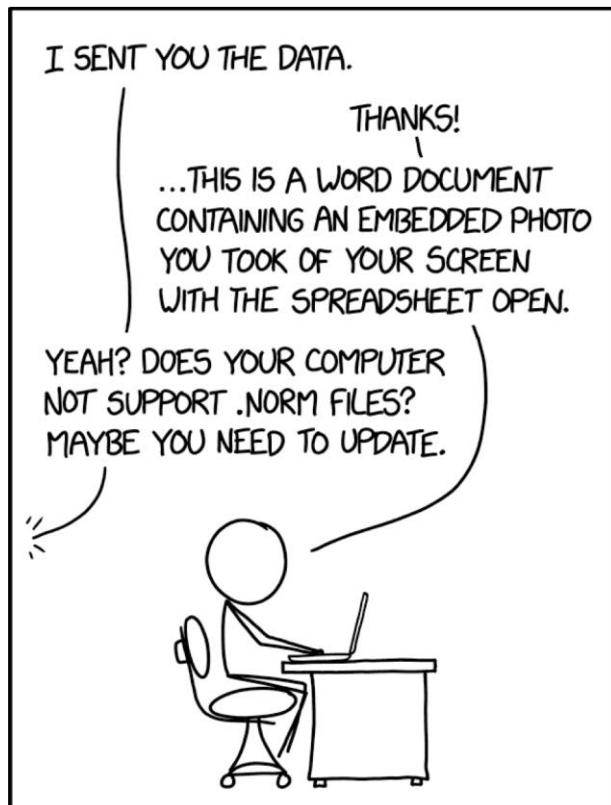
Common elements ([UK Data Service](#)):

- 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?



SINCE EVERYONE SENDS STUFF THIS WAY ANYWAY, WE SHOULD JUST FORMALIZE IT AS A STANDARD.



SINCE EVERYONE SENDS STUFF THIS WAY ANYWAY, WE SHOULD JUST FORMALIZE IT AS A STANDARD.

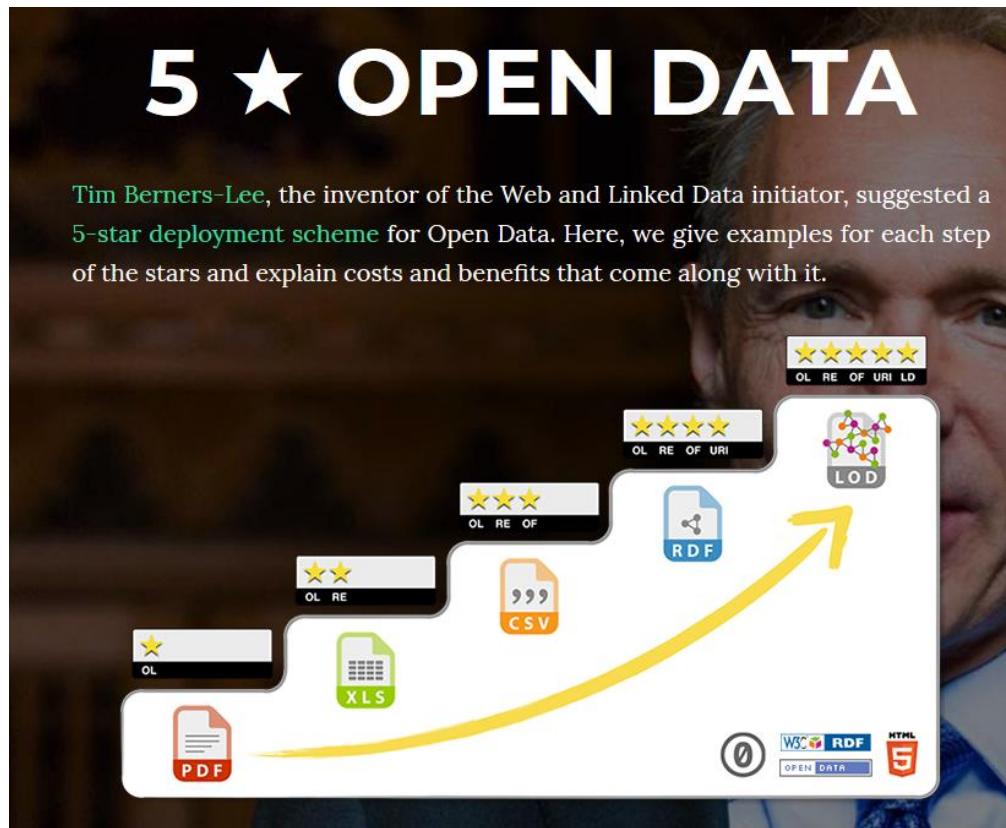
Image source: [Stanford Library](#)

Open formats

Statistical data	Type	• Preferred format(s)	• Non-preferred format(s)
Programming languages	Text documents	<ul style="list-style-type: none"> • PDF/A (.pdf) • ODT (.odt) 	<ul style="list-style-type: none"> • Microsoft Word (.doc) • Office Open XML (.docx) • Rich Text File (.rtf) • PDF other than PDF/A (.pdf)
Images (raster)			<ul style="list-style-type: none"> • Non-Unicode text (.txt)
Images (vector)			<ul style="list-style-type: none"> • SGML (.sgml) • Markdown (.md)
Audio	Plain text	<ul style="list-style-type: none"> • Unicode text (.txt) 	
Video	Markup language	<ul style="list-style-type: none"> • XML (.xml) • HTML (.html) • Related files: .css, .xslt, .js, .es 	
Computer Aided Design (CAD)		<ul style="list-style-type: none"> • MATLAB • NetCDF • TextFabric 	<ul style="list-style-type: none"> • Microsoft Excel (.xls) • Office Open XML Workbook (.xlsx) • PDF/A (.pdf)
Geographical information (GIS)	Programming languages		
Images (georeference)			
Raster grid	Spreadsheets	<ul style="list-style-type: none"> • ODS (.ods) • CSV (.csv) 	
3D			<ul style="list-style-type: none"> • Microsoft Access (.mdb, .accdb) • dBase (.dbf) • HDF5 (.hdf5, .he5, .h5)
RDF	Databases	<ul style="list-style-type: none"> • SQL (.sql) • SIARD (.siard) • CSV (.csv) 	<ul style="list-style-type: none"> • SPSS Portable (.por) • STATA (.dta) • DDI (.xml)
Computer Assisted Qualitative Data Analysis (CAQDAS)			
Abbreviations and acronyms	Statistical data		<ul style="list-style-type: none"> • SPSS (.sav) • SAS (.7dat; .sd2; .tpt)

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 license¹ [example ...](#)
- ★★ make it available as structured data (e.g., Excel instead of image scan of a table)² [example ...](#)
- ★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)³ [example ...](#)
- ★★★★ use URIs to denote things, so that people can point at your stuff⁴ [example ...](#)
- ★★★★★ link your data to other data to provide context⁵ [example ...](#)

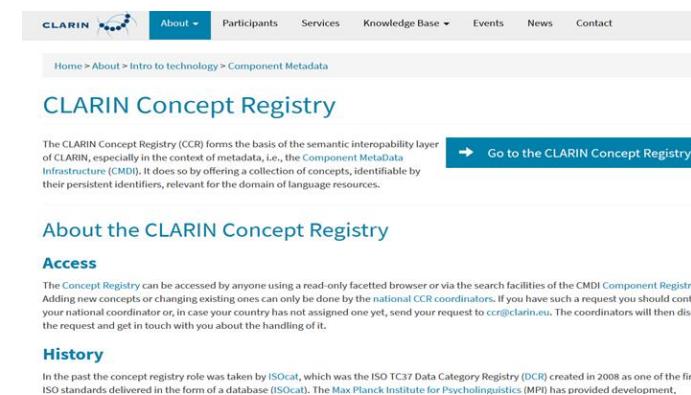
Source: <https://5stardata.info/en/>

Ontologies, vocabularies

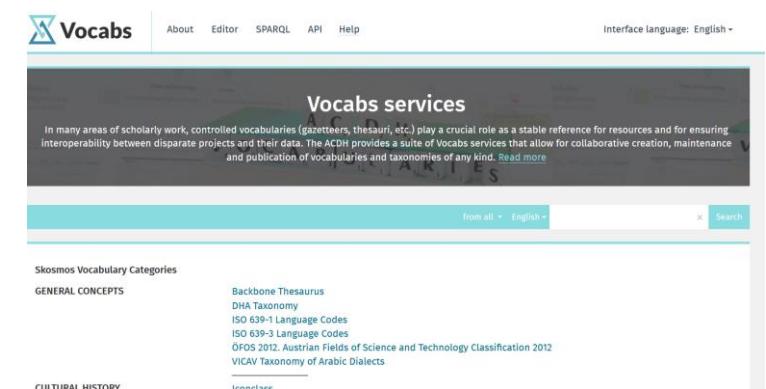
- Put structure on your messy data so that it opens up to others (people, machine, applications, related databases).
- Check whether some of the general topics and terms (persons, locations, concepts) that you focus on have already been assigned persistent identifiers or URIs in one of the ontologies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



The screenshot shows the BARTOC.org homepage. At the top, there's a navigation bar with links for ILC, Title Finder, KOS Registries, and Download. Below the navigation is a search bar with a red 'GO' button. To the right of the search bar are two green boxes: 'Currently indexed vocabularies' (2,986) and 'Currently indexed registries' (89). Underneath these are dropdown menus for 'Content by discipline' (Social sciences, General works, Computer science and Information, Technology, History and Geography) and 'Browse DDC' (with categories like <Any>, EuroVoc, and Language). A sidebar on the left shows a list of categories under 'Browse DDC'.



The screenshot shows the CLARIN Concept Registry homepage. It features a header with the CLARIN logo and a 'About' menu. Below the header, a breadcrumb trail shows 'Home > About > Intro to technology > Component Metadata'. The main content area is titled 'CLARIN Concept Registry' and includes a button to 'Go to the CLARIN Concept Registry'. Below this, there's a section about the registry's role in providing a semantic interoperability layer for CLARIN metadata. Further down are sections for 'About the CLARIN Concept Registry', 'Access', and 'History'.

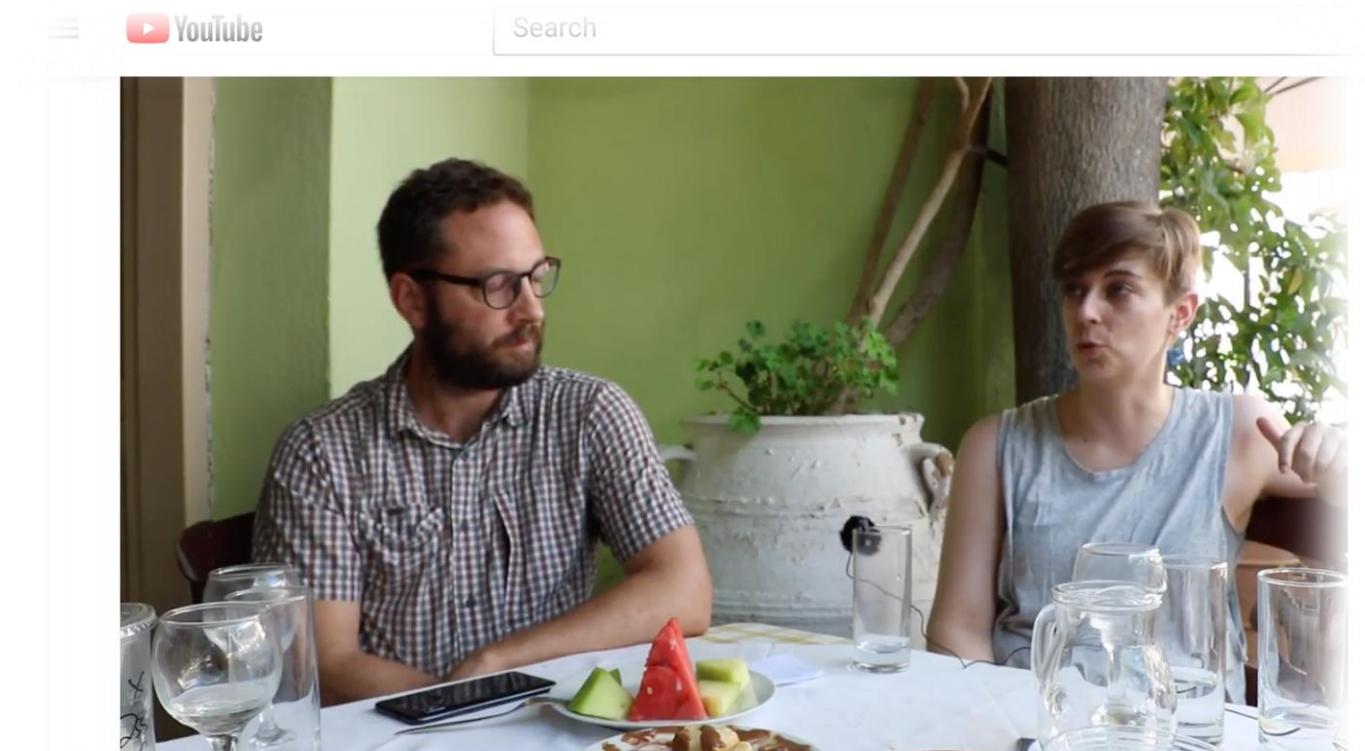


The screenshot shows the Vocabs services homepage. It has a header with the Vocabs logo and a 'About' menu. The main content area is titled 'Vocabs services' and discusses the importance of controlled vocabularies for scholarly work. It includes a search bar and a sidebar with 'Skosmos Vocabulary Categories' and 'GENERAL CONCEPTS' (listing Backbone Thesaurus, DHA Taxonomy, ISO 639-1 Language Codes, ISO 639-1 Language Codes, OFOS 2012, Austrian Fields of Science and Technology Classification 2012, VICAT Taxonomy of Arabic Dialects, and Iconclass).

The Basel Register of Thesauri, Ontologies & Classifications (*BARTOC*)
 The CLARIN Concept Registry and the DARIAH/ACDH collection of vocabularies.

PARTHENOS YouTube Videos

‘An Ontologist and a Data Scientist walk into a bar: Data in Research Projects’



An Ontologist and a Data Scientist walk into a bar: Data in Research Projects
Unlisted

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/>

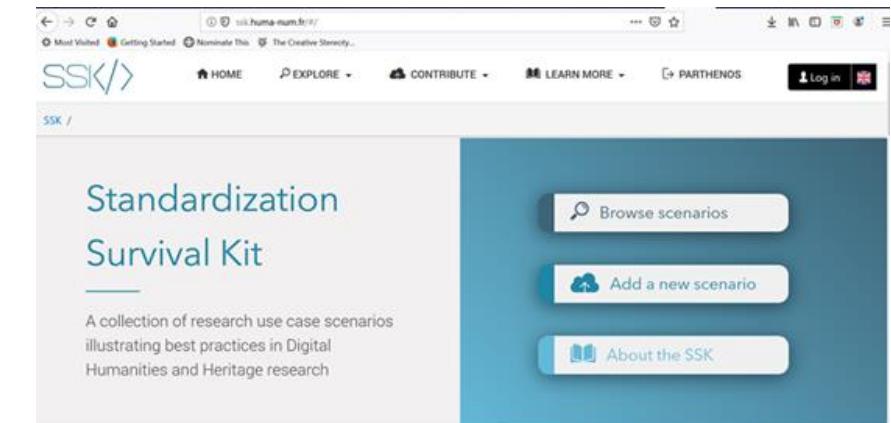
The choice of appropriate standards is more of a community issue than a technical one.

The Standardization Survival Kit

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/#/>



Extract textual content from images
Saranya Balasubramanian, Peter Catrie, Dario Kampkaspar, Tomasz Parkoła, Charles Riondet, Pavel Stranak, Daniel Schopper
[Images](#) [Manuscript](#) [Text](#) [Text Bearing Objects](#) [Literature](#) [History](#)

[+ READ DESCRIPTION](#)

Last updated: 8 months ago

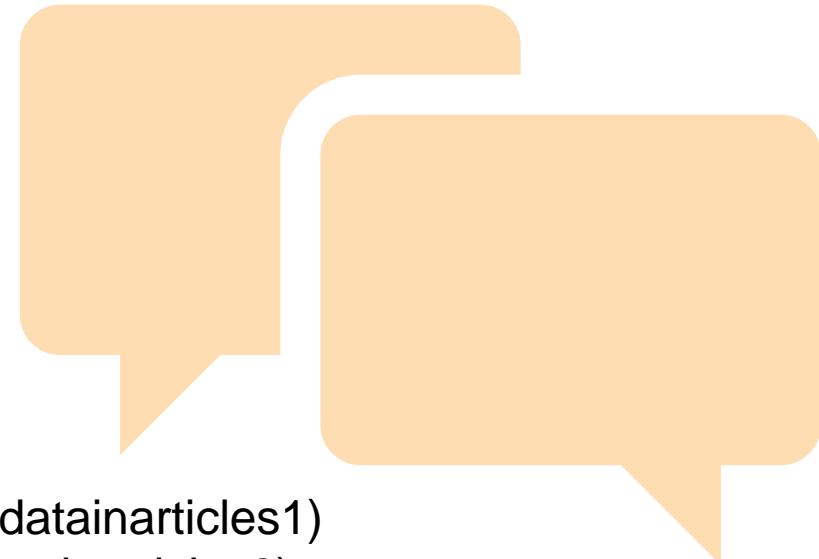
Linking datasets with publications

Excercise:

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?



<https://doi.org/10.1080/0969594X.2016.1194257> (<https://tinyurl.com/datainarticles1>)

<http://dx.doi.org/10.1371/journal.pone.0139563> (<https://tinyurl.com/datainarticles2>)

<http://doi.org/10.1111/lang.12172> (<https://tinyurl.com/datainarticles3>)

<http://doi.org/10.21627/2019cd> (<https://tinyurl.com/datainarticles4>)

Excercise adopted from [Kristina Hettne](#), [Peter Verhaar](#) al. (2019, February). Top 10 FAIR Data & Software Things. Humanities- Historical research. Zenodo. <http://doi.org/10.5281/zenodo.2555498>

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.

Excercise:

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>



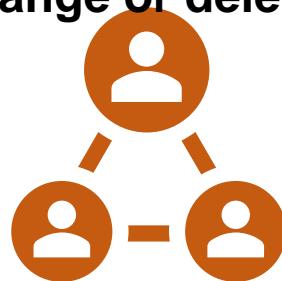
Image source: <https://www.feeldesain.com/travel-tag-texts.html>

3. Data sharing, storing and publication

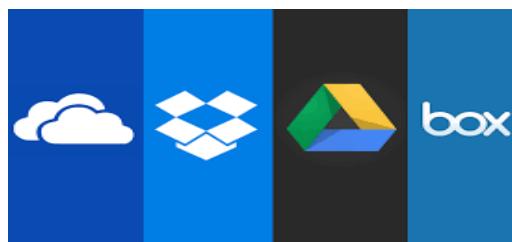
Data sharing and storing

With collaborators while research is active

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



Data are mutable



(Open) data sharing

- Institutional, disciplinary or generic repository



Data are stable, searchable, citable, clearly licensed



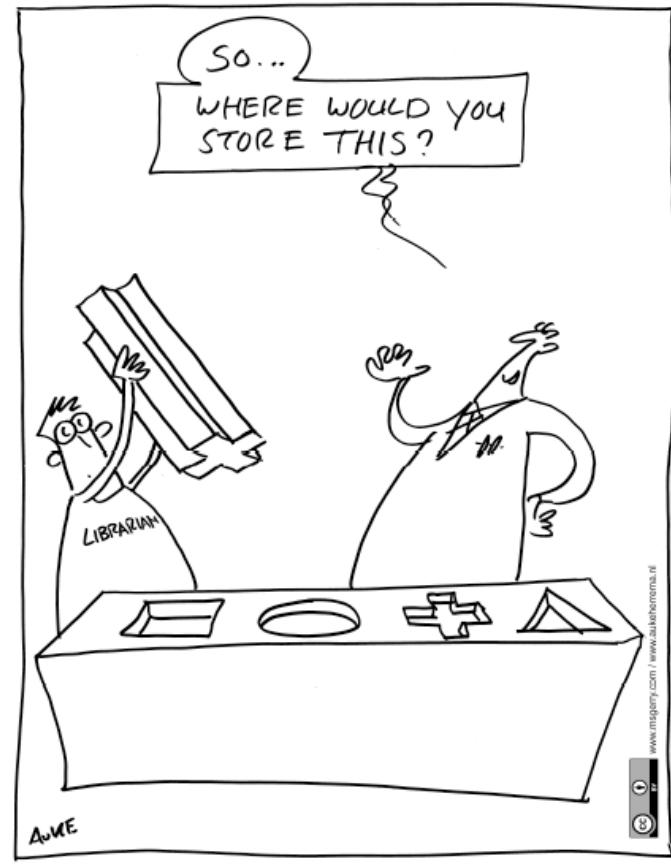
Excercise

- 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.

<https://dans.knaw.nl/en/deposit/information-about-depositing-data/before-depositing>

OR: <https://tinyurl.com/DANSmetadata>

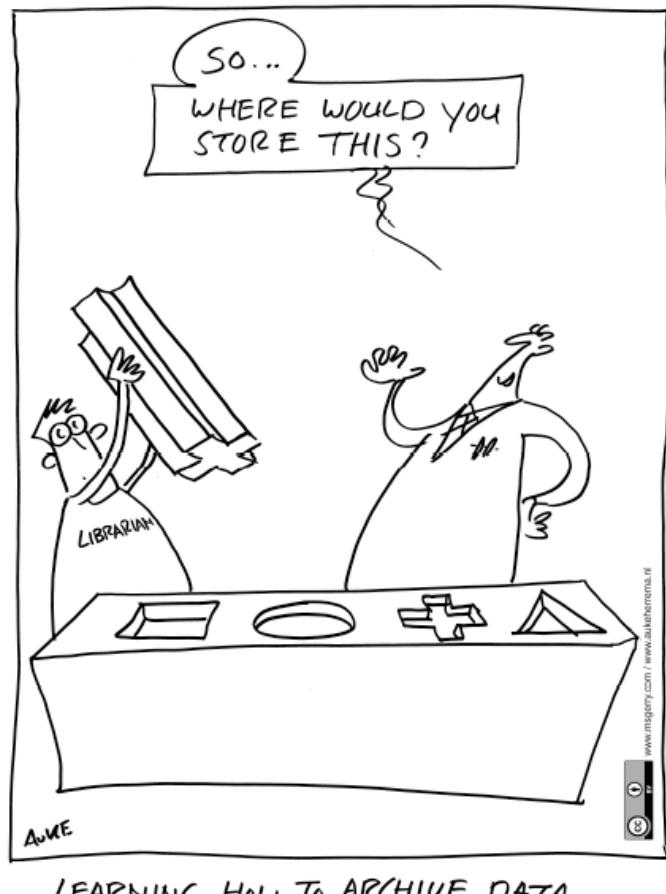




How to select a repository that best fits your research?

What are your governing criteria when selecting a repository?

How to select a repository that best fits your research?



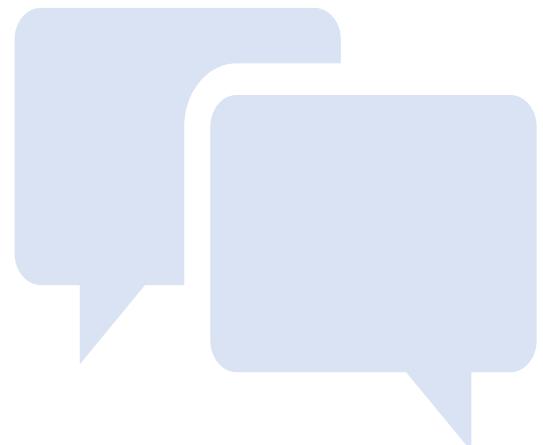
1. Use an external data archive or **repository** already established for your research **domain** to preserve the data according to recognised standards in your discipline.

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

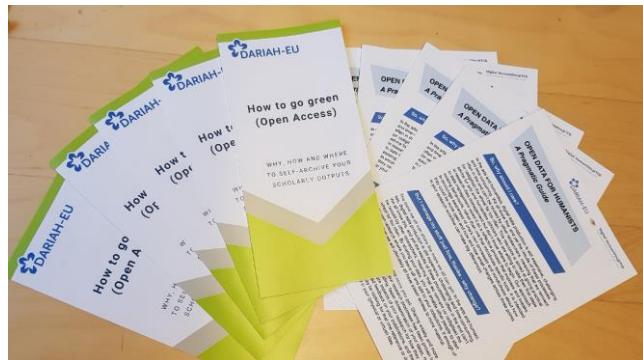
3. Use a cost-free data repository such as **Zenodo**.
4. Search for other data repositories here:
re3data.org.

Pick a position and discuss !

- 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 dat 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!



Adapted from: 'Future Proof and FAIR Research Data: Open Data Management Best Practices and First Steps',
Ulrike Wuttke: <https://www.fosteropenscience.eu/node/2603>



Self-archiving and open data management flyers.
 DOI: [10.5281/zenodo.2657248](https://doi.org/10.5281/zenodo.2657248) and
[10.5281/zenodo.3070069](https://doi.org/10.5281/zenodo.3070069)

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

How to Facilitate Cooperation between Humanities
 Researchers and Cultural Heritage Institutions. Guidelines
[10.5281/zenodo.2587480](https://doi.org/10.5281/zenodo.2587480).

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

How to Facilitate Cooperation between Humanities
 Researchers and Cultural Heritage Institutions. Guidelines
[10.5281/zenodo.2587480](https://doi.org/10.5281/zenodo.2587480).

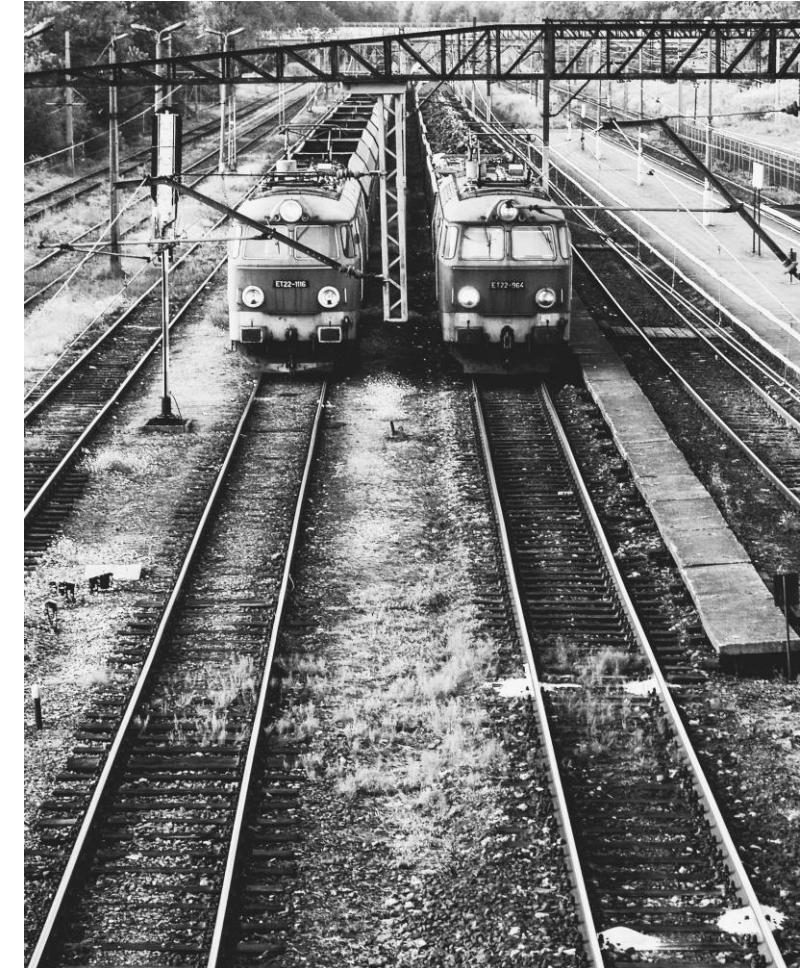
Ready, set, explore, reuse !

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

The Standardization Survival Kit
<http://ssk.huma-num.fr/#/>

Further useful resources

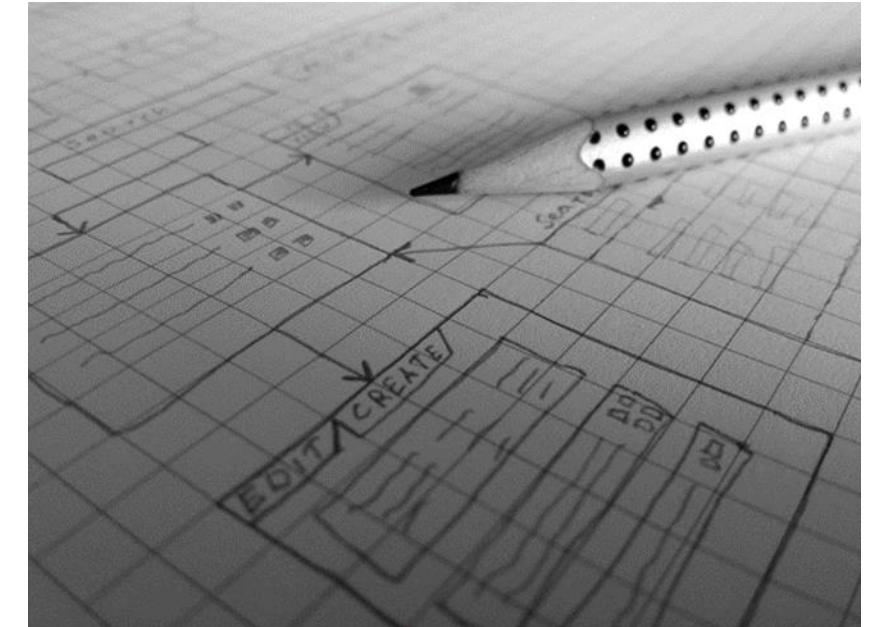
- Erzsébet Tóth-Czifra. The risk of losing thick description: Data management challenges Arts and Humanities face in the evolving FAIR data ecosystem. 2019. [⟨halshs-02115505⟩](https://halshs.archives-ouvertes.fr/halshs-02115505)
- DARIAH Pathfinder to Data Management Best Practices in the Humanities:
<https://campus.dariah.eu/resource/dariah-pathfinder-to-data-management-best-practices-in-the-humanities>



Source: [Pixabay](#), CC0.

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 **on-going 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?**



Source: [Pixabay](#), CC0

An example:

Performance Criteria		Performance Levels		
		Detailed	Incompletely addressed	Not addressed
Section 1	What types of data will be created?	Clear descriptions of the types of data being created with details of formats and approximate dataset sizes where appropriate.	Some description of the data being created but it is unclear or incomplete (based on subsequent answers)	Minimal information about what data types are being created.
Section 1	Why are these data types being proposed?	The plan explains why the particular data type and format is being used and how it will contribute to the project/answering the research question.	Some mention of why these data types are being used but it is not clear how they will contribute to the project.	No explanation given for why the proposed data types are required for the research or how they will aid in answering the research questions.
Section 2	What methodology is being proposed?	Methodologies are described clearly for each data type and information given about any intermediate data type produced before the final data.	Methodologies are described but either not clearly or not for all data types outlined in section 1.	Minimal description of the methodologies being proposed
Section 2	How is this methodology suited to the data types or digital outputs?	As well as linking methodologies to the data types the plan explains how the methodology will produce the data type.	It is clear which methodology is related to each data type but not necessarily why that methodology suits a data type	No mention of how the methodologies proposed connect to the data being collected or created.
Section 2	How is the project team suitable for the digital/data aspects of the work?	The plan details the skills necessary to deliver the data and digital aspects of the project, and explains how the project team either deliver these skills, or will be trained to do so (eg transcribers will be trained in TEI/XML).	The plan mentions skills necessary to deliver the project, but does not indicate how the project team deliver these skills. OR the plan mentions skills necessary but does not cover the full range of the project.	There is no mention of the digital or data skills needed to deliver the project, or how the project team will provide these skills.
Section 2	How will the institution's support teams be supporting these methodologies?	Where methodologies use specialist techniques, equipment or processes it is clearly described how the research will be supported in implementing their proposal.	There is some mention of institutional support for the methodologies involved in the project, but it is not clear if the support covers all the specialist methodologies involved in the project.	No mention of how the researcher will be supported in using these methodologies by their institution

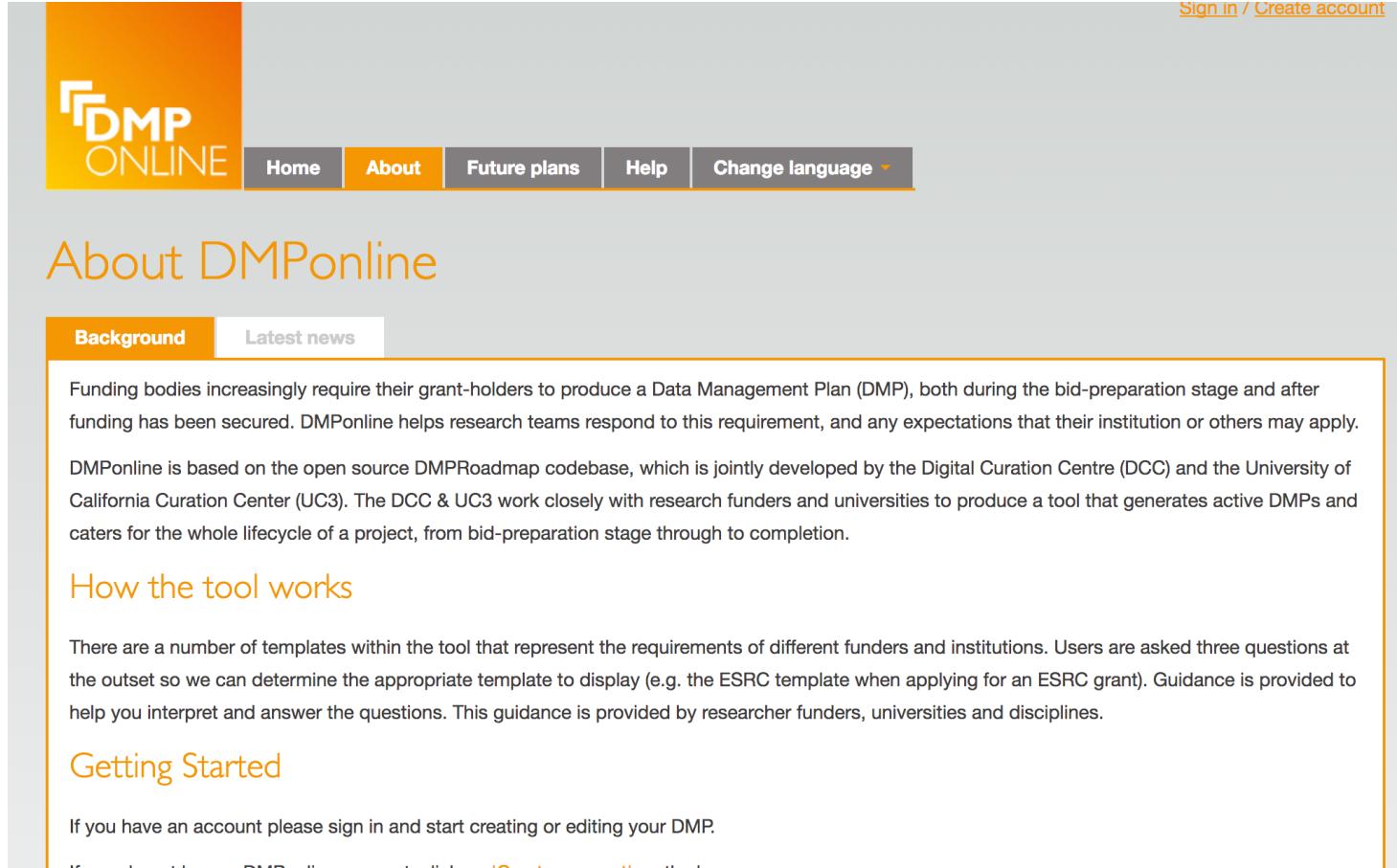
This rubric is designed as a checklist or marking aid for those reviewing data management plans for submission to the Arts and Humanities Research Council (AHRC).

Source: Donaldson, Mary, & Higman, Rosie. (2018, November). Arts and Humanities Research Council Data Management Plan Rubric. Zenodo. <http://doi.org/10.5281/zenodo.1745533>

Performance Criteria		Performance Levels		
	Detailed	Incompletely addressed	Not addressed	
Section 3	Has an appropriate storage solution for the duration of the project been described?	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.	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.	It is not clear where the data will be stored during the project or the storage solution proposed is inappropriate.
Section 3a	Is the backup process described appropriate?	The back-up process for active data storage is clearly described or referenced, and is appropriate for the data to be collected.	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.	No backup process is described or the one described is inappropriate or inadequate for the data that is being collected.
Section 4a	Has an appropriate long-term storage solution been described?	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.	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.	No long-term storage plan is mentioned, or the solution proposed is inappropriate, either for the data to be retained, or does not comply with funder requirements.
Section 4b	How long will the data be stored for and is this appropriate to the project?	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.	The long-term retention schedule is mentioned, but may not cover all data or may be inadequate or inappropriate.	No long-term retention schedule is mentioned.
Section 4c	Has long-term storage costs been described in the plan?	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.	Cost for long-term storage are mentioned, but no detail. Costs may not appear to cover all the data. Costs may appear to be inappropriate for the storage option indicated.	No costs for long-term storage are mentioned.
Section 5a	Has the value of the data to the disciplinary area been outlined?	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	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	The value of the data to the disciplinary community is not mentioned.

Use Tools for Data Management Planning

e.g. DCC DMPOonline
<https://dmponline.dcc.ac.uk/>



The screenshot shows the DMPOonline website's "About" page. The header features the "DMPOonline" logo with a stylized orange and yellow square icon. The navigation bar includes links for Home, About, Future plans, Help, and Change language. The main content area has a light gray background and displays the following text:

Funding bodies increasingly require their grant-holders to produce a Data Management Plan (DMP), both during the bid-preparation stage and after funding has been secured. DMPOonline helps research teams respond to this requirement, and any expectations that their institution or others may apply.

DMPOonline is based on the open source DMRoadmap codebase, which is jointly developed by the Digital Curation Centre (DCC) and the University of California Curation Center (UC3). The DCC & UC3 work closely with research funders and universities to produce a tool that generates active DMPs and caters for the whole lifecycle of a project, from bid-preparation stage through to completion.

How the tool works

There are a number of templates within the tool that represent the requirements of different funders and institutions. Users are asked three questions at the outset so we can determine the appropriate template to display (e.g. the ESRC template when applying for an ESRC grant). Guidance is provided to help you interpret and answer the questions. This guidance is provided by researcher funders, universities and disciplines.

Getting Started

If you have an account please sign in and start creating or editing your DMP.

If you do not have a DMPOonline account, click on "Create account" on the homepage.

Research Data Management Organiser

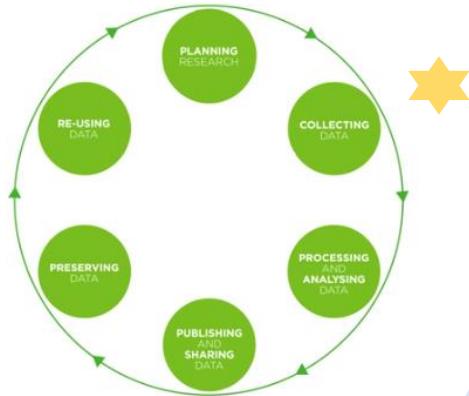
<https://rdmorganiser.github.io/en/>



The screenshot shows the RDMO homepage. At the top, there is a dark banner with the DARIAH-EU logo and the text "RDMO - Research Data Management Organiser". Below the banner, the main title "RDMO" is displayed in large white letters, with "Research Data Management Organiser" in smaller white text underneath. The background of the page is a grayscale photograph of a library or archive with shelves filled with books and containers. At the bottom of the page, there is a light gray box containing two paragraphs of text.

The Research Data Management Organiser (RDMO) enables institutions as well as researchers to plan and carry out their management of research data. RDMO can assemble all relevant planning information and data management tasks across the whole life cycle of the research data.

RDMO is ready for application in smaller or bigger projects. In the next projectphase, which started November 2017, the RDMO tool will be extended and the project partners AIP, FHP, and KIT Library will collaborate with the RDMO users to improve its usage. The tool will be extended by enhancing its implementation of roles and interfaces to institutional infrastructure, e.g. repositories, ticketing systems, and the infrastructure for authentication and autorization. Tutorials, documentation and other material are planned for dissemination, and workshops for users and developers.



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?

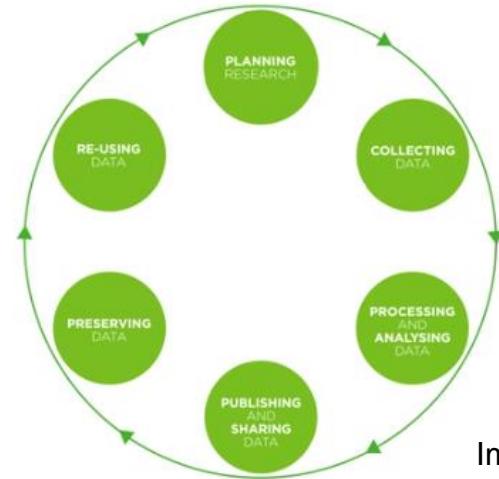


Image source: [UK Data Archive](#)



Casey Parker @caseynparker · Jun 8

For all my married science friends: I'm getting married in April and am thinking about changing my last name. Tell me your personal experience with this, your decision, and why you made that decision. Looking forward to seeing everyone's perspective! #WomenInSTEM

263

36

178



Melanie Imming

@MellImming

Replying to @caseynparker

Get an **#ORCID** and all your publications are recognised even if you change your name. Solves a great deal of the problems?

1:47 PM · Jun 8, 2019 · Twitter for Android

How do we find data for reuse?

- **Persistent Identifiers (PIPs)** ensure that online references to publications, research data, and persons remain stable and available in the future **even if their location changes**.
- A PIP is a specific type of a [Uniform Resource Identifier](#) (URI), 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.

<https://twitter.com/MellImming/status/1137325232396615680>



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](https://doi.org/10.14293/S2199-1006.1.SOR-UNCAT.CL49QV3.v1)

hal-01836189