# Open Data for Humanists, A Pragmatic Guide

Jennifer Edmond, Erzsébet Tóth-Czifra

DARIAH-EU
Digital Research Infrastructure for the Arts and Humanities

Digital Humanities@TCD
Digital Humanities at Trinity College Dublin

# Open Data for Humanists, A Pragmatic Guide

## *So, why should I care?*

Most humanists would agree that sharing knowledge with other researchers is a cornerstone of academic life. Many will also fear that sharing too much, too early can be professionally damaging, however. And many also would not find much resonance between how they see their work and the discourses of Open Data, with its emphasis on particular approaches to Data Management Planning that have been adapted from other, more data intensive, disciplines.

The conflict between these positions is in part a semantic one. The many things that would be seen as data in another discipline are often called something else in the humanities. We resist the blanket term ‘data’ for the very good reason that we have more and precise terminology (e.g. primary sources, secondary sources, theoretical documents, bibliographies, critical editions, annotations, notes, etc.) available to us to describe and make transparent our research processes.

The conflict is, however, also a material one. Our relationships to our sources tend to be different. We seldom create them from scratch for the purpose of a project: instead, we start from a firm basis or historical, cultural or creative practices carried out by others. We don’t own our data, not solely, and this changes what we can and cannot do with it. Data Management Planning protocols tend to start from an assumption about data as having a scope the researcher controls. This is seldom actually the case in the humanities, a fact that renders a lot of what are seen as fundamental tools for data management very limited in their utility for the humanities.

But there are huge benefits to overcoming the barriers to sharing data in the humanities. Through it, we will all gain better access to rare materials, making the representation of marginal or underrepresented positions stronger.

What we recommend here proposes a different approach to data management, viewing 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.

## But I manage my stuff just fine, thanks - why change?

**The more we all can share the more we all can get:** 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. By sharing your data, you will enable other people to start their own research where you leave off and add to your collections. If enough of us do this, then we will ultimately all have easier access to more material.

**But also, the more <u>you</u> share the more <u>you</u> get:** Sharing widely, and more than just final publications, from your research processes has been proven to have lots of positive effects, including: faster dissemination of your ideas, new collaborations, recognition of more aspects of your research for how they advance knowledge, more informed feedback on your publications, and the gratitude of your future self, who may be looking for that one reference or document that could otherwise be buried in your (physical or virtual) files.

## *OK, so what do I do?*

Small changes across three points in your research workflow can make big differences. More information is available in the FAQs section on items in **blue**.

**In Publications, as Communicators of Data:** *because the better people can see how you validate your conclusions, the more confidently they can reuse your data*

1. Have an **ORCID ID**. Linking your stuff to your other stuff benefits you and your readers.
2. Know and use your green OA (free deposit) options, e.g. through the **SHERPA-Romeo** service
3. Make your personal knowledge provenance clear in your work: cite and acknowledge broadly: the source, the digital source, the archive, the website, the infrastructure...
4. Cite the published version of a source, but if you used a preprint or copy, cite that too.

**Through Processural Phase Data:** *because the more people can see of how you reached a conclusion, the more confidently they can reuse your data*

1. Be aware of your processes, their strengths and weaknesses, their hybridity, the tools you use, and potential for future nudges toward openness (such as an open, annotated Zotero bibliography, or a file-naming convention).
2. Be generous with your processural data, as and when opportunity arises.
3. Make explicit how your source data has been 'cooked'. Share or at least indicate, to the greatest extent that you can, the contextual components and contributors to your understanding of the material and realization of your conclusions etc.. How has your

source data has been collected, annotated, combined with other sources and organised? How far back can you trace the influences contributing to your own thinking about how to resolve your research questions?

4. If you use any services, tools, code, corpora or virtual research environments to read, organise, mine or visualise your data, then cite them. If possible, use **Persistent Identifiers** in citations to help locations of and key information about these resources easily identifiable both for humans and machines.
5. If you are building a digital project, inform yourself about how to use standards and open software options to make your end result maximally reusable, and how to build community around to make sure it gets reused.

**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 (eg. personal data) in the data I should be aware of, and what are the best practices for using this data for research purposes?
    d. Are there any specific references or identifiers you recommend I capture in my personal metadata to facilitate reuse?
    e. Can I deposit any data surrogates I create with you or link mine with somehow with your collection? Would you like me to make you aware if I deposit them elsewhere (e.g. in my institutional repository or other data repositories)? Whom should I contact with this information?
    f. How would you like your institution to be cited/acknowledged in any publications in which I use them?

2. When you can deposit, know what your options are and how you make the most of a deposit (for example by including related publications, explanatory notes, codes lists, and/or other paradata alongside the sources deposited).
3. Be aware of what common templates for Data Management Plans are and do, and when you might need to use one of them.
4. To maximize the potentials of your publications, interlink them with any processural or source data you are able to make open. Some journals specialise in making this possible, but you can also find repositories to use for this, or at least provide links to one from the other. All of the components of your scholarly production can be reused and recognised more effectively when they are visible as parts of the same project.

## *Useful Resources and FAQs*

How do I get an ORCID? https://orcid.org/

Where can I learn more about managing data in humanities research? http://training.parthenos-project.eu/sample-page/manage-improve-and-open-up-your-research-and-data/

How do I know a journal's Open Access Policy? http://www.sherpa.ac.uk/romeo/index.php

Where can I learn the basics about persistent identifiers and data citation? https://www.youtube.com/watch?v=PgqtiY7oZ6k

How are the various Creative Commons Licenses different? https://creativecommons.org/choose

Where can I find a data repository I can use?

https://www.re3data.org/

Where can I find a repository for my research papers? http://v2.sherpa.ac.uk/opendoar/

How do I know what standard to use? http://www.parthenos-project.eu/portal/ssk-2

Under what conditions can I reuse cultural data? https://www.kl.nl/wp-content/uploads/2015/09/150617-Europeana-Food-and-Drink-IPR-Guides-FINAL-.pdf

What guidelines should I follow when citing data? https://www.force11.org/datacitationprinciples

Authored by Jennifer Edmond and Erzsébet Tóth-Czifra.