DESIR WINTER SCHOOL - COLLABORATIVE NOTEBOOK

10-13 December 2019 | NOVA FCSH

10 December 2019 - Day 1

What is data in the humanities?: Erzsébet Tóth-Czifra

Specificity of data in the Humanities

- multilingualism
- continuum of data curation between Cultural Heritage Institutions and scholars
- difficult to distinguish primary/secondary sources

Does the availability/ online visibility of data distort the research being done i.e. do researchers follow the easily available digital data (e.g. by the British Library) and ignore the long tail?

A success story that highlights how the availability of good-quality, standardized, machine-readable data can give rise to tools that change the ways scholarship can be conducted: example of Bentham project: https://www.ucl.ac.uk/bentham-project/, used Citizen Science and produced tools (Transkribus, link below):

https://transkribus.eu/Transkribus/

Assessing the FAIRness of data: https://www.fosteropenscience.eu/learning/assessing-the-fairness-of-data/#/id/5c52e8cf0d3def29462d8cb5

Does the availability/ online visibility of data distort the research being done i.e. do researchers follow the easily available digital data (e.g. by the British Library) and ignore the long tail?

Sharing is not giving away!

FAIR principles (https://www.go-fair.org/fair-principles/ or https://www.ands.org.au/working-with-data/fairdata/training). The original article: https://www.nature.com/articles/sdata201618

Accessible is not necessarily open (it just means I have to know where to find the data and under well-documented access conditions)

EXERCISE: Linking datasets with publications

Data in article 1: https://www.tandfonline.com/doi/full/10.1080/0969594X.2016.1194257

- data accessible only via the publisher's website
- ownership issue
- data is not raw, we can reuse only for the same purpose
- only data chunks
- licence unclear

Data in article 2: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0139563

- PLoS has a data policy
- data are linked via DOI to a repository, so they are independently accessible
- clear documentation

Data in article 3: https://onlinelibrary.wiley.com/doi/full/10.1111/lang.12172

- data accessible via a repository
- they have a readme file to explain/document the data

Data in article 4: http://chinesedeathscape.org/index.html

- this is an ongoing project
- important to notice it is a scholarly monograph: you can make your data visible not only in journal articles
- data is still embargoed, under 12 month embargo

FAIR data means that your data can have an independent life from its original context. That's why you need rich metadata on provenance, methods, protocols etc.

- A good example https://www.fosteropenscience.eu/content/future-proof-and-fair-research-data-open-data-management-best-practices-and-first-steps

DANS Information on how depositing data:
https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data

Importance of

- file naming/folder organization
- formats e.g. DANS preferred formats https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/before-depositing/file-formats?set_language=en
- increasing openness/ reusability in the 5 star development scheme for Open Data: https://5stardata.info/en
- ontologies, standards and controlled vocabularies, e.g. BARTOC https://bartoc.org/ Basel Register of Thesauri, Ontologies & Classifications [vocabularies are just controlled list of terms, ontologies make relations possible among entities - same difference between a vocabulary and the whole grammar]

If you have ontologies or standards please contribute to FAIRsharing, as there is still a lack of resources from Humanities/Social Science: https://fairsharing.org/

ONTOHGIS: ontological foundations for historical GIS (settlements and administrative units : https://ontohgis.pl/, https://github.com/ontogeohist/ontology
FISH (Forum on Information Standards in Heritage): http://www.heritage-standards.org.uk/fish-vocabularies/
Getty Vocabularies: https://www.getty.edu/research/tools/vocabularies/
Data for history http://ontome.dataforhistory.org/

PARTHENOS project http://www.parthenos-project.eu/
Standardization Survival Kit (**application of standards presented as a "recipe"**): http://ssk.huma-num.fr/#/

EXERCISE: https://cornell.app.box.com/v/ReadmeTemplate

Repositories:

- how to ensure sustainability/long term support? Too many projects ends and what about the data/data repositories?

- be sure that you deposit in the repository where your peers go (like e.g. Humaniites Commons https://hcommons.org/ Think of the “social life” of your data
- Nakala FR https://www.nakala.fr/ and data model https://www.nakala.fr/NAKALA_RDF_model_en.pdf

Repository finding service: www.re3data.org

DARIAH community on Zenodo: https://zenodo.org/communities/dariah/?page=1&size=20

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⟩

11 December 2019 - Day 2

Data and Software citation practices and PIDs: Frances Madden

Introductory video: https://www.ukdataservice.ac.uk/citethedata

Persistent identifiers (PIDs):
- long lasting reference to a resource
- composed by "persistent" as an organization committed to keep it alive and "identifier" globally unique string

Journal articles, people, datasets, software, organisations etc. may have a PID.

Why PIDs?
- identify
- give persistent reference
- is independent from languages
- trustworthiness
- disambiguation e.g. same name different affiliation; married women who changed surname...
- allow linking
- to make resources FAIR (all 4 principles supported by assigning PIDs)

How to get a PID?
Mainly by depositing in a repository (Zenodo, Figshare, Institutional repositories, Github…)

When to use a PID?

Example of (removed) citations: A Culture of non-citation: Assessing the digital impact of British History Online and the Early English Books Online Text Creation Partnership http://www.digitalhumanities.org/dhq/vol/11/1/000282/000282.html

Elements of a citation:
- creators
- year
- title
- format [dataset]
- source [Zenodo]
- DOI
- location
- version

FORCE 11 principles on **data citation**: https://www.force11.org/datacitationprinciples

Software citation principles: https://doi.org/10.7717/peerj-cs.86

Chue Hong, Neil P., Allen, Alice, Gonzalez-Beltran, de Waard, Anita, Smith, Arfon M., Robinson, Carly, … Pollard, Tom. (2019, October 15). Software Citation Checklist for Developers (Version 0.9.0). Zenodo. http://doi.org/10.5281/zenodo.3482769

Open Research Notebooks: Javier de la Rosa

Wolfram Mathematica 1988: https://www.wolfram.com/mathematica/

Tools mentioned in the proceedings of the annual ADHO conferences (2015–2019)
https://lehkost.github.io/tools-dh-proceedings/index.html

Jupyter and Open Notebook are virtual environments for research.

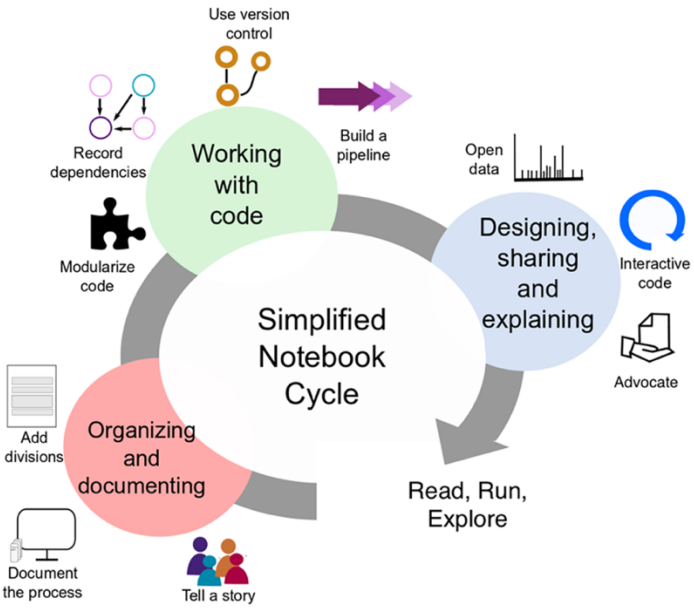
Jupyter works with around 40 programming languages.

“Literate computing” documents that can be read like texts but also contain the code and the executed code

Reproducible research using Jupyter https://reproducible-science-curriculum.github.io/workshop-RR-Jupyter/
Jake VanderPlas: Jupyter advocate - https://twitter.com/jakevdp/

Python data science handbook
https://jakevdp.github.io/PythonDataScienceHandbook/

Ten simple rules to writing a Jupyter notebook
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007007



Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks, Rule A, Birmingham A, Zuniga C, Altintas I, Huang SC, et al. (2019) Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. PLOS Computational Biology 15(7): e1007007. https://doi.org/10.1371/journal.pcbi.1007007

Document step by step is important in the first place for your future self, to remember why you made some choices!
https://pypi.org/ repository of software for the Python programming language.

https://github.com/versae/open-research-notebooks

Quinn Dombrowsky: Natural language processing resources for multiple languages (GitHub): https://github.com/multilingual-dh/nlp-resources

12 December 2019 - Day 3

Copyright and (Open) licensing: Walter Scholger

Legal system:

Common Law: Utilitarianism (right of public to better themselves) [Copy-right: right to copy]
Civil Law: Natural rights attached to human beings [Author's right]

Differences:

COMMON LAW
- author can transfer or waive their rights
- work for hire
- no neighbouring rights
- "fair use"

CIVIL LAW
- personality rights are not transferable
- focus on creativity and human action
- neighbouring rights
- no fair use

Commonalities

Work:
- original works
- expressions, not ideas are protected

Territoriality principle
- national law stops at border
- the law of the country of the user applies in case of infringement

WIPO 1996: making available to the public = uploading

"**Work**"= original intellectual creation

[Creation mean that your particular expression is protected not the idea in inself]

Not every work is an artwork and not every artwork is a work

"**Author**" is always a natural person

Institutions can claim exploitation rights, never the copyright

Artificial intelligence: discussed, but the latest directive did not take it into account.

Moral rights can never be waived

But in Common law you have work for hire, copyright waiver, public domain

Exploitation rights vary among different legislations

what particularly interests us is Reproduction and Making available (i.e. download and upload)

Case studies:

- Digitization:
- public domain works remain in public domain
- libraries are entitled ONLY to fair compensation (not exaggerated fees!)

- Teaching: when the law says “clearly defined circle” = requires a login [for instance Moodle is ok as it requires login]
  Problem is that even the exceptions often comprise “reproduction” but not “making available”, which means you can download but not upload

“A license is a formalised promise not to sue”
Only the rights holder can license
Attribution cannot be waived

Software: machine-language, can be understood as text, Creative Commons may apply to it, but other licenses works better with software

Licenses never override copyright law!

“Non-commercial” is focused on reuse (nobody will sell something freely available) e.g. an article on a subject included in a database which will be sold: if you apply a NC licence your articles won’t be included

Legally-binding text in Creative Commons that applies in international law

CC zero, not legally valid in the EU for “works”

A licence is supposed to clarify what you can do with a work not to create legal uncertainty

A useful guide on data and their legal aspects, and the possible applicable protection: Thomas Margoni, Data ownership http://eprints.gla.ac.uk/171314/

Data Management Plans: Antonia Correia

PARTHENOS module on Ethics and Research: http://training.parthenos-project.eu/sample-page/manage-improve-and-open-up-your-research-and-data/

CESSDA training: https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide

5 steps to decide what data to keep:
http://www.dcc.ac.uk/resources/how-guides/five-steps-decide-what-data-keep

Open Science also to check results
See Retraction watch on articles retracted for fraudes/scientific miscounduct:
https://retractionwatch.com/

H2020 Open Research Data Pilot: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

Principle: as open as possible, as closed as necessary

In H2020, research data is open by default

Thematic repository for depositing data, if not, institutional repository, if not Zenodo or other repositories on Re3data

When opt-out:

- commercial or industrially exploited
- confidentiality
- protection of personal data
- security issues
- no research data to be collected

But, important to explain why data is not possible to be shared.

Open Data Excuse Bingo: http://tinyurl.com/cvkba2h

Charges for storing datasets may be charged to the project! Anticipation is crucial.

Timeline: proposal (DMP + planned budget), 1st version in the first 6 months, update (changes in data, policy, consortium), final review.

An increasing number of funding entities asks for a DMP:

1. Dataset description
2. Documentation and quality of data
3. Backup and storage

4. Ethical requirements and code of conduct
5. Data sharing and long time preservation
6. Responsibilities and resources in data management

FAIR DATA Management Guidelines
https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

FAIR data checklist https://zenodo.org/1065991 FAIR data ≠ open data

DMP online (template based on funding agencies tools) https://dmponline.dcc.ac.uk

13 December 2019 - Day 4

Innovative Publishing Practices in the Arts and Humanities: Delfim Leão

A LIST OF USEFUL RESOURCES ON HOW TO MAKE THE MOST FROM YOUR PUBLICATIONS. FAIR in the SSH scholarly communication:
https://docs.google.com/document/d/13Md2tUGNHzOZxNWnRRtE4O-28E2tmkoGkIP9riLgv2E/edit?usp=sharing

OPERAS (https://operas.hypotheses.org/) - distributed research infrastructure (SSH)

The 'long tail' of scholarly communication:

- cultural diversity
- linguistic diversity (not a new thing in the history of humankind, a challenge but also an opportunity)
- small size of players
- diverse levels of skills and resources (some people remain unaware of what is needed in the new publishing, affects the way in which scholarly communication can be promoted)
- lack of information in the market (difficult to follow what is happening)
- v. soft type of coordination between players (in OPERAS there is a core group and a clear governance, yet the players have freedom to make their own decisions)

https://openmethods.dariah.eu/

400+ Tools and innovations in scholarly communication:
https://docs.google.com/spreadsheets/d/1KUMSeq_Pzp4KveZ7pb5rddcssk1XBTiLHniD0d3nDqo/edit#gid=1519702055

OPERAS: bringing the long tail of SSH into EOSC (ElenaGiglia's paper on JLIS):
http://dx.doi.org/10.4403/jlis.it-12523

High Integration of Research Monographs in the European Open Science (**HIRMEOS**) project: https://www.hirmeos.eu

OPERAS white papers (and other works) in the Zenodo OPERAS Community
https://zenodo.org/communities/operaseu/?page=1&size=20

ISIDORE: https://isidore.science/

b-on: https://www.b-on.pt/en/__

UC Digitalis: https://digitalis.uc.pt/en

Open Monograph Press: https://pkp.sfu.ca/omp/