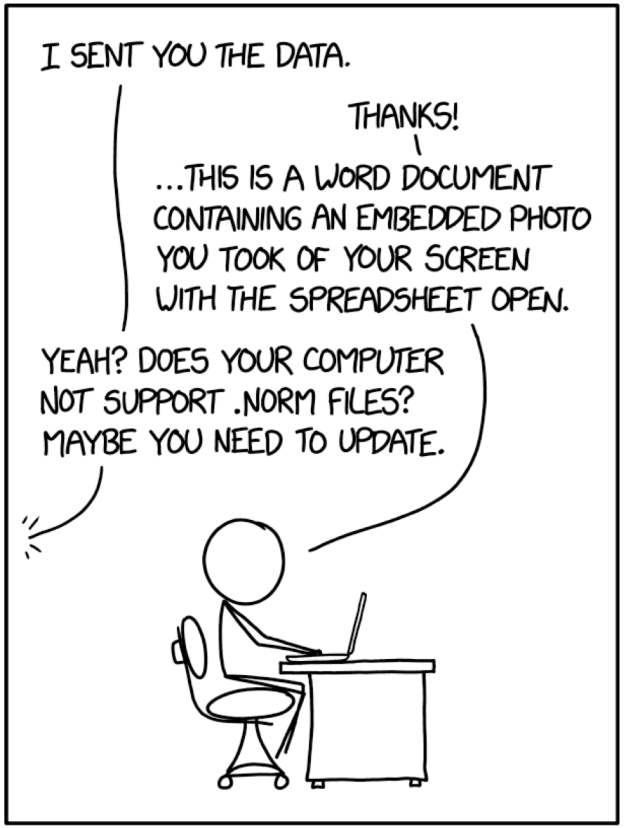
Documenting workflow and data management

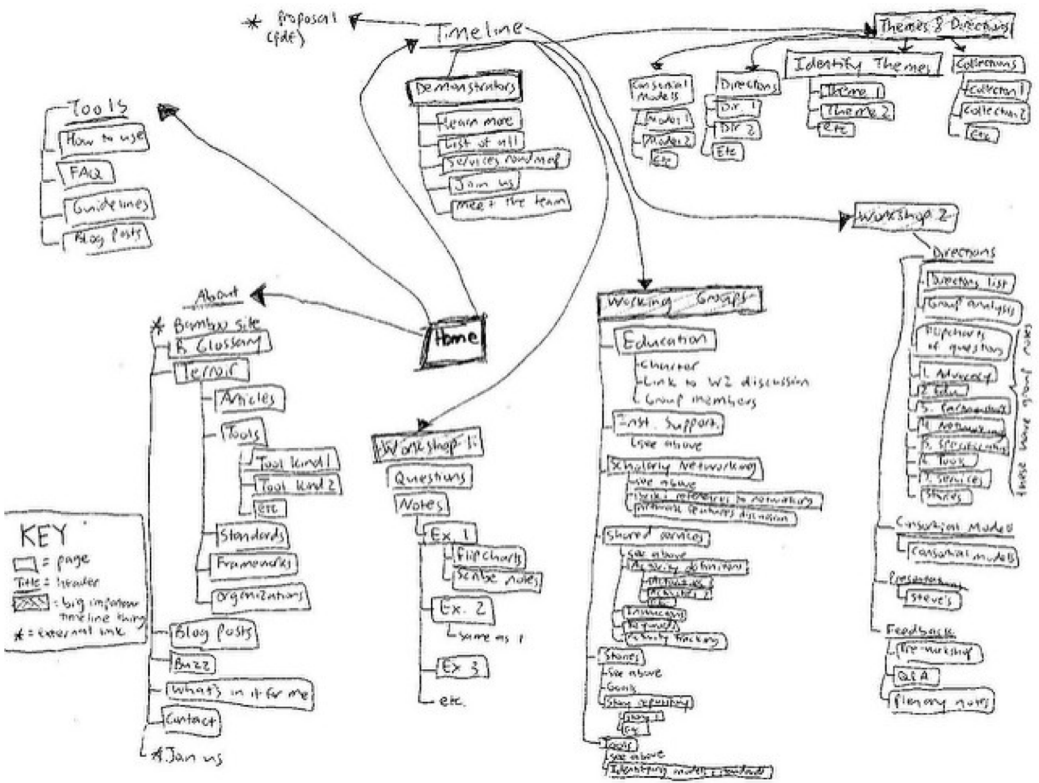
Quinn Dombrowski
Ron Nakao

CIDR Workshop - February 26, 2019

Does this look familiar?



A documented workflow is the best present you can give your future self



Data lifecycle

Source: https://www.ukdataservice.ac.uk/manage-data/plan/checklist.aspx

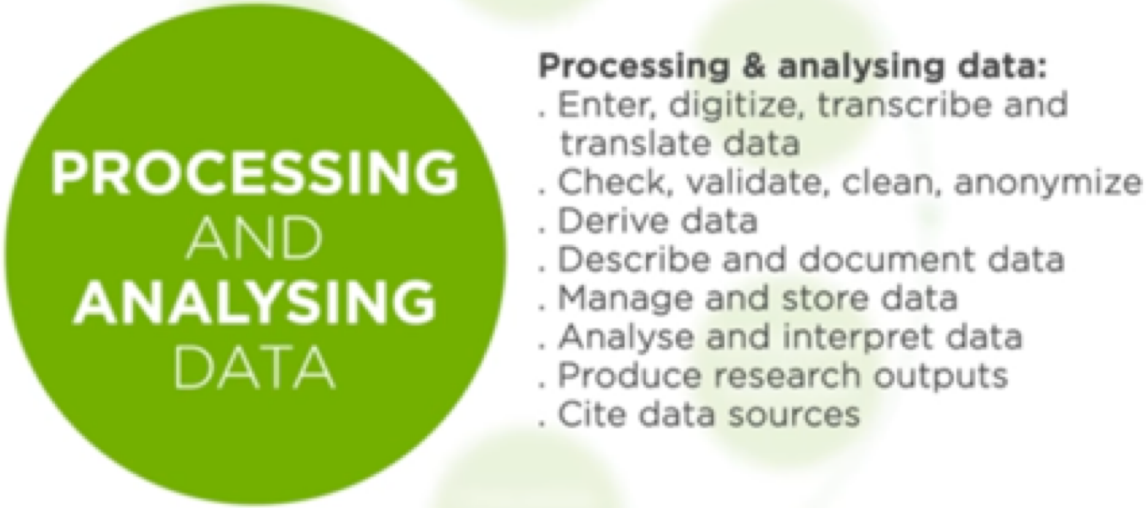
Planning Research

Source: https://www.ukdataservice.ac.uk/manage-data/plan/checklist.aspx

Collecting Data

Source: https://www.ukdataservice.ac.uk/manage-data/plan/checklist.aspx

Processing and Analyzing Data

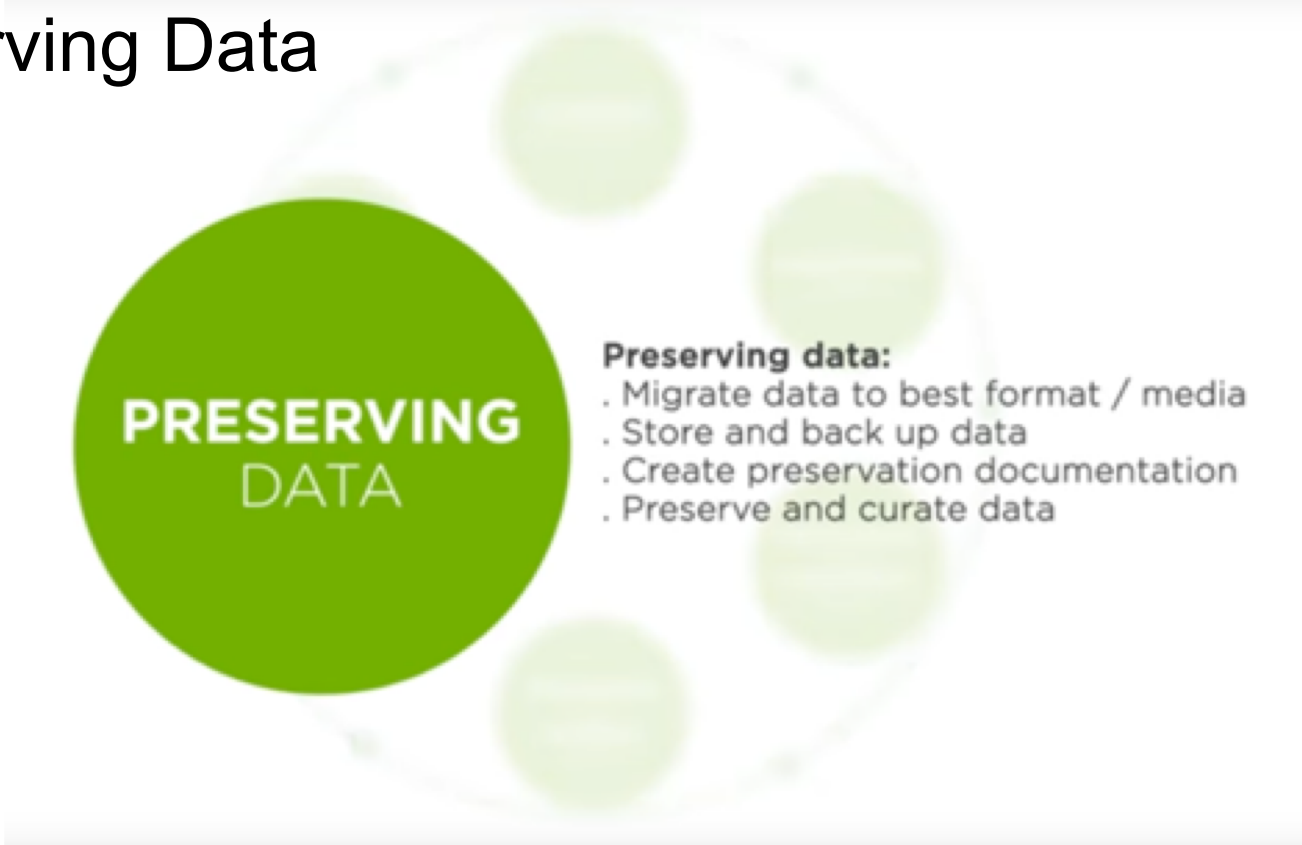


Source: https://www.ukdataservice.ac.uk/manage-data/plan/checklist.aspx

Publishing and Sharing Data

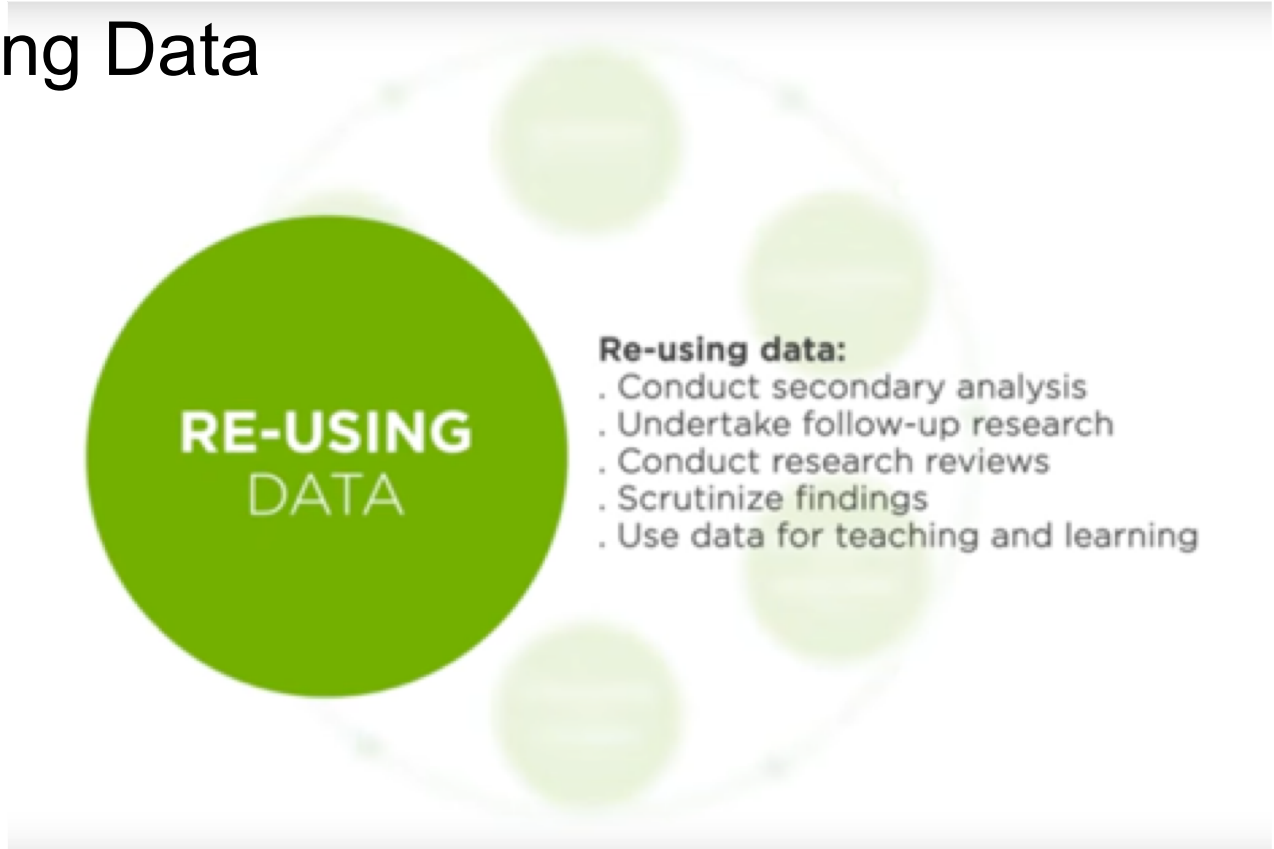
Source: https://www.ukdataservice.ac.uk/manage-data/plan/checklist.aspx

Preserving Data



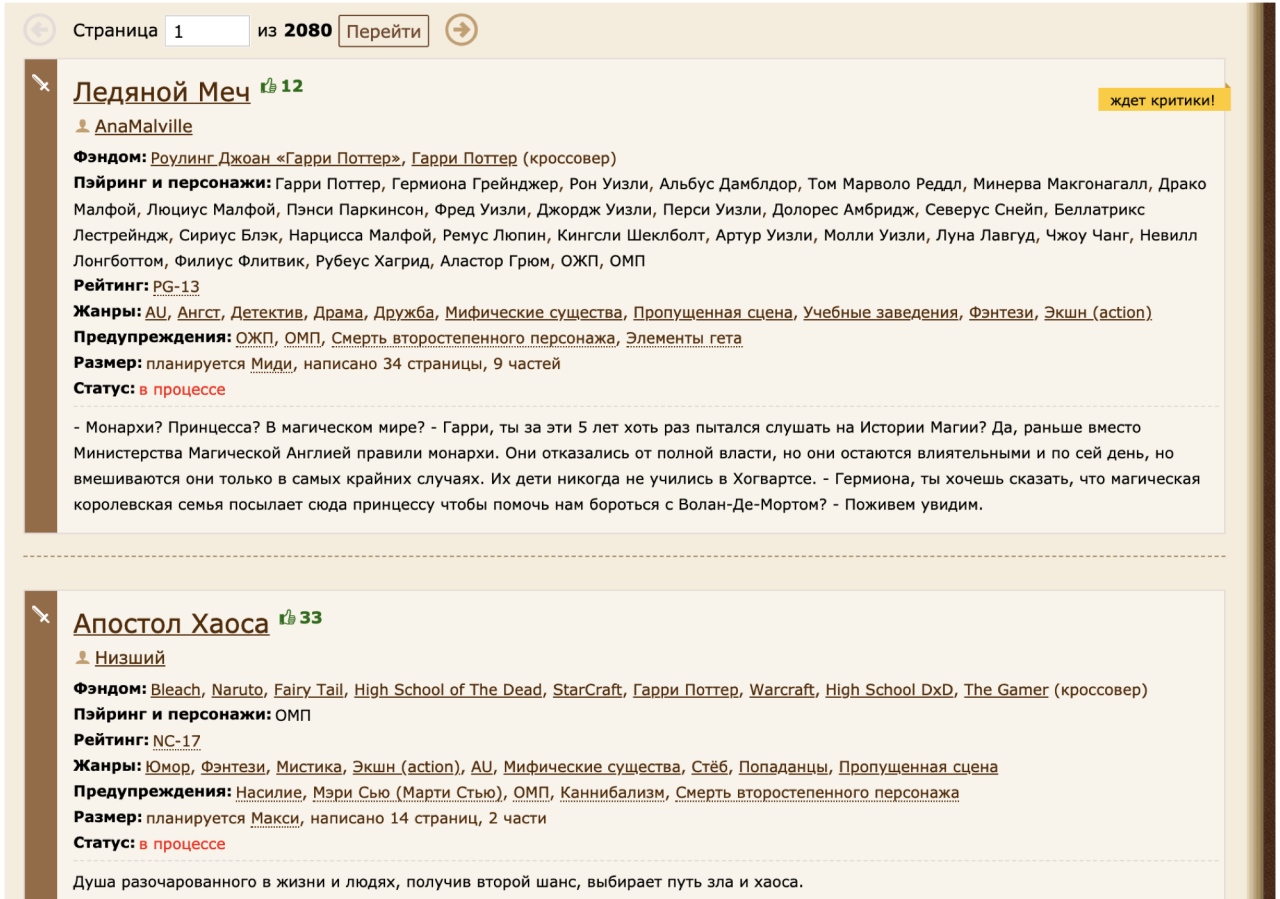
Source: https://www.ukdataservice.ac.uk/manage-data/plan/checklist.aspx

Re-Using Data



Source: https://www.ukdataservice.ac.uk/manage-data/plan/checklist.aspx

Case study: Multilingual Harry Potter fanfic



Страница 1 из **2080** Перейти

Ледяной Меч 12

ждет критики!

AnaMalville

Фэндом: Роулинг Джоан «Гарри Поттер», Гарри Поттер (кроссовер)
Пэйринг и персонажи: Гарри Поттер, Гермиона Грейнджер, Рон Уизли, Альбус Дамблдор, Том Марволо Реддл, Минерва Макгонагалл, Драко Малфой, Люциус Малфой, Пэнси Паркинсон, Фред Уизли, Джордж Уизли, Перси Уизли, Долорес Амбридж, Северус Снейп, Беллатрикс Лестрейндж, Сириус Блэк, Нарцисса Малфой, Ремус Люпин, Кингсли Шеклболт, Артур Уизли, Молли Уизли, Луна Лавгуд, Чжоу Чанг, Невилл Лонгботтом, Филиус Флитвик, Рубеус Хагрид, Аластор Грюм, ОЖП, ОМП
Рейтинг: PG-13
Жанры: AU, Ангст, Детектив, Драма, Дружба, Мифические существа, Пропущенная сцена, Учебные заведения, Фэнтези, Экшн (action)
Предупреждения: ОЖП, ОМП, Смерть второстепенного персонажа, Элементы гета
Размер: планируется Миди, написано 34 страницы, 9 частей
Статус: в процессе

- Монархи? Принцесса? В магическом мире? - Гарри, ты за эти 5 лет хоть раз пытался слушать на Истории Магии? Да, раньше вместо Министерства Магической Англией правили монархи. Они отказались от полной власти, но они остаются влиятельными и по сей день, но вмешиваются они только в самых крайних случаях. Их дети никогда не учились в Хогвартсе. - Гермиона, ты хочешь сказать, что магическая королевская семья посылает сюда принцессу чтобы помочь нам бороться с Волан-Де-Мортом? - Поживем увидим.

Апостол Хаоса 33

Низший

Фэндом: Bleach, Naruto, Fairy Tail, High School of The Dead, StarCraft, Гарри Поттер, Warcraft, High School DxD, The Gamer (кроссовер)
Пэйринг и персонажи: ОМП
Рейтинг: NC-17
Жанры: Юмор, Фэнтези, Мистика, Экшн (action), AU, Мифические существа, Стёб, Попаданцы, Пропущенная сцена
Предупреждения: Насилие, Мэри Сью (Марти Стью), ОМП, Каннибализм, Смерть второстепенного персонажа
Размер: планируется Макси, написано 14 страниц, 2 части
Статус: в процессе

Душа разочарованного в жизни и людях, получив второй шанс, выбирает путь зла и хаоса.

Case study: Multilingual Harry Potter fanfic

7 sort by characters (E). English values aren't characters; select those rows and drag them over to the next column.

6 copy just the bad ones to a new doc, add headers

5 sort by status — 1835 don't need cleanup woo

4 add headers in excel

3 Use text mate to copy to file: https://stackoverflow.com/questions/4146421/textmate-save-result-of-regex-search-into-new-window

Case study: Multilingual Harry Potter fanfic

E12 =D12/691

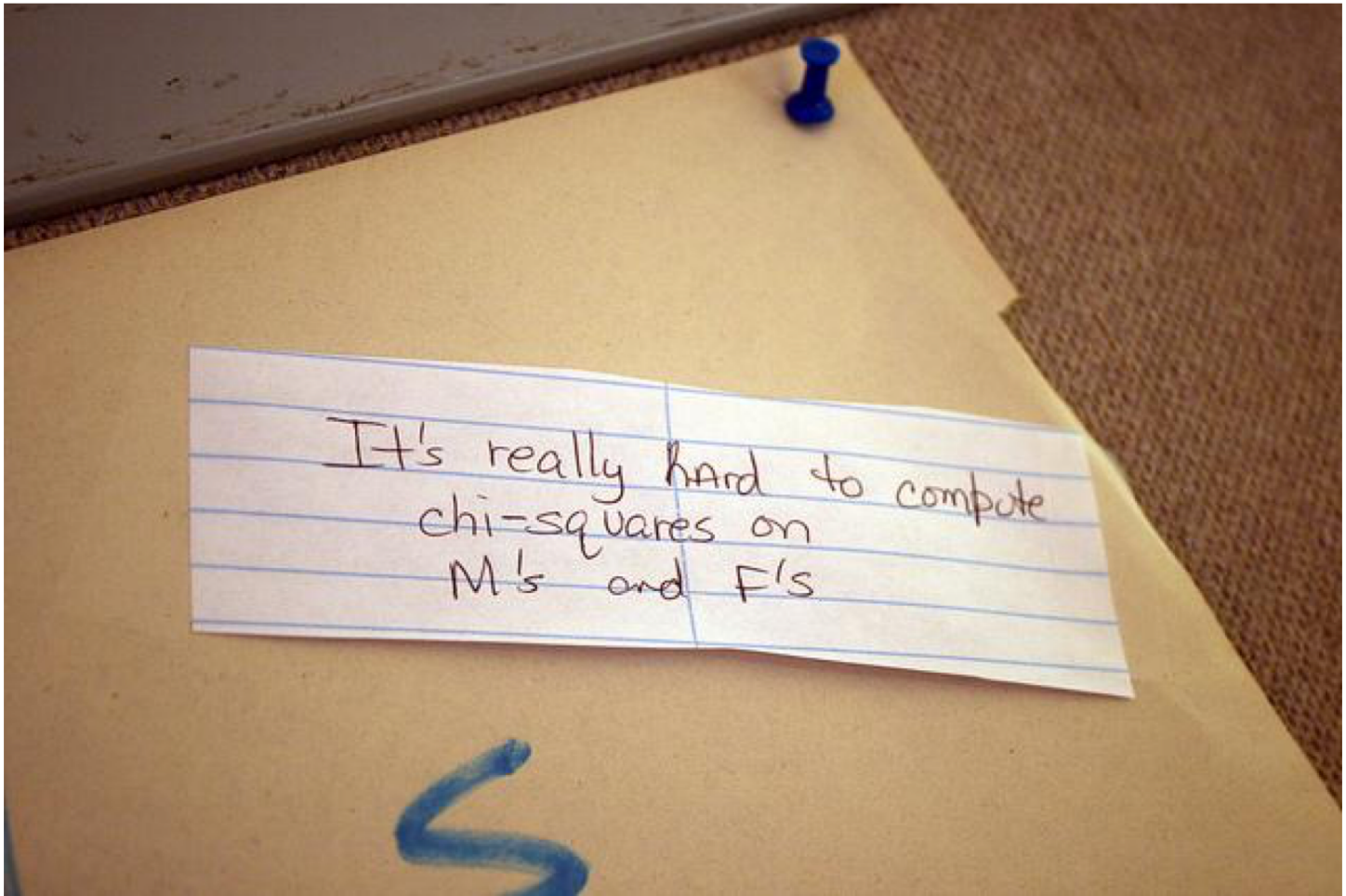
	A	B	C	D	E	F	G	H	I	J	K
1		HP first-eng	Other first-eng	Total En	TOTAL % EN	HP first-Ru	Other first-Ru	Total Ru	TOTAL % RU	Russian name	Fem HP
2	Ginny W.	101	14	115	16.64%	78	12	90	19.87%	Джинни Уизли	
3	Hermione G.	129	10	139	20.12%	57	9	66	14.57%	Гермиона Грейнджер	
4	Amelia B.	1	0	1	0.14%			0	0.00%		
5	Astoria G.	2	0	2	0.29%			0	0.00%		
6	Bellatrix L.	1	0	1	0.14%	4		4	0.88%	Беллатрикс Лестрейндж	1
7	Cedric D.	6	1	7	1.01%			0	0.00%		
8	Charlie W.	4	0	4	0.58%			0	0.00%		
9	Daphne G.	42	2	44	6.37%	15	2	17	3.75%	Дафна Гринграсс	
10	Draco M.	96	50	146	21.13%	74	30	104	22.96%	Драко Малфой	
11	Fenrir G.	5	0	5	0.72%			0	0.00%		
12	Fleur D.	16	0	16	2.32%	14	0	14	3.09%	Флёр Делакур	
13	Gabrielle D.	2	0	2	0.29%			0	0.00%		
14	George W.	3	0	3	0.43%			0	0.00%		
15	Hedwig	1	0	1	0.14%			0	0.00%		
16	Katie B.	1	0	1	0.14%			0	0.00%		
17	Kingsley S.	1	0	1	0.14%	1		1	0.22%	Кингсли Шеклболт	
18	Lucius M.	5	0	5	0.72%			0	0.00%		
19	Luna L.	14	1	15	2.17%	2	0	2	0.44%	Луна Лавгуд	
20	N. Tonks	9	0	9	1.30%	4	0	4	0.88%	Нимфадора Тонкс	
21	Neville L.	4	0	4	0.58%			0	0.00%		
22	OC	45	6	51	7.38%	15	2	17	3.75%	ОЖП (f), НЖП	1
23	Oliver W.	4	0	4	0.58%			0	0.00%		
24	Pansy P.	5	0	5	0.72%	4	1	5	1.10%	Панси Паркинсон	
25	Rabastan L.	1	0	1	0.14%			0	0.00%		
26	Remus L.	2	0	2	0.29%			0	0.00%		
27	Salazar S.	1	0	1	0.14%			0	0.00%		
28	Seamus F.	1	0	1	0.14%			0	0.00%		
29	Severus S.	34	5	39	5.64%	36	53	89	19.65%	Северус Снейп	2
30	Sirius B.	2	0	2	0.29%			0	0.00%		
31	Susan B.	2	0	2	0.29%			0	0.00%		
32	Theodore N.	4	0	4	0.58%			0	0.00%		
33	Tom R. Jr.	33	3	36	5.21%	11	16	27	5.96%	Том Марволо Реддл	5
34	Voldemort	20	2	22	3.18%	2	2	4	0.88%	Волан-де-Морт	
35	harem			0	0.00%	2		2	0.44%	гарем	
36	Cho C.			0	0.00%	5	1	6	1.32%	Чжоу Чанг	

Case study: Multilingual Harry Potter fanfic

- Date web scraper was run
  - Web scraper JSON code
- Output CSV from web scraper
- Workflow clean-up steps (text/PDF)
- Output file from clean-up
- Workflow for analysis
- Spreadsheets used for analysis (Excel, CSV)
- Any additional code used for analysis

What even **is** “data”? Do you have any?

In DH, “data” is contextual



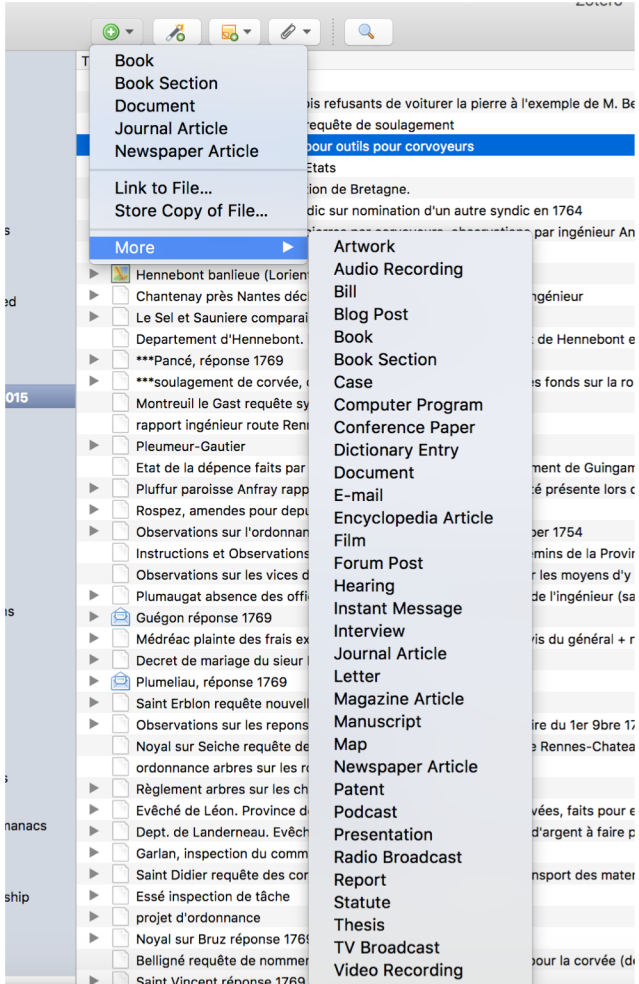
“Data” creates fissures

Your documentation is a map to reconstruct reality

Special considerations for archival materials

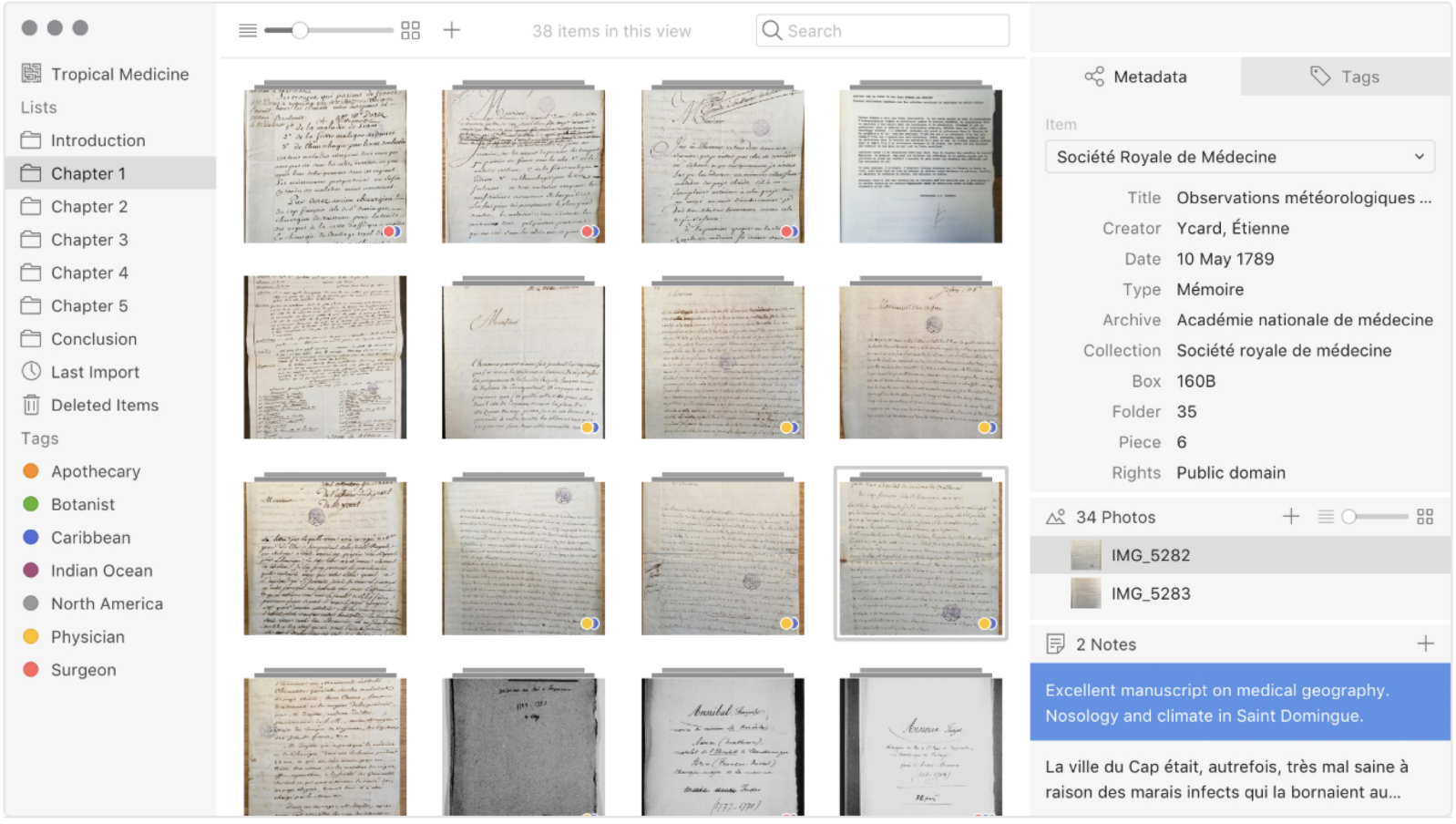
- How can you minimize the hassle for other people to find these materials?
  - Your documentation should indicate finding aids and other resources used
- Keep note of what boxes/files/etc. do *not* have what you're looking for
- Follow conventions for your field (or related fields) for granularity of citation:
  - Box #
  - Dossier/files #
  - Document #
  - Page # (recto/verso)
  - Paragraph/article/chapter/etc.

Tools for archival materials



Item Type	Document
Title	Reglement pour les digues de Dol (Arrest du Conseil touchant l'adjuication et renable des ouvrages Publics...Extrait des Registres du conseil d'Etat du Roy)
Author	Colbert, (first)
Author	Guillard (commis a... , (first)
Abstract	manuscript copy
Publisher	
Date	26 October 1701 d m y
Language	
Short Title	
URL	
Accessed	
Archive	AD 35
Loc. in Archive	C 2267
Library Catalog	
Call Number	
Rights	
Extra	
Date Added	7/26/2010, 4:58:22 AM
Modified	2/11/2013, 2:31:59 PM

Tools for archival materials



Tools for archival materials

ZOTERO

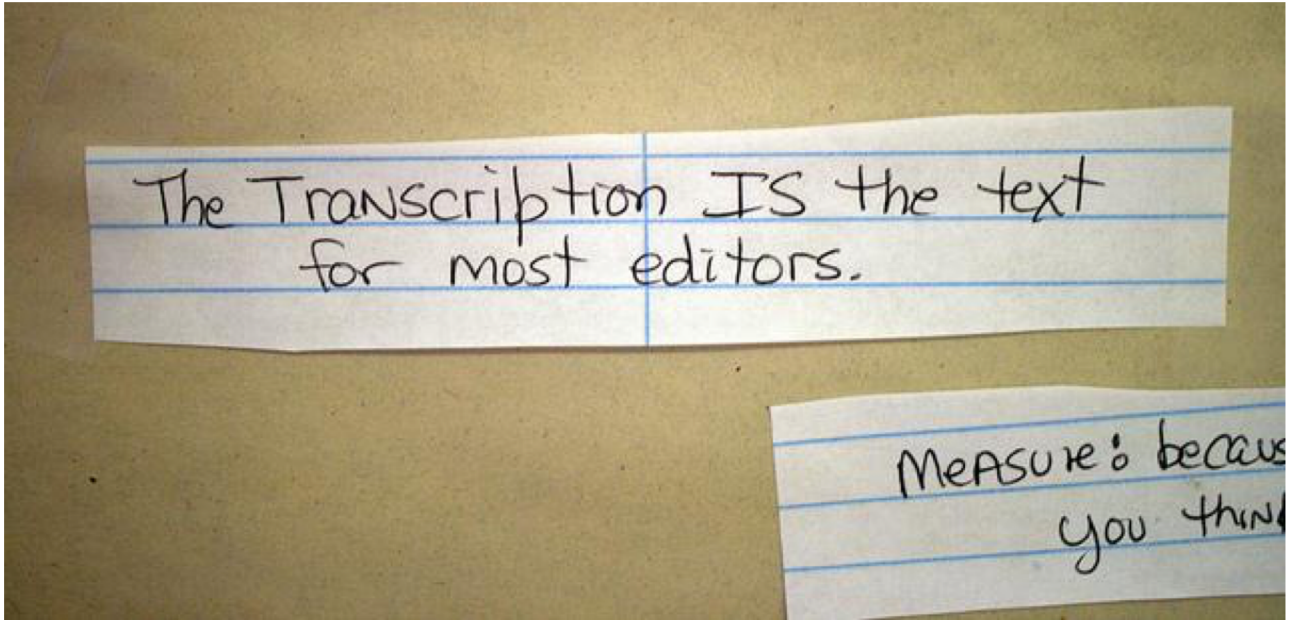
- Free!
- Accommodates many defined types of items
- Can automatically import metadata w/browser connector or identifiers (doi)
- Items can be linked to local files or attached directly in Zotero database
- Notes field
- Tags and related items – Sort & search functions
- Shareable (online libraries for groups)
- Export bibliographies, reports

TROPY

- Also free!
- Permits metadata field customization
- Connects to locally stored photos

Page images for archival material

- You may not need page images: notes could be enough
- 300 dpi for 8-12 pt font
- 400-600 dpi for very small text
- Color is better than grayscale is better than B&W



Social Science Data is...

- the evidence required to answer your research question
- measures and constructions (operationalization) of concepts
- dependent on the methodology used to answer your research question
- found in many places, sizes, shapes, and formats
- costly to collect, curate, archive, and share
- often not available

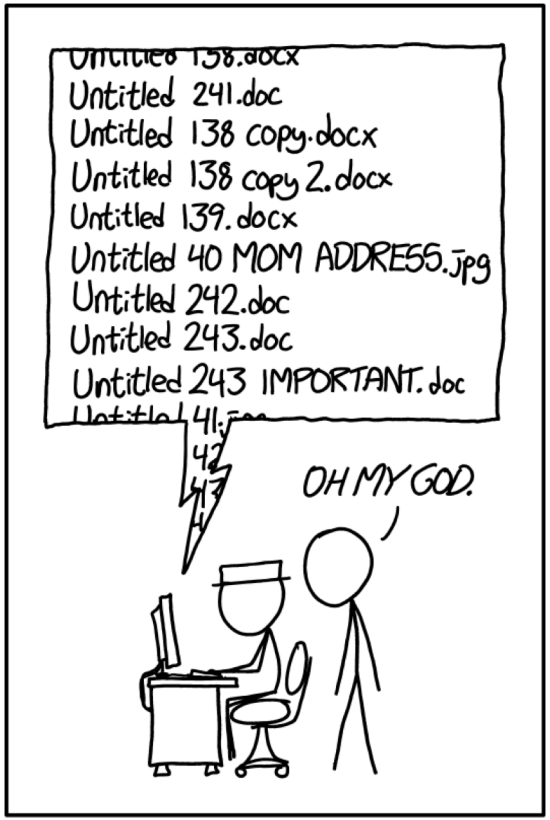
Different Data Types

- Research vs. Administrative
- Primary vs. Secondary
- Aggregate/Tabular/Macrodata vs. Sample/Microdata
- Quantitative (numeric) vs. Qualitative (non-numeric)
- Structured vs. Unstructured
- Restricted/Confidential/Licensed/Copyright vs. Public/Open Access/Open Source

Numeric Data Storage (Form Factors)

- Printed page
- Punch cards
- Tapes
- Floppy disks
- CD/DVD
- Internal/External hard drives or USB flash drives
- Internet/Cloud

File naming conventions



The specifics usually matter less than just having some.

File naming conventions: general guidelines

- Files should be named consistently
- File names should be short but descriptive (<25 characters)
- Avoid special characters or spaces in a file name
- Use capitals and underscores instead of periods or spaces or slashes
- Use date format ISO 8601: YYYYMMDD
- Include a version number
- Write down naming convention in your documentation
    a. Describe the logic behind the file naming system for your project.
    b. Give examples of the file names, from different types of digital data used in your research.
    c. Explain any non-obvious coding or numbering systems

File structure

- Have a folder for your research project
- If you’re collaborating with others, make sure they’re saving their files in the same place, with the same naming and organization conventions
- Document your file structure

Where can you store data? How do you do it?

Data Storage Options at Stanford & Beyond

- Local computer, external Hard Drive, USB flash drive
- Stanford Google Drive, Google Team Drive (L,M,H)
  - https://uit.stanford.edu/service/gsuite/drive
  - https://uit.stanford.edu/service/gsuite/teamdrive
- Stanford Box (L,M), Stanford Medicine Box (H)
  - https://uit.stanford.edu/service/box
- Stanford Google Drive, Team Drive, Box Comparison
  - https://uit.stanford.edu/service/gsuite/drive/comparison
- Stanford Medicine Redcap (H)
  - https://med.stanford.edu/researchit/infrastructure/redcap.html

Data Storage Options at Stanford & Beyond (cont'd)

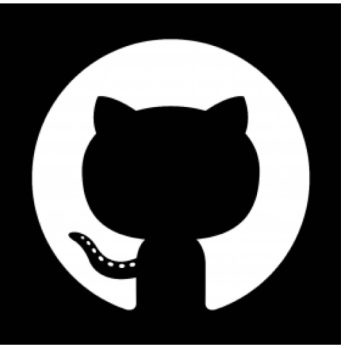
- Stanford Shared Computing Environment: Farmshare
  - https://uit.stanford.edu/service/sharedcomputing
- Stanford UIT File and Data Storage
  - https://uit.stanford.edu/service/storage
- Stanford Servers and Data
  - https://uit.stanford.edu/services/category/servers-and-data
- School/Departmental Servers
- Cloud: AWS, Microsoft Azure, Google Cloud
  - https://uit.stanford.edu/cloud-vendor/reduce-cost

Some Data Archives and Data Sharing Sites

- Stanford Digital Repository (SDR)
  - http://library.stanford.edu/research/stanford-digital-repository
- Stanford Social Science Data and Software (SSDS) Collection
  - https://data.stanford.edu/
- Inter-university Consortium for Political and Social Research (ICPSR) Archive – Self Deposit
  - https://www.icpsr.umich.edu/icpsrweb/deposit/
- Harvard Dataverse — Self Deposit
  - https://dataverse.harvard.edu/

Github

https://github.com



tapas-TEI-files/

Samples, templates, and test files for use in TAPAS.

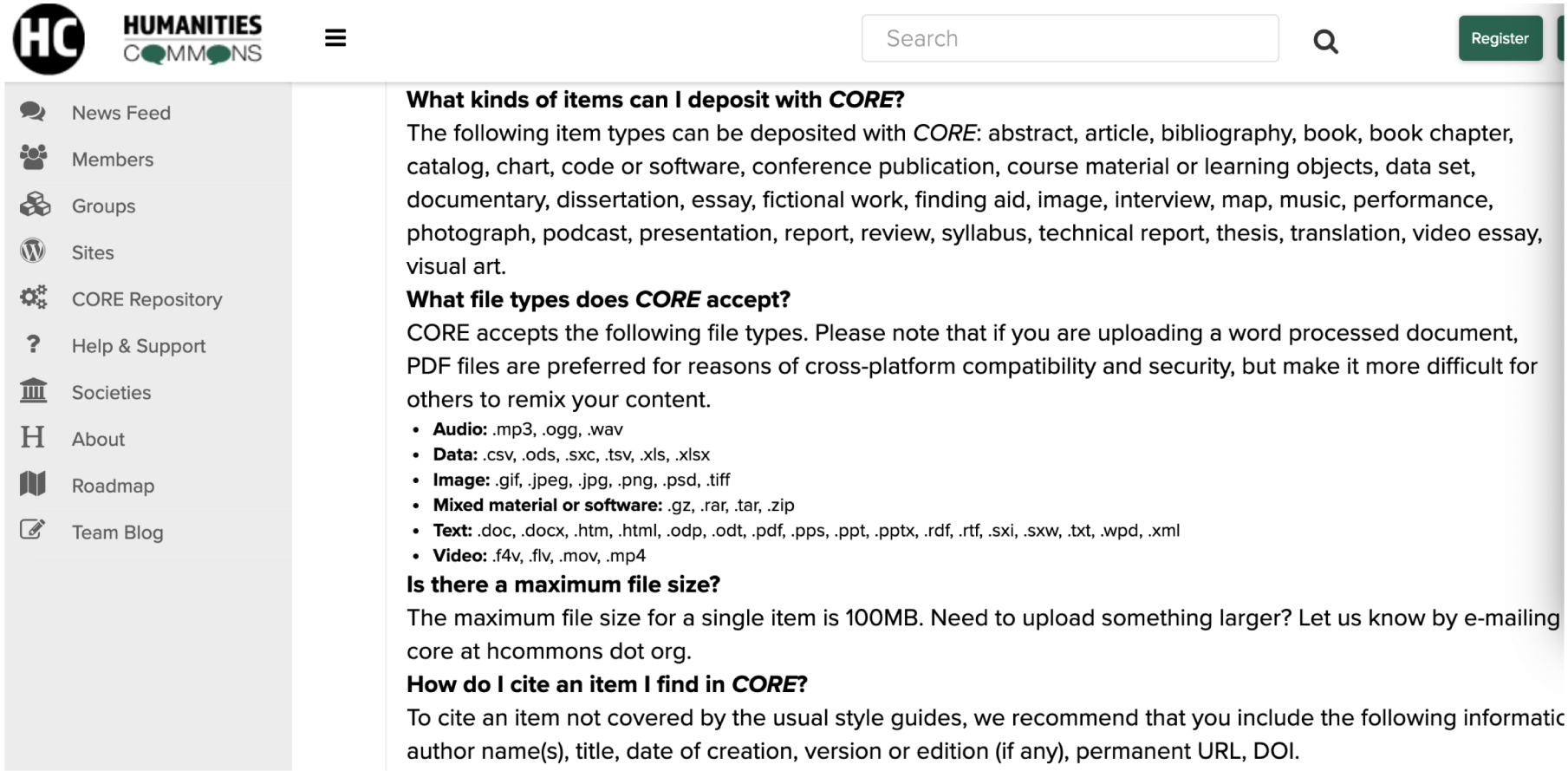
This repository contains files for use in testing TAPAS itself, and the source files used to generate the samples and templates that may be used to teach TEI or as a starting point for using TEI.

General layout

- Raw files: These are the files from which the samples and templates (below) are generated. If you want to make changes to a sample or template file, make the changes here.
- Samples: These are simple examples of valid TEI files and should look reasonably good in at least one TAPAS viewpackage. They are generated (from the files in raw_files/, see above), so you probably don't want to make any changes here. (Note: sample files may be hand-tweaked after generation, usually to improve whitespace.)
- Templates: These are templates of TEI files. I.e., they are incomplete TEI files that may be used as a starting point and "filled out". Typically these files have comments indicating where you would want to insert data. Because there is information missing, these files may be invalid and may not look acceptable in TAPAS if used as-is. (But should be valid and look OK after having been filled out.) These files are generated (from the files in raw_files/, see above), so you probably don't want to make any changes here. (Note: template files may be hand-tweaked after generation, usually to improve whitespace.)
- Tests: Files used internally to test TAPAS. See the README file in that directory for more details.

Humanities Commons

https://hcommons.org



Secure Computing

Secure Computing at Stanford

- Stanford Research Compliance Office: Human Subjects and IRB
  - https://researchcompliance.stanford.edu/panels/hs
- Stanford Secure Computing Practices
  - https://uit.stanford.edu/guide/security
- *Stanford Data/Systems Risk Classifications
  - https://uit.stanford.edu/guide/riskclassifications
- *Stanford Minimum Security Standards for Endpoints, Servers, and Applications
  - https://uit.stanford.edu/guide/securitystandards#security-standards-endpoints

Secure Computing at Stanford (cont'd)

- Stanford Encryption Options (for your computer)
  - https://uit.stanford.edu/guide/encrypt
- Stanford Secure File Transfer Protocol (FTP) Software: Fetch (Mac) & SecureFX (Windows)
  - https://uit.stanford.edu/software
- Stanford VPN (Virtual Private Network)
  - https://uit.stanford.edu/service/vpn/

When and how do you cite data?

Joint Declaration of Data Citation Principles

1. **Importance**
Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

2. **Credit and Attribution**
Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

https://doi.org/10.25490/a97f-egyk

Joint Declaration of Data Citation Principles

3. **Evidence**
In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

4. **Unique Identification**
A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

5. **Access**
Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

https://doi.org/10.25490/a97f-egyk

Joint Declaration of Data Citation Principles

6. **Persistence**
Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.

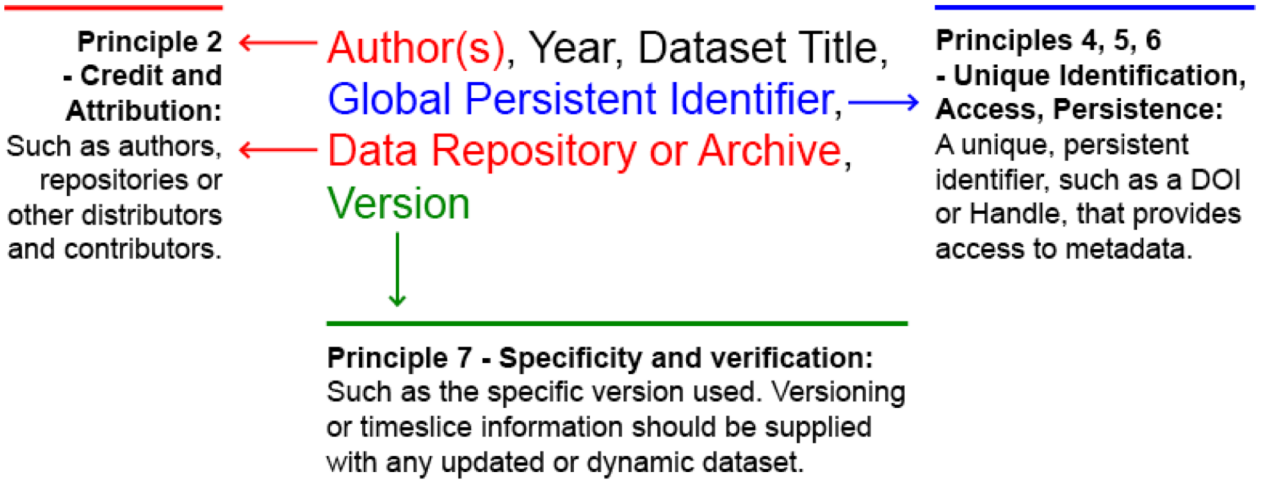
7. **Specificity and Verifiability**
Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

https://doi.org/10.25490/a97f-egyk

Joint Declaration of Data Citation Principles

8. **Interoperability and Flexibility**

Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.



https://doi.org/10.25490/a97f-egyk

Questions?