Data Management and working with texts in Digital Humanities

Christopher Ohge, Martin Steer
Riga Technical University, September 2019

Outline

- What is Data Management?
- Data formats, metadata, coding
- Storage, preservation and sustainability
- Sharing and reuse

What is Data Management?

What is Data Management?

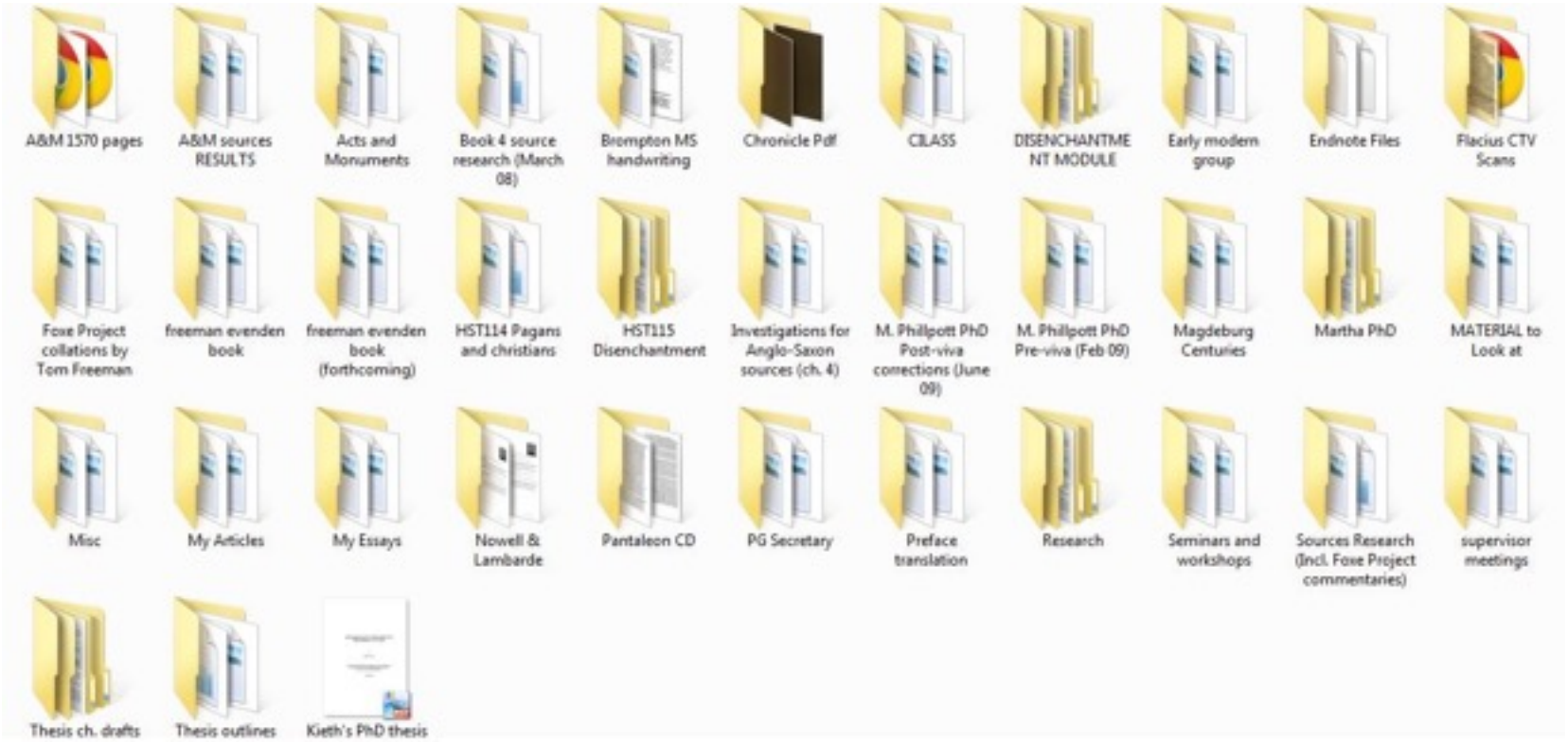
"If you go into a project not understanding how you are going to manage that data or how you are going to organise it and retrieve it, then you are destined to struggle to keep the research under control and on track."

- Professor Glenn Burgess Pro-vice Chancellor, University of Hull

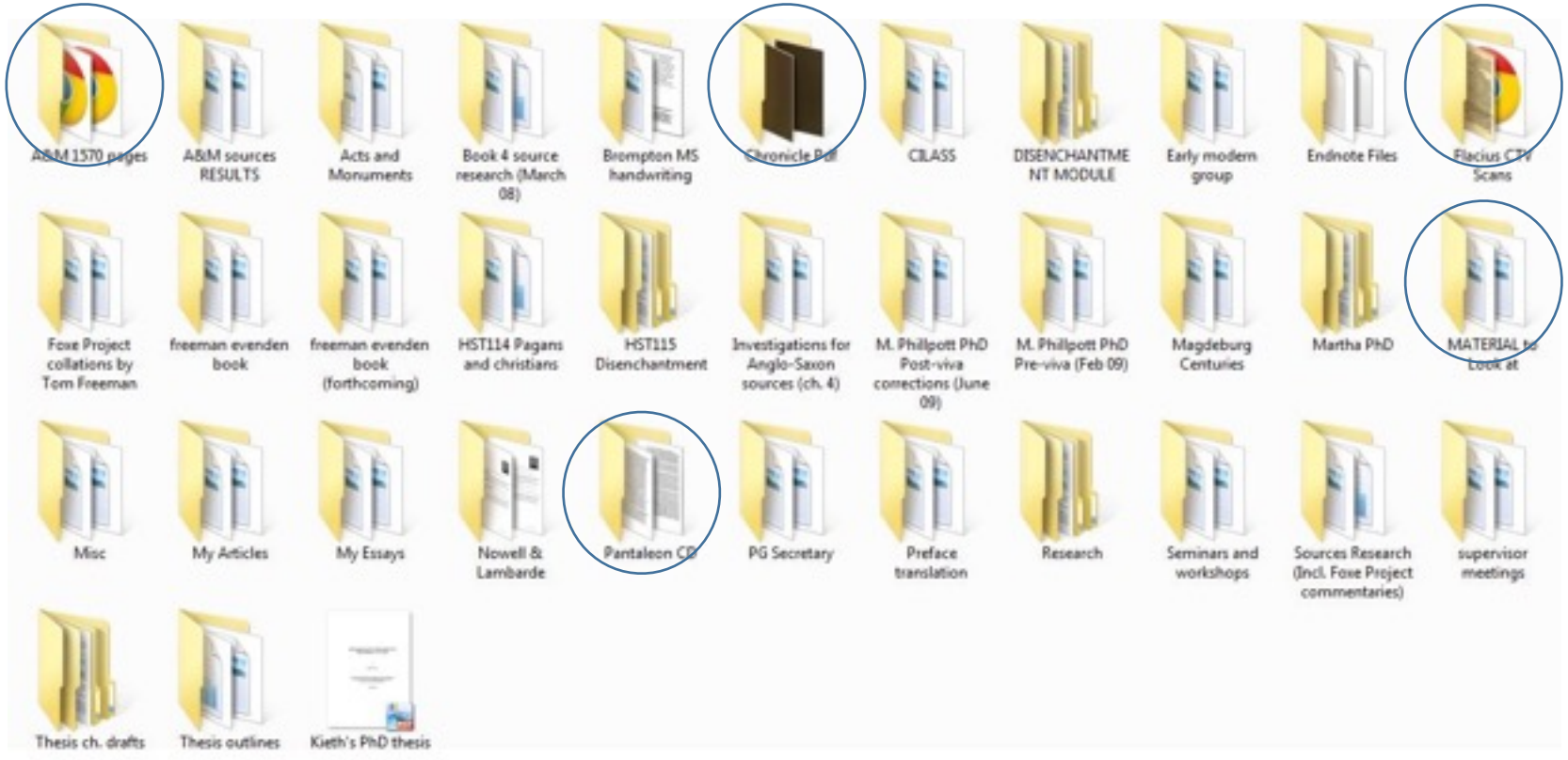
Good data management

- Saves time
- Increases your efficiency
- Helps you to preserve and protect your data
- Helps you to view data as an output in its own righ
- Meets grant requirements
- Helps to meet requirements of Open Acces
- Enables transparency and research

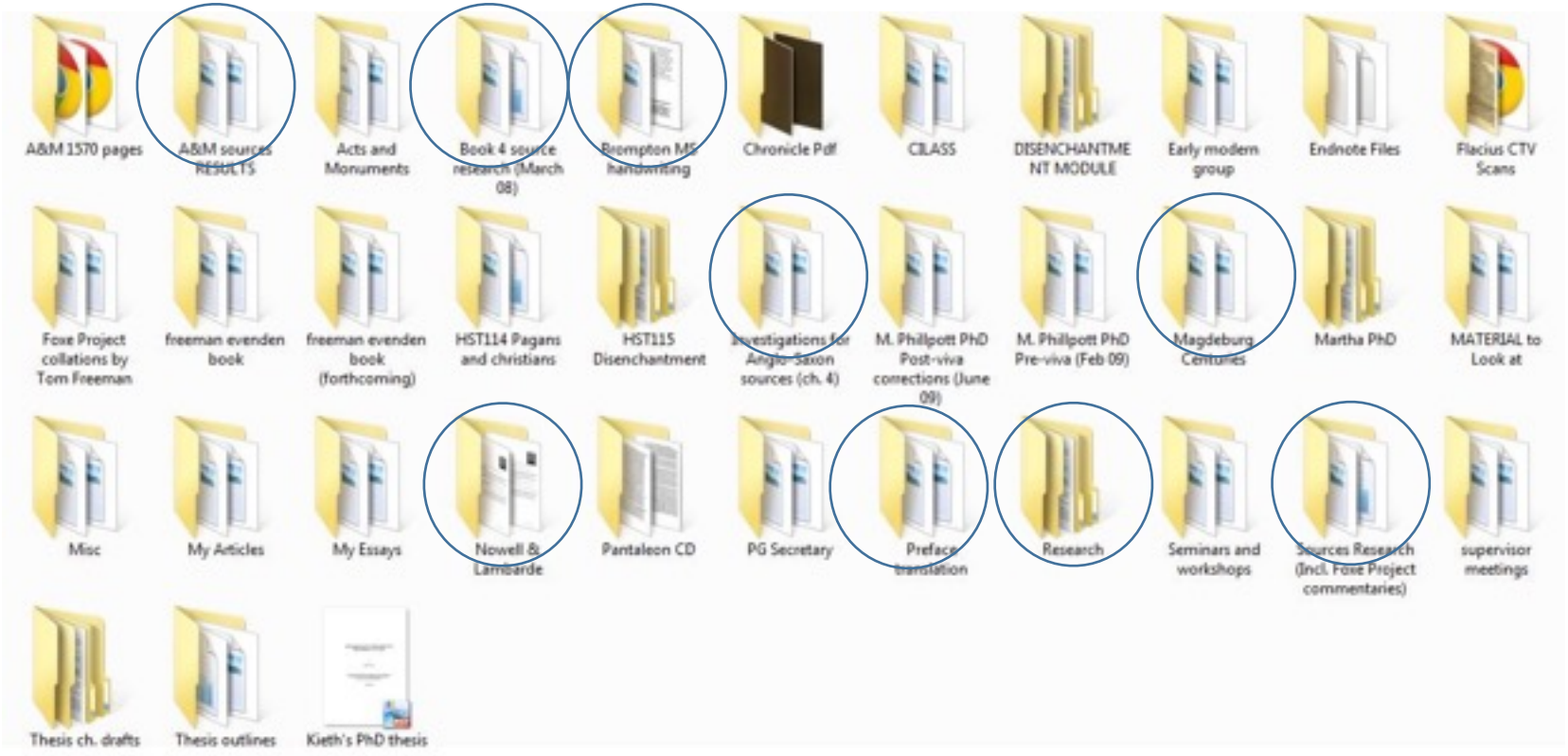
Example


A&M 1570 pages
A&M sources RESULTS
Acts and Monuments
Book 4 source research (March 08)
Brompton MS handwriting
Chronicle Pdf
CILASS
DISENCHANTMENT MODULE
Early modern group
Endnote Files
Flacius CTV Scans
Foxe Project collations by Tom Freeman
freeman evenden book
freeman evenden book (forthcoming)
HST114 Pagans and christians
HST115 Disenchantment
Investigations for Anglo-Saxon sources (ch. 4)
M. Phillpott PhD Post-viva corrections (June 09)
M. Phillpott PhD Pre-viva (Feb 09)
Magdeburg Centuries
Martha PhD
MATERIAL to Look at
Misc
My Articles
My Essays
Nowell & Lambarde
Pantaleon CD
PG Secretary
Preface translation
Research
Seminars and workshops
Sources Research (Incl. Foxe Project commentaries)
supervisor meetings
Thesis ch. drafts
Thesis outlines
Kieth's PhD thesis

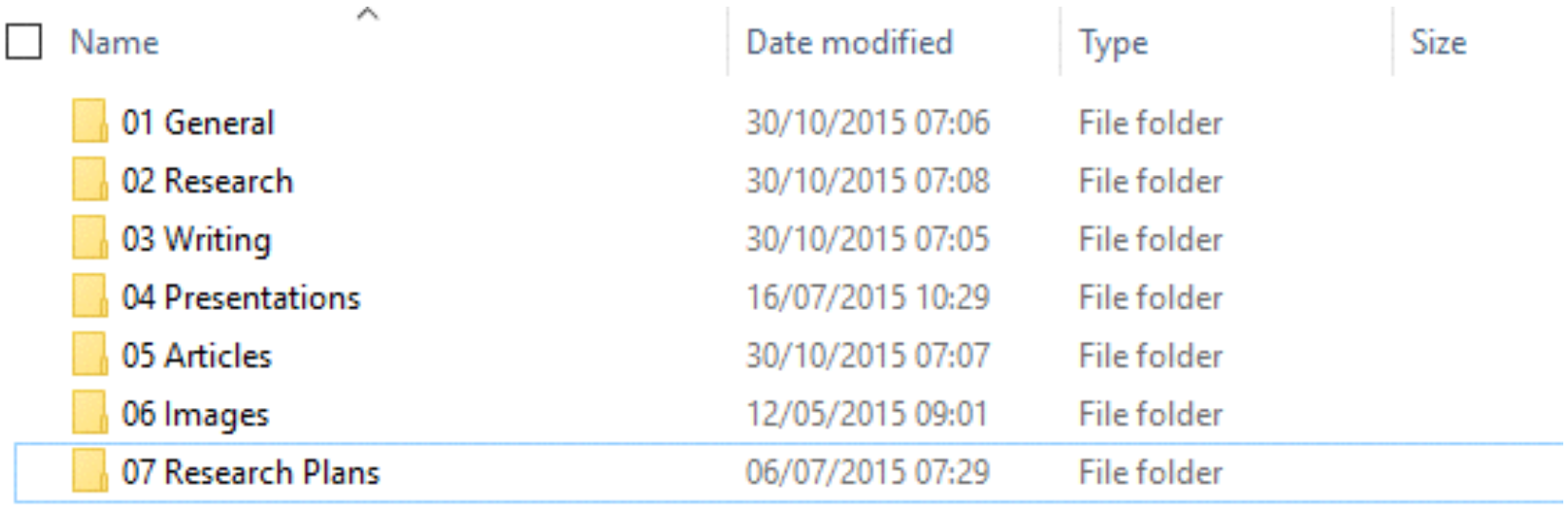
Gathered data



Created data



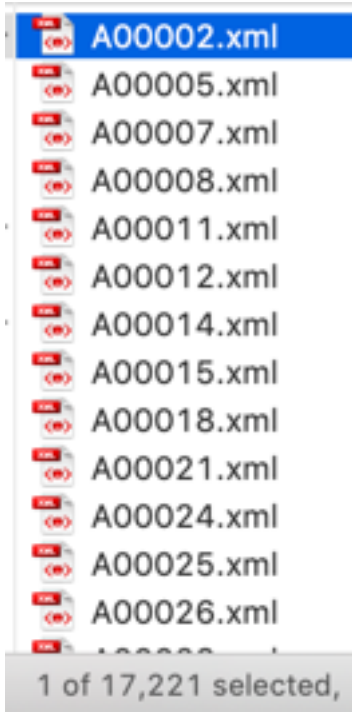
My folder system today



Name	Date modified	Type	Size
01 General	30/10/2015 07:06	File folder	
02 Research	30/10/2015 07:08	File folder	
03 Writing	30/10/2015 07:05	File folder	
04 Presentations	16/07/2015 10:29	File folder	
05 Articles	30/10/2015 07:07	File folder	
06 Images	12/05/2015 09:01	File folder	
07 Research Plans	06/07/2015 07:29	File folder	

EEBO-TCP corpus structure

Name	Date Modified	Size	Kind
cloneall.sh	29 Oct 2015 at 17:24	2 MB	Shell Script
graball.sh	29 Oct 2015 at 17:24	2.9 MB	Shell Script
identifiers.txt	29 Oct 2015 at 17:24	4 KB	Plain Text Document
LICENSE	29 Oct 2015 at 17:24	7 KB	TextEdit.app Document
README.md	29 Oct 2015 at 17:24	54 bytes	Markdown File
TCP.csv	29 Oct 2015 at 17:24	29.3 MB	CSV Document
TCP.json	29 Oct 2015 at 17:24	40.2 MB	JSON
tcpchars.xml	29 Oct 2015 at 17:24	112 KB	XML Document



Scripts and metadata to download entire corpus:

https://github.com/textcreationpartnership/Texts

Corpora/Data packages – many standards!

- There is a movement towards Frictionless data
- Digital archives standards – MODS, MADS, PREMIS
- Still a lot of non-standard textual corpora

- You will have to plan how to structure and use your text!
- R libraries to help ease use – datapackage.r, tidytext, tibble, etc.

Data Management Plan

- *What is needed to validate the results of your research?*
- What data would you need to include for someone else to replicate your results?
  - Bibliography/citations
  - Access to the raw data
  - Access to modified/created data
  - Documentation (methodology/process used for creating datasets)
  - Data package standards
  - Which repository to publish

Data Management Plan

A data management plan (DMP) can perform a number of roles over the course of a research project.

- **A checklist** – a DMP acts a means of checking that everything that needs to be done to effectively manage the data you are working with is being done. It can be particularly useful at the start of a project to ensure you get up and running smoothly, but can also be applied at different stages of the project to check everything is proceeding as it should be.
- **A manual** – a DMP can go beyond a checklist and be used as a manual to guide you through different aspects of managing your data when needed. Establishing how different aspects of data management can or should be undertaken as part of setting up your research will enable you to confidently address data management steps and issues as they arise.
- **A record** – whilst a DMP is predominantly used for the purposes described above, it can also be used as a record of the data management activity you have undertaken. This can then act as a demonstration of good research practice, and also be part of the overall project documentation and output.

AHRC DMP Template

- Data Summary
- Data Collection
- Short-term Data Storage
- Long-term Data Storage
- Data Sharing
- Ethical and Legal Considerations

Arts and Humanities Research Council (AHRC): AHRC Data Management Plan

Data Summary

1. Briefly introduce the types of data the research will create. Why did you decide to use these data types?

Guidance:
When defining data types, consider the format/quality of the data and how you will make it as easy as possible to access the data?
Consult with yourinstitution's data support (e.g. library services, IT department)

Data Collection

2. Give details on the proposed methodologies that will be used to create the data. Advise how the project team selected will be suitable for the data/digital aspects of the work, including details of how the institution's data support teams may need to support the project

Short-term Data Storage

3. How will the data be stored in the short term?

Guidance:
You should consult with the institution's data support (e.g. library services, IT department).
By submitting the DMP you are confirming that:

- The institution is able to store the data appropriately during the lifecycle of the grant, the relevant people have been consulted and this has been considered and agreed
- The institution has considered all the risks, and storage will be in line with the institution's data management policy (provide a link to the policy if applicable)

3a What backup will you have in the in-project period to ensure no data is lost?

Long-term Data Storage

4. How the data will be stored in the long term?

Guidance:
For advice on data storage and sharing, including future planning for the data, see:
Digital Preservation Coalition Knowledge Base
Digital Curation Centre

4a. Where have you decided to store it, why is this appropriate?

4b. How long will it be stored for and why?

4c. Costs of storage – why are these appropriate? Costs related to long term storage will be permitted providing these are fully justified and relate to the project. Full justification must be provided in Justification of Resources (JoR)

Guidance:
Costs of preserving the data:
4C (Collaboration to clarify the costs of Curation)

https://dmponline.dcc.ac.uk/template_export/1148994747.p

AHRC DMP Template

- Data Summary
- Data Collection
- Short-term Data Storage
- Long-term Data Storage
- Data Sharing
- Ethical and Legal Considerations

- Project planning and lifecycles
- Data formats, metadata, coding
- Storage, preservation and sustainability
- Sharing and reuse
- Policies, ethics and security

https://dmponline.dcc.ac.uk/template_export/1148994747.p

AHRC DMP Template

- Data Summary
- Data Collection
- Short-term Data Storage
- Long-term Data Storage
- Data Sharing
- Ethical and Legal Considerations

- Project planning and lifecycles
- Data formats, metadata, coding
- Storage, preservation and sustainability
- Sharing and reuse
- Policies, ethics and security

https://dmponline.dcc.ac.uk/template_export/1148994747.p

AHRC DMP Template

- Data Summary
- Data Collection
- Short-term Data Storage
- Long-term Data Storage
- Data Sharing
- Ethical and Legal Considerations

- Project planning and lifecycles
- Data formats, metadata, coding
- Storage, preservation and sustainability
- Sharing and reuse
- Policies, ethics and security

https://dmponline.dcc.ac.uk/template_export/1148994747.p

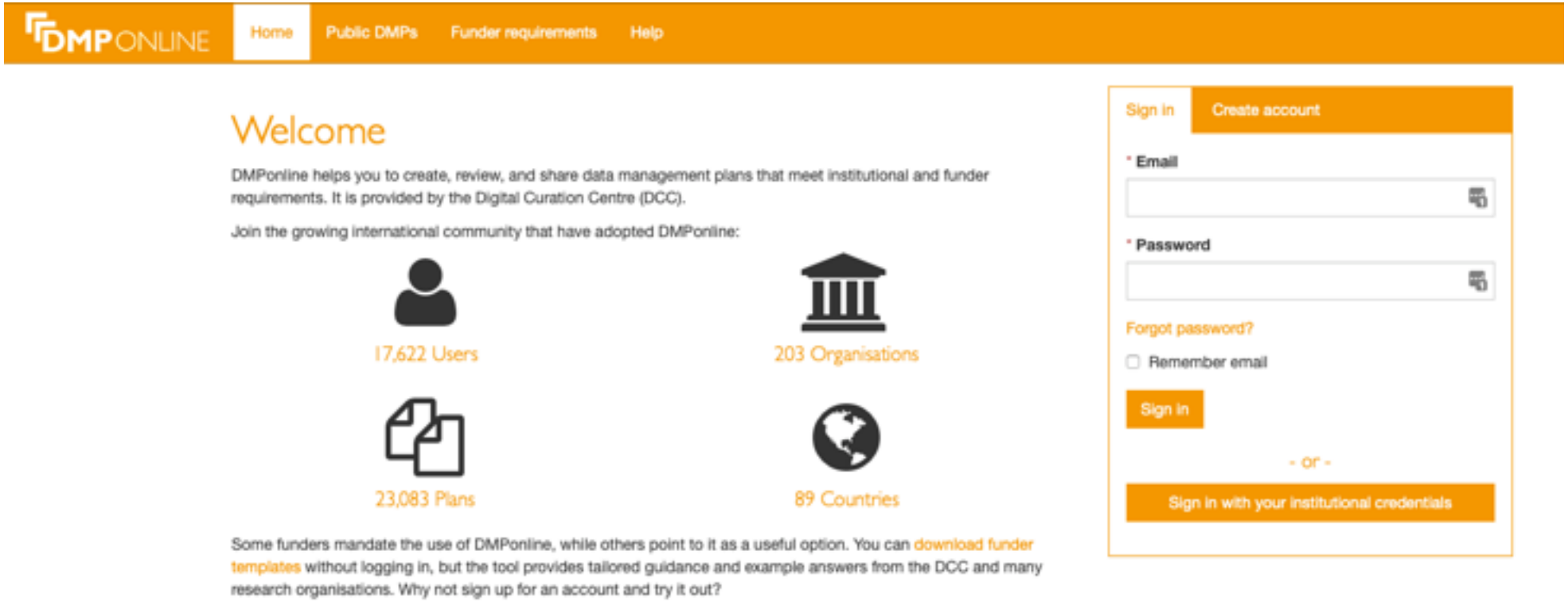
AHRC DMP Template

- Data Summary
- Data Collection
- Short-term Data Storage
- Long-term Data Storage
- Data Sharing
- Ethical and Legal Considerations

- Project planning and lifecycles
- Data formats, metadata, coding
- Storage, preservation and sustainability
- Sharing and reuse
- Policies, ethics and security

https://dmponline.dcc.ac.uk/template_export/1148994747.p

DMP Online - https://dmponline.dcc.ac.uk



Data formats, metadata, coding

Data formats, metadata, coding

- Data deluge
- Information architecture
- Naming things
- Providing structures
- Make it usable

Data deluge

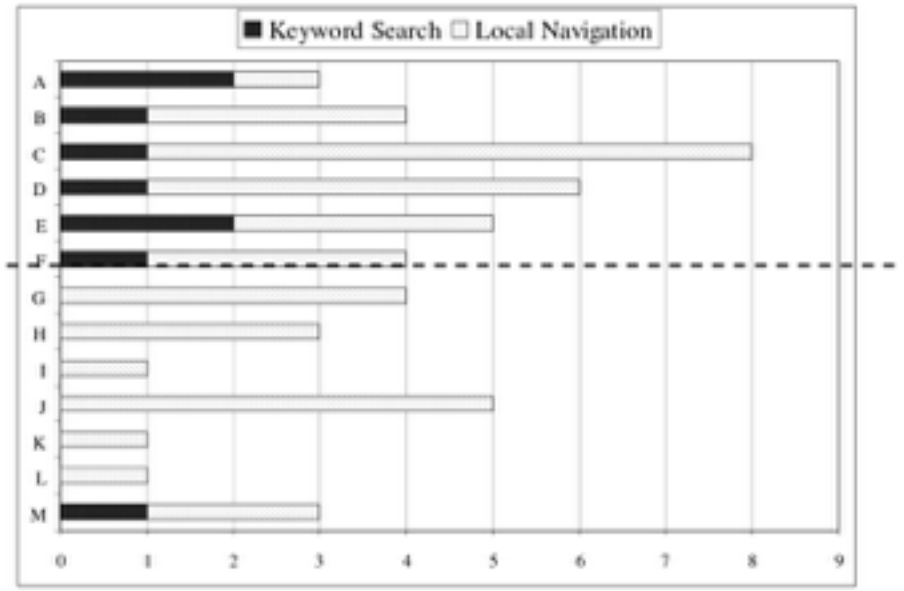


Figure 2: The number of times participants used each search tactic in their files.

Pilers ←→ Filers

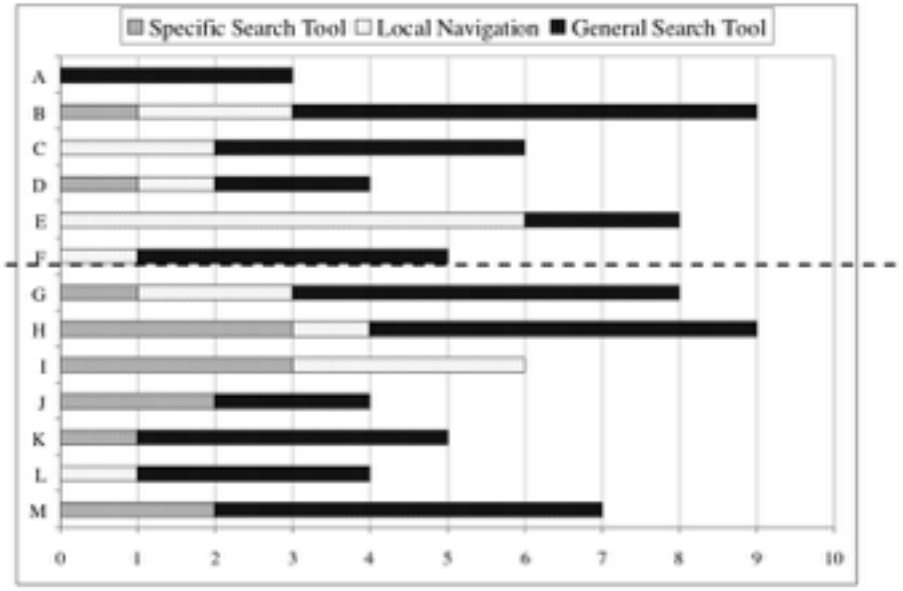
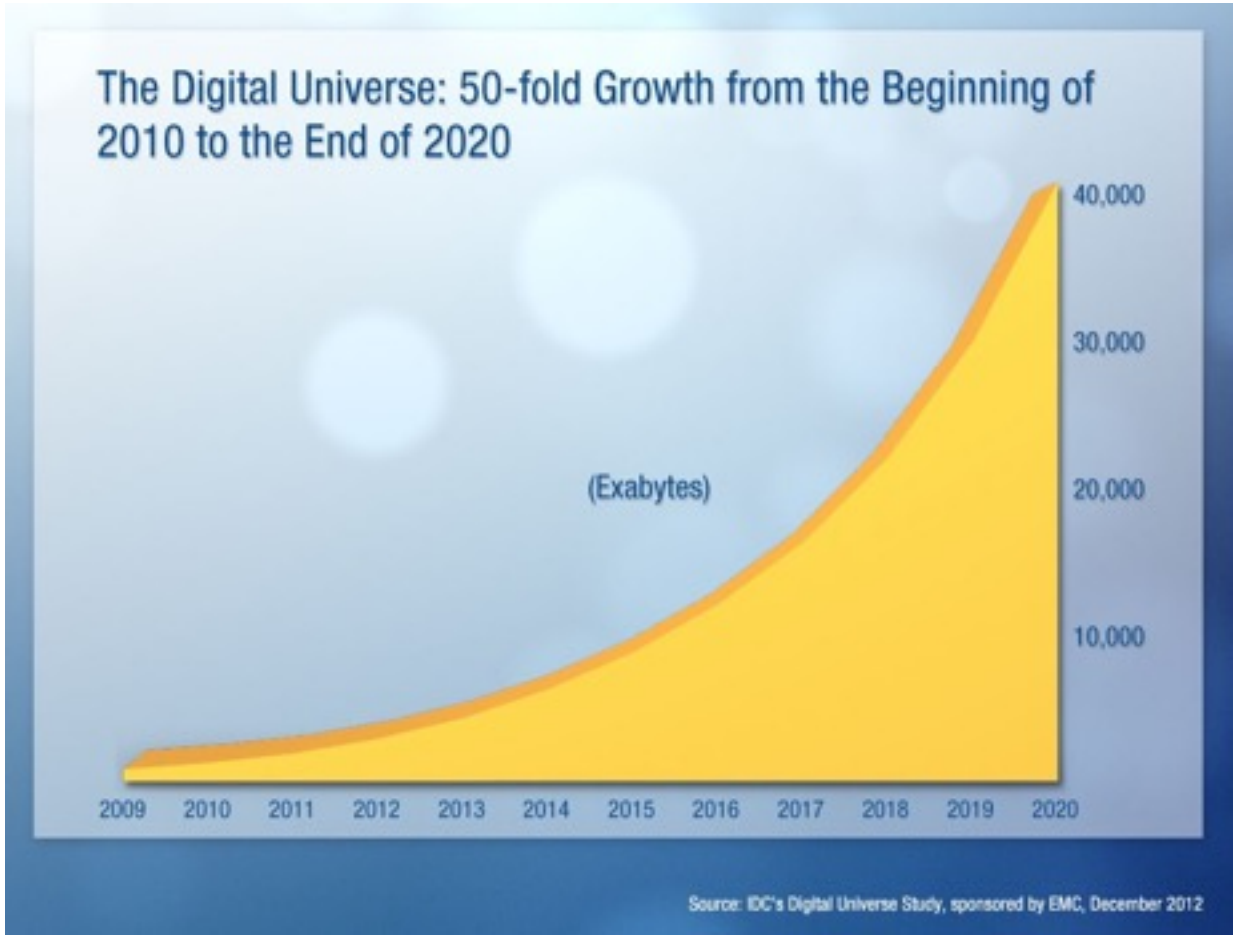


Figure 3: The number of times participants used each tactic on the Web. Note that pilers appear to use specific search tools more often.

Malone, T. E. (1983). How do people organize their desks? ACM Transactions on Office Information Systems 1, 1, 99-112.
Alvarado, Christine, et al. "Surviving the information explosion: How people find their electronic information." (2003).

Data deluge

- Your research
  - More data to sort through



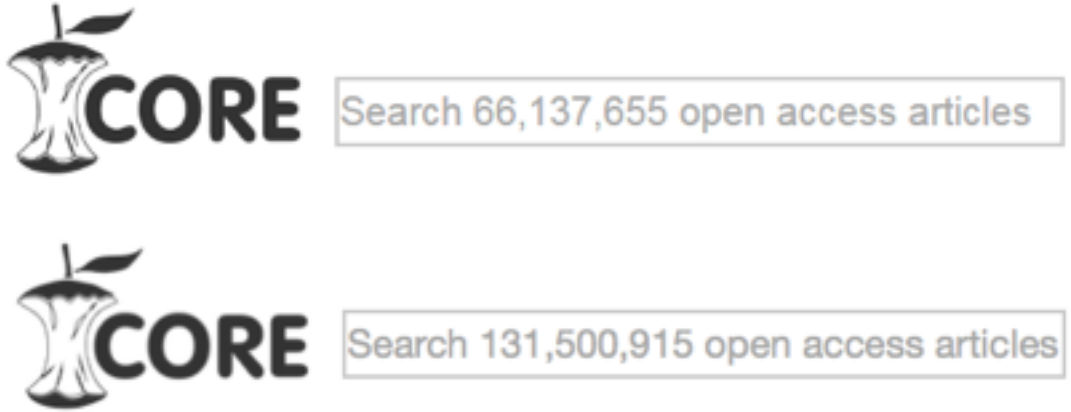
"It is **doubling in size every two years**, and by 2020 the digital universe – the data we create and copy annually – will reach 44 zettabytes, or 44 trillion gigabytes."

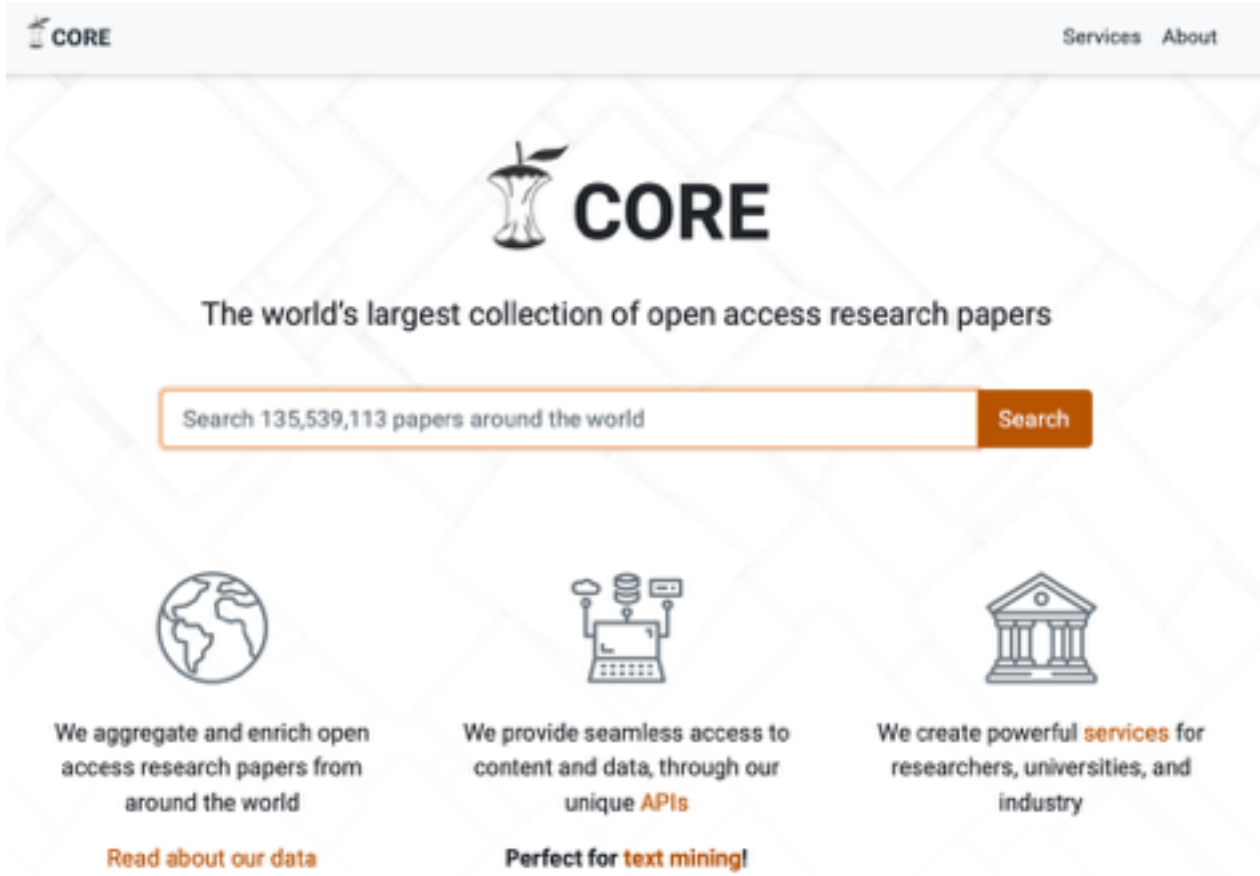
– EMC Digital Universe, Executive Summary, 2014

https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Data deluge

- Your research
  - More data to sort through
  - More data to use in your research





https://core.ac.uk *Accessed: 2017-01 and 2018-05 and 2019-0*

Data deluge

- Your research
  - More data to sort through
  - More data to use in your research
  - More data to manage

- billions of individual assets (pages, images, videos, pdfs etc.)
- As of 2017, collected approximately 500TB of data.
- Increasing by over roughly 60 – 70 TB a year.

https://webarchive.org.uk/en/ukwa/info/faq

Data deluge

- Your research
  - More data to sort through
  - More data to use in your research
  - More data to manage (ML, AI and Informatics)

https://www.quora.com/Is-there-really-a-boom-in-big-data

Data deluge

- Your research
  - More data to sort through
  - More data to use in your research
  - More data to manage
- DIKW Pyramid

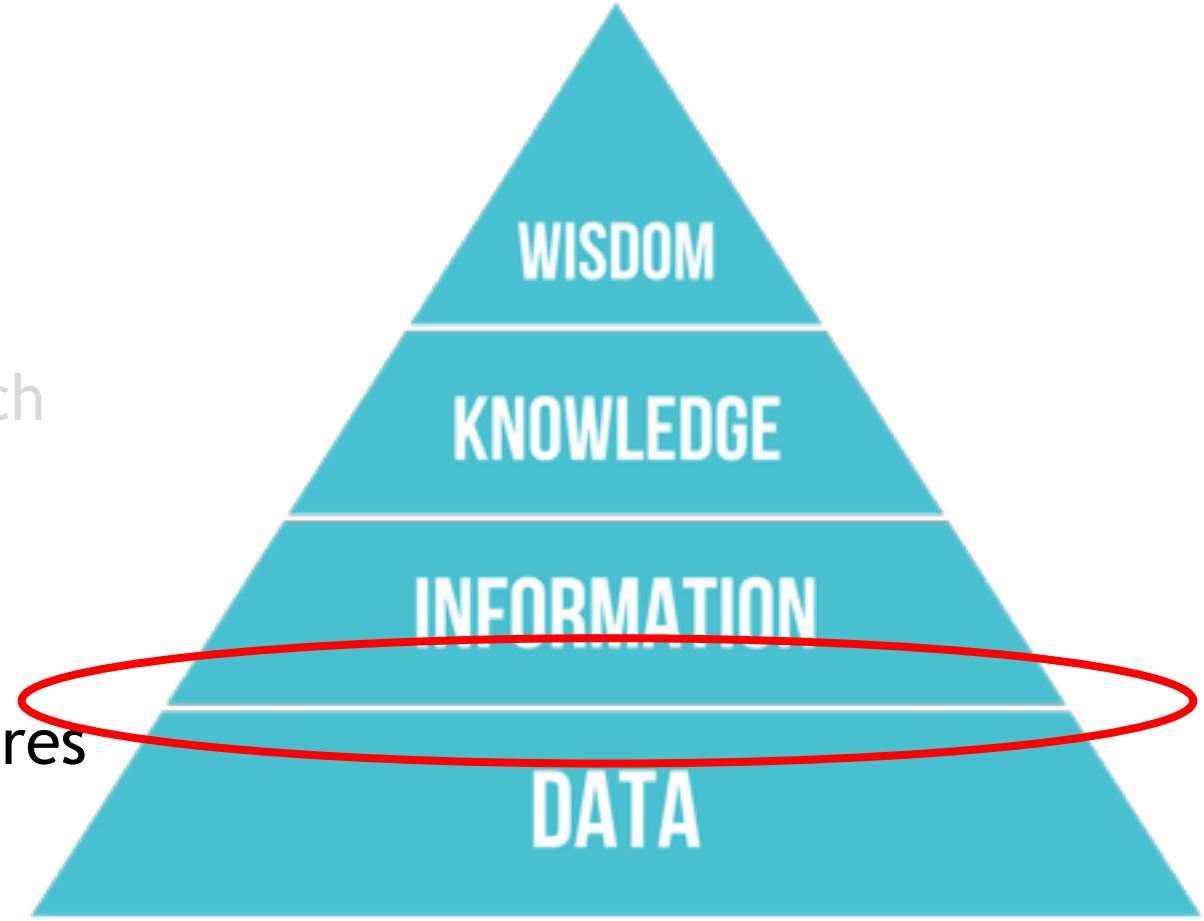
Data deluge

- Your research
  - More data to sort through
  - More data to use in your research
  - More data to manage
- DIKW Pyramid
  - Used in Information Science

https://en.wikipedia.org/wiki/DIKW_Pyramid

Data deluge

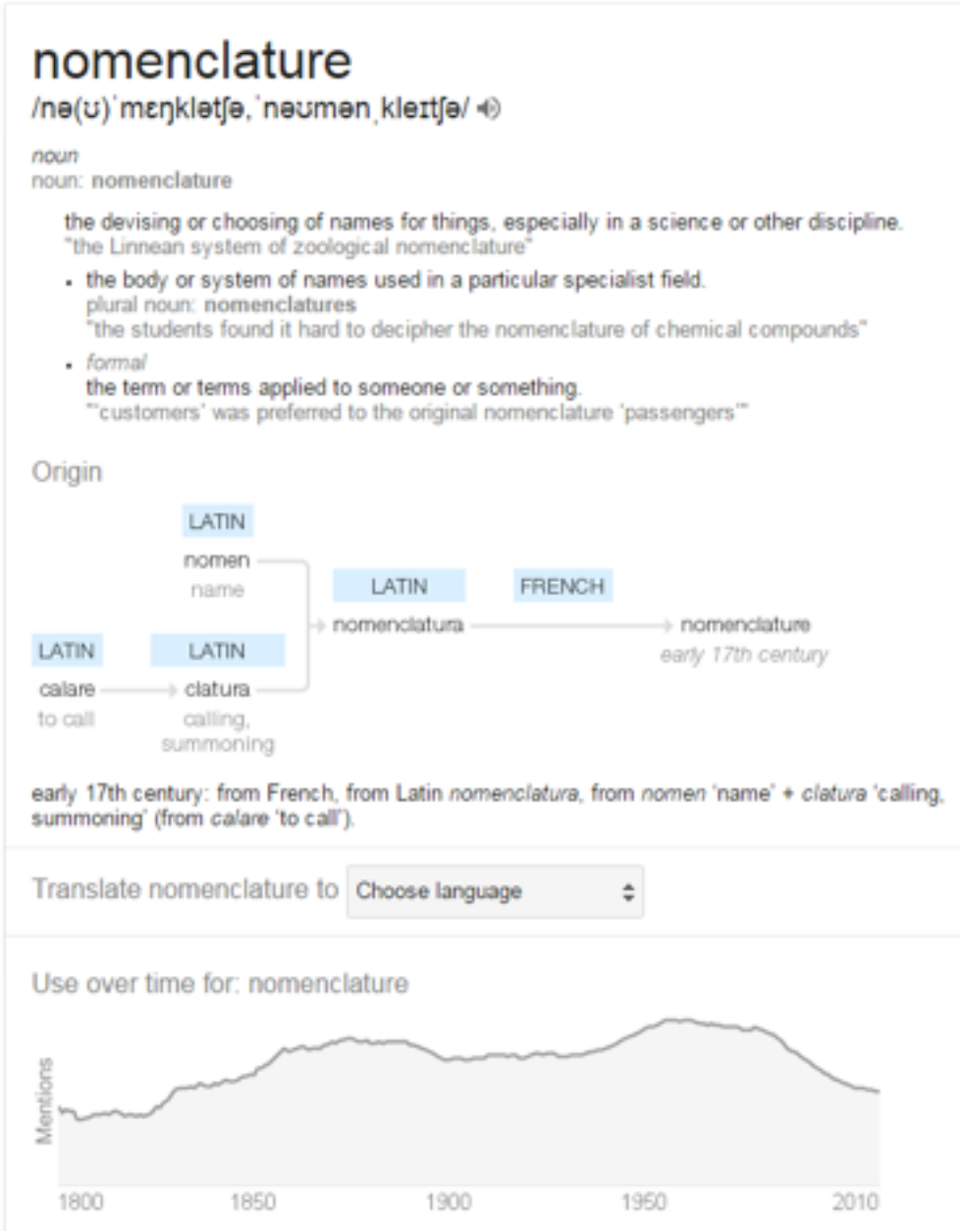
- Your research
  - More data to sort through
  - More data to use in your research
  - More data to manage
- DIKW Pyramid
  - Used in Information Science
  - Naming conventions and structures



https://en.wikipedia.org/wiki/DIKW_Pyramid

Data deluge

- Your research
  - More data to sort through
  - More data to use in your research
  - More data to manage
- DIKW Pyramid
  - Used in Information Science
  - Naming conventions and structure
  - Nomenclature



Google search card for 'Nomenclature'

Information Architecture (IA)

- Richard Saul Wurman, Information architect and graphic designer

That's why I've chosen to call myself an Information Architect … I mean architect as in the creating of systemic, structural, and orderly principles to make something work … I use the word information in its truest sense. Most of the word information contains the word inform, so I call things information only if they inform me, not if they are just collections of data, of stuff.

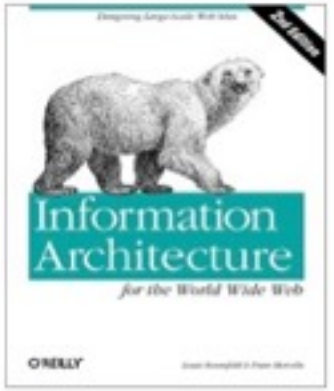
\- *Richard Saul Wurman, 1996*

Resmini, A. & Rosati, L. (2012). A Brief History of Information Architecture. Journal of Information Architecture. Vol. 3, No. 2

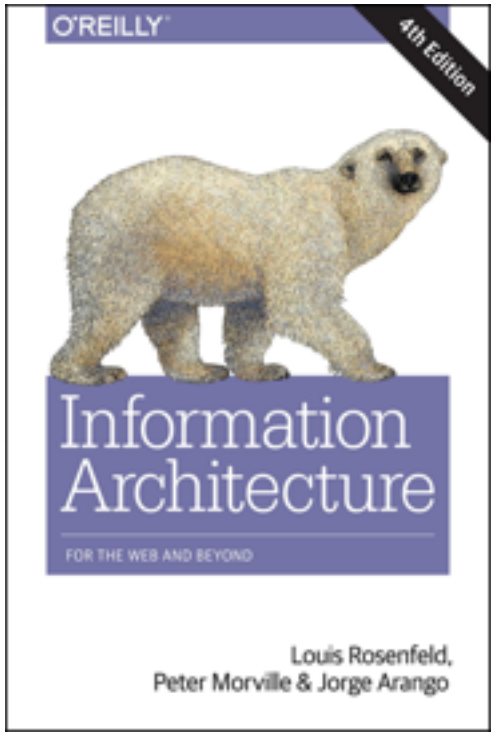
Information Architecture (IA)

- IA is usually associated with:
  - Website development
  - Taxonomy design

1998



2002

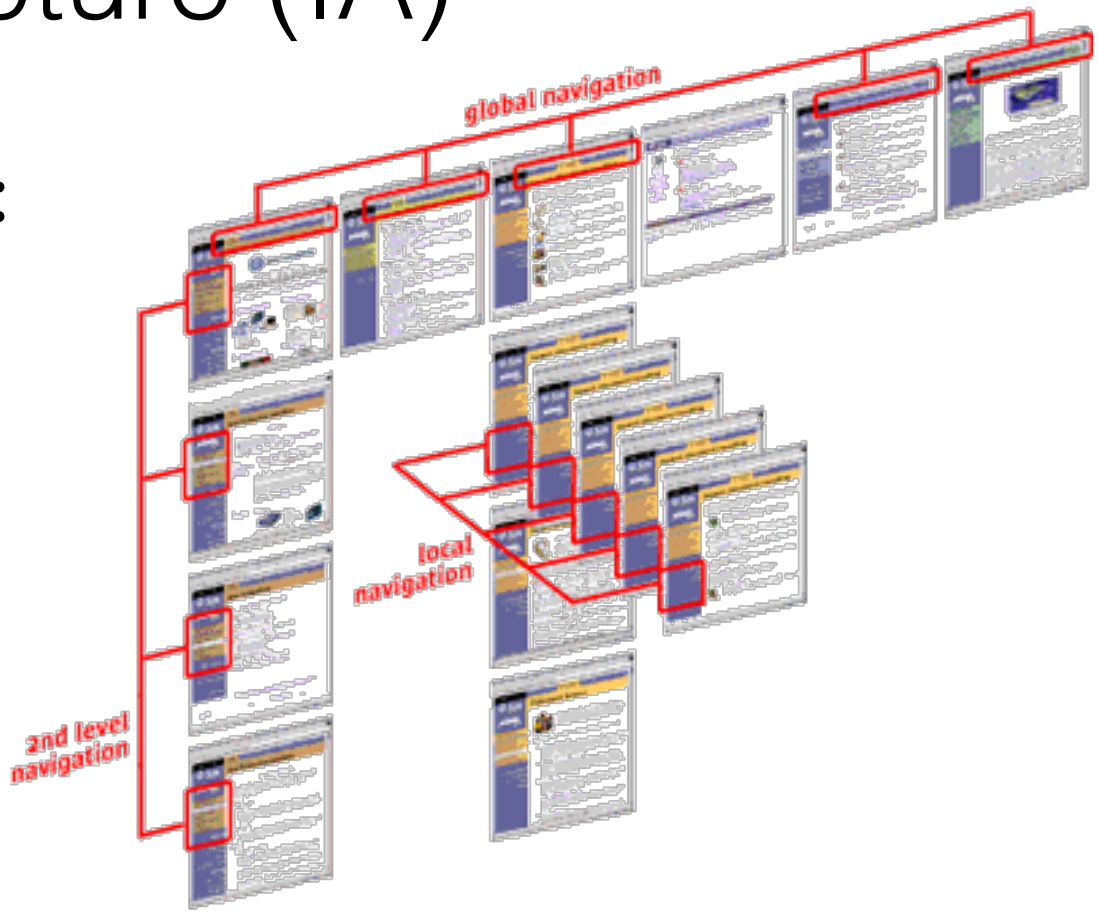


The polar bear book

2018

Information Architecture (IA)

- IA is usually associated with:
  - Website development
  - Taxonomy design
  - Navigation menus
  - Web page hierarchies



Information Architecture (IA)

- IA is:
  - The structural design of shared information environments

Information Architecture (IA)

- IA is:
  - The structural design of shared information environments
  - Naming things
  - Using structures
  - Making it usable

Information Architecture (IA)

- IA is:
  - The structural design of shared information environments
  - Naming things
  - Using structures
  - Making it usable
- Text corpora/Data packages!

Creating Data Packages in R

Open Knowledge Greece was one of 2017's Frictionless Data Tool Fund grantees tasked with extending implementa
core Frictionless Data libraries in R programming language. You can read more about this in their grantee profile. In t
tutorial, Kleanthis Koupidis, a Data Scientist and Statistician at Open Knowledge Greece, explains how to create Data
Packages in R.

This tutorial will show you how to install the R library for working with Data Packages and Table Schema, load a CSV file, infe
and write a Tabular Data Package.

https://frictionlessdata.io/docs/creating-tabular-data-packages-in-r/

Information Architecture (IA)

- IA is:
  - The structural design of shared information e
  - Naming things
  - Using structures
  - Making it usable
- Text corpora/Data packages
- Tidy text in R

The tidy text format

```
library(tidytext)

text_df %>%
  unnest_tokens(word, text)
```

```
## # A tibble: 20 x 2
##     line word
##    <int> <chr>
##  1     1 because
##  2     1 i
##  3     1 could
##  4     1 not
##  5     1 stop
##  6     1 for
##  7     1 death
##  8     2 he
##  9     2 kindly
## 10     2 stopped
## # … with 10 more rows
```

https://www.tidytextmining.com/tidytext.html

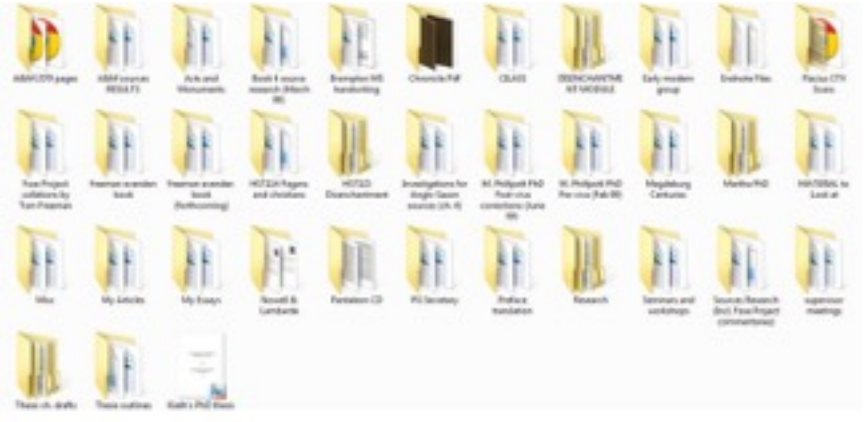
Naming things

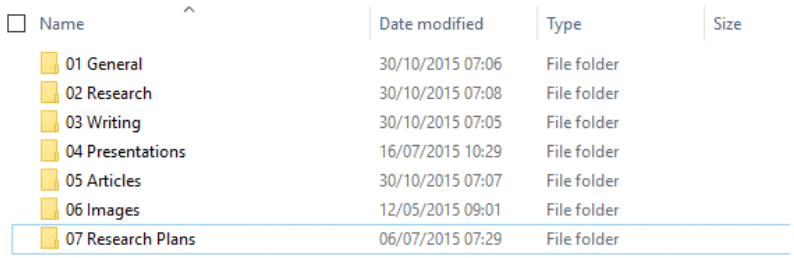
- Importance of naming things
  - Long lived

    Usually hard to change

Naming things

- Importance of naming things
  - Long lived
  - Reduces complexity





Naming things

- Importance of naming things
  - Long lived
  - Reduces complexity
  - Describes the object

```
library(dplyr)
library(tidytext)

ap_td <- tidy(AssociatedPress)
ap_td
```

```
## # A tibble: 302,031 x 3
##    document term       count
##       <int> <chr>      <dbl>
##  1        1 adding          1
##  2        1 adult           2
##  3        1 ago             1
##  4        1 alcohol         1
##  5        1 allegedly       1
##  6        1 allen           1
##  7        1 apparently      2
##  8        1 appeared        1
##  9        1 arrested        1
## 10        1 assault         1
## # … with 302,021 more rows
```

Tidy a list of terms up and turn into data frame

Naming things

- Importance of naming things
  - Long lived
  - Reduces complexity
  - Describes the object
  - Informs about structure

http://www.british-history.ac.uk/source.aspx?pubid=739

Naming things

- Importance of naming things
  - Long lived
  - Reduces complexity
  - Describes the object
  - Informs about structure

    Series, Volume, Pages

Providing structures

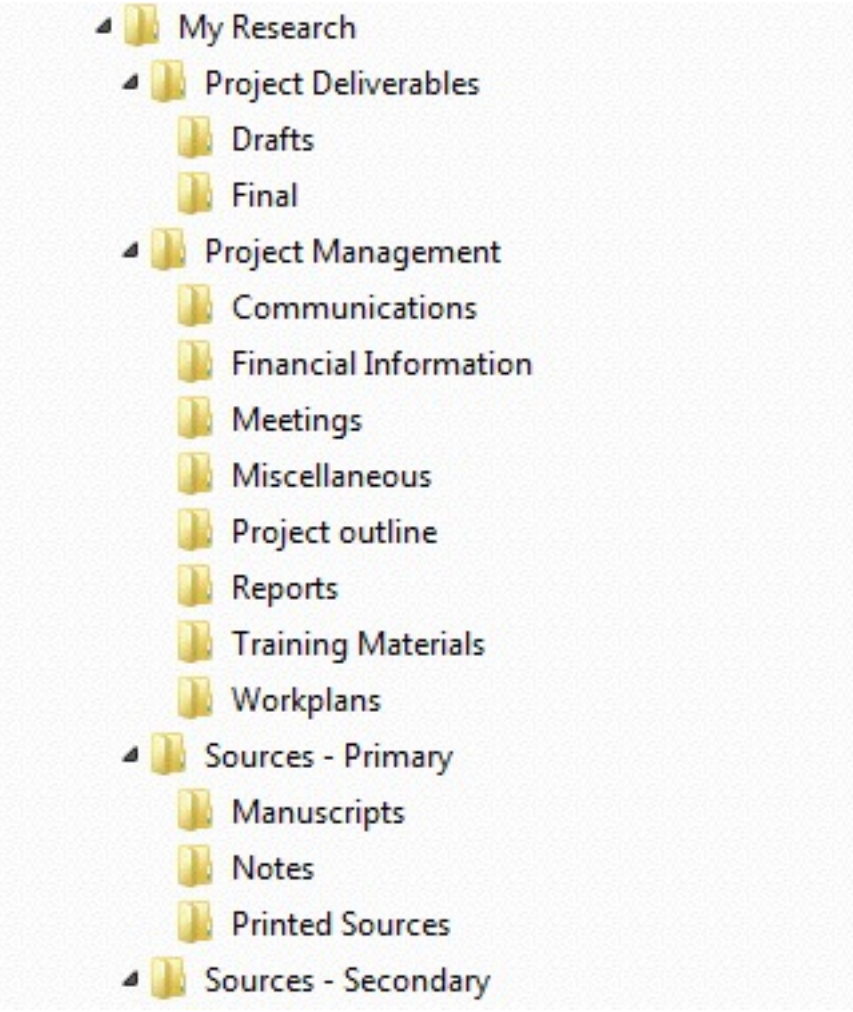
- Hierarchies
- Collations
- Conventions

Using structures

- Hierarchies
  - Simple form of classification

Using structures

- Hierarchies
  - Simple form of classification
  - Subject/Function
    - Type/categories next level down



Using structures

- Hierarchies
  - Simple form of classification
  - Subject/Function
    - Type/categories
  - Your own vocabulary

Browse by Department

Please select a value to browse from the list below.

- (346)
- (54)
- ? (2)
- CN (1)
- Colonial Office (4)
- FR (1)
- Foreign Office (3)
- Government of New Zealand (1)
- Helbert, Wagg & Company, Limited (1)
- IN (5)
- NZ (1)
- Royal Navy UK (5)
- Trinidad (4)
- UK (631)
- UK Embassy, Washington, D.C. (81)
- UK Embassy, Washington, D.C.; UK Embassy, Washington, D.C. (2)
- UK Embassy, Washington, D.C.; UK Embassy, Washington, D.C.; UK Embassy, Washington, D.C.
- UK Embassy, Washington, D.C.; US Department of State (1)
- UK Embassy, Washington, D.C.; Unknown (1)
- UK Embassy, Washington, D.C.; Unknown; Unknown (3)
- US (81)
- US Department of State (20)
- US Department of the Interior (1)
- US Executive Office (1)
- US Executive Office; California State government; San Francisco city government (1)
- US Executive Office; US Executive Office (1)
- US House of Representatives (1)
- Unknown (4)
- us (2)

Using structures

- Hierarchies
  - Simple form of classificatio
  - Subject/Function
    - Type/categories
  - Your own vocabulary
  - Controlled Vocabulary

Using structures

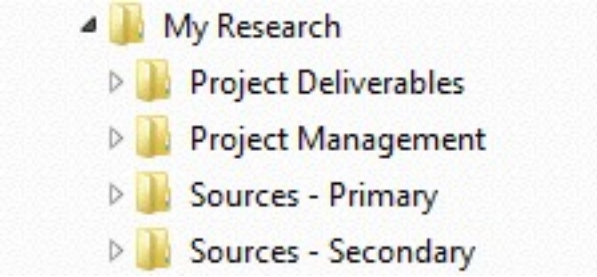
- Hierarchies
- Collations

Using structures

- Hierarchies
- Collations
  - Putting things into order

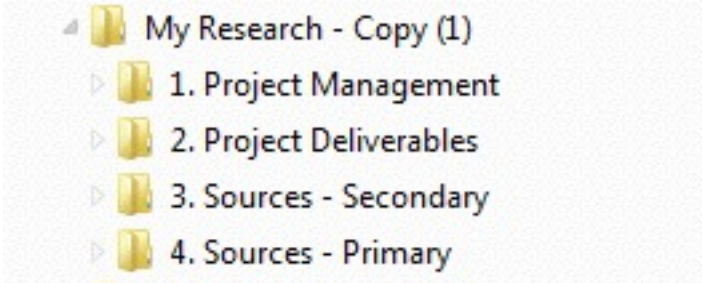
Using structures

- Hierarchies
- Collations
  - Putting things into order
  - Alphabetical
    - Deliver first. Manage later



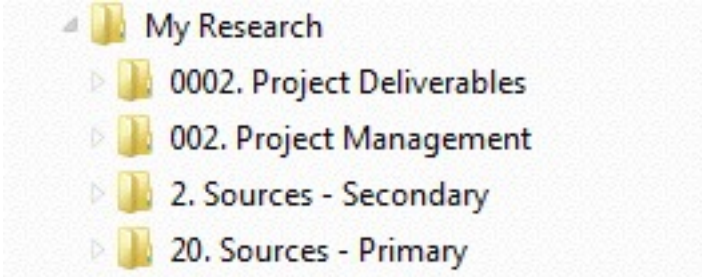
Using structures

- Hierarchies
- Collations
  - Putting things into order
  - Alphabetical
  - Numerical
    - Force the order we want



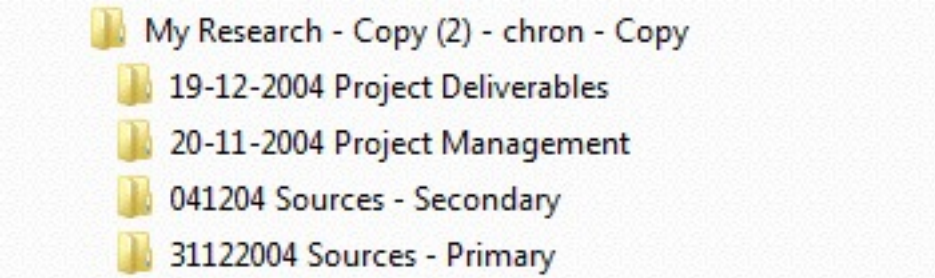
Using structures

- Hierarchies
- Collations
  - Putting things into order
  - Alphabetical
  - Numerical
    - Force the order we want
    - Zero prefix
    - Affects order differently on Mac/PC



Using structures

- Hierarchies
- Collations
  - Putting things into order
  - Alphabetical
  - Numerical
  - Chronological
    - At the end of the name
    - At the start of the name
    - Date formats matter
      - DD-MM-YYYY, MMDDYY, DDMMYYYY, etc.



Using structures

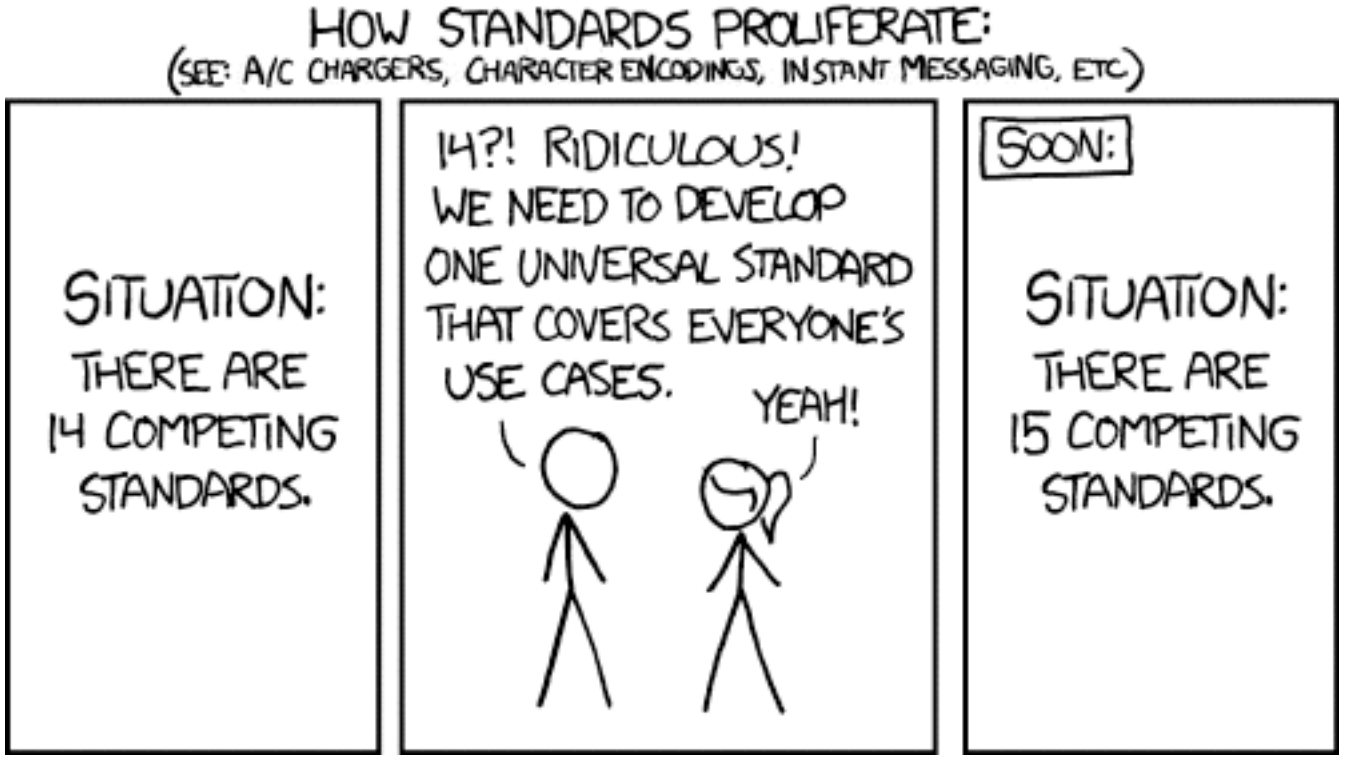
- Hierarchies
- Collations
  - Putting things into order
  - Alphabetical
  - Numerical
  - Chronological

A tidy text dataframe

```
## # A tibble: 50 x 16
##    author  datetimestamp       description heading  id    language origin topics l
##    <chr>   <dttm>              <chr>       <chr>    <chr> <chr>    <chr>  <chr>  <
##  1 <NA>    1987-02-26 15:18:06 ""          COMPUTE… 10    en       Reute… YES    T
##  2 <NA>    1987-02-26 15:19:15 ""          OHIO MA… 12    en       Reute… YES    T
##  3 <NA>    1987-02-26 15:49:56 ""          MCLEAN'… 44    en       Reute… YES    T
##  4 By Cal… 1987-02-26 15:51:17 ""          CHEMLAW… 45    en       Reute… YES    T
##  5 <NA>    1987-02-26 16:08:33 ""          <COFAB … 68    en       Reute… YES    T
##  6 <NA>    1987-02-26 16:32:37 ""          INVESTM… 96    en       Reute… YES    T
##  7 By Pat… 1987-02-26 16:43:13 ""          AMERICA… 110   en       Reute… YES    T
##  8 <NA>    1987-02-26 16:59:25 ""          HONG KO… 125   en       Reute… YES    T
##  9 <NA>    1987-02-26 17:01:28 ""          LIEBERT… 128   en       Reute… YES    T
## 10 <NA>    1987-02-26 17:08:27 ""          GULF AP… 134   en       Reute… YES    T
```

Using structures

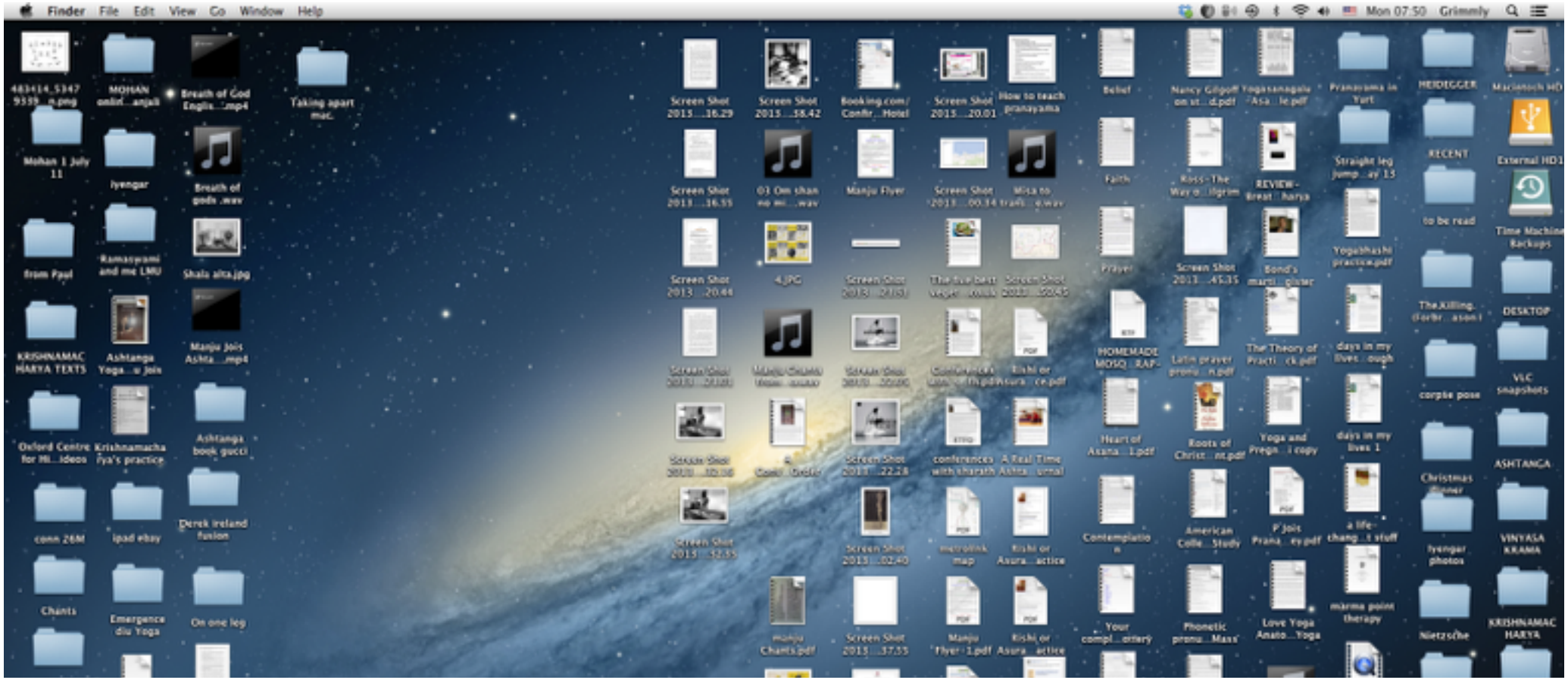
- Hierarchies
- Collations
  - Putting things into order
  - Alphabetical
  - Numerical
  - Chronological
- Consistency is key!



https://xkcd.com/927/

Is it usable?

- Can you find the majority of your files without searching?



Storage, preservation and sustainability

Storage, preservation and sustainability

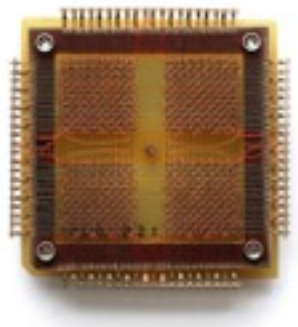
- Backups
- Types of storage
- Documentation
- Version control
- Reproducibility

Storage, preservation and sustainability

- Backups
- Types of storage
- Documentation
- Version control
- Reproducibility

Storage, preservation and sustainability

- Backups
- Types of storage
- Documentation
- Version control
- Reproducibility



OneDrive

SharePoint

Storage, preservation and sustainability

- Backups
- Types of storage
- Documentation
- Version control
- Reproducibility

Storage, preservation and sustainability

- Backups
- Types of storage
- Documentation
- Version control
- Reproducibility

"Obsolete power corrupts obsoletely."
- Ted Nelson

The technology associated with interpreting the representation at each of the layers can change or become less available

Storage, preservation and sustainability

- Backups
- Types of storage
- Documentation
- Version control
- Reproducibility

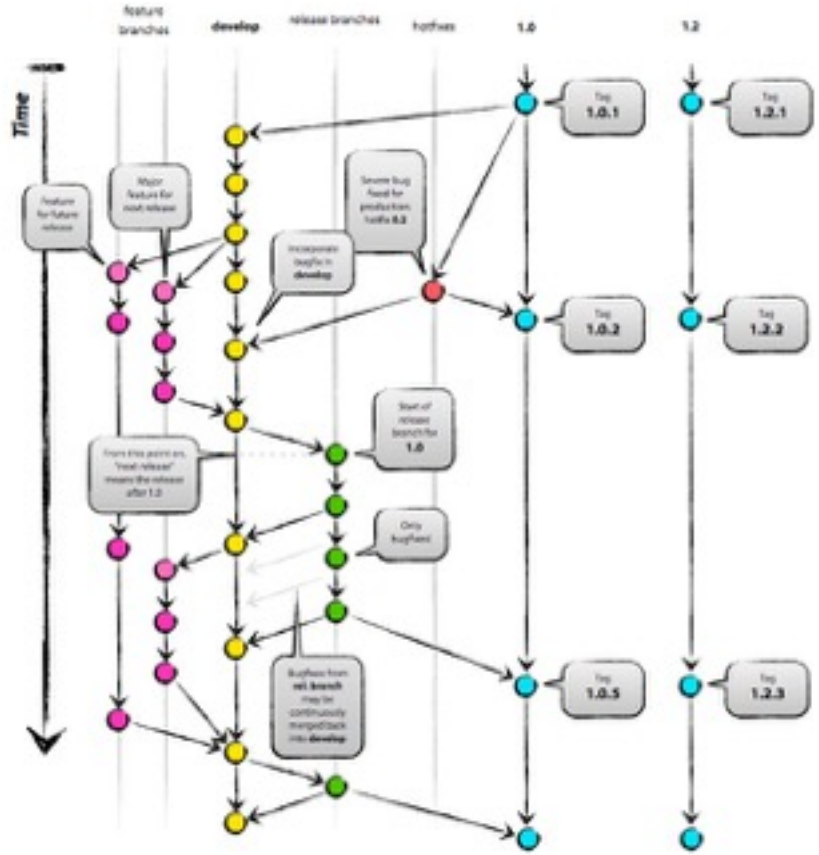
Storage, preservation and sustainability

- Backups
- Types of storage
- Documentation
- Version control
- Reproducibility

Closed
vs.
Open
standards

Storage, preservation and sustainability

- Backups
- Types of storage
- Documentation
- Version control
- Reproducibility



Storage, preservation and sustainability

- Backups
- Types of storage
- Documentation
- Version control
- Reproducibility

RefWorks

Endnote

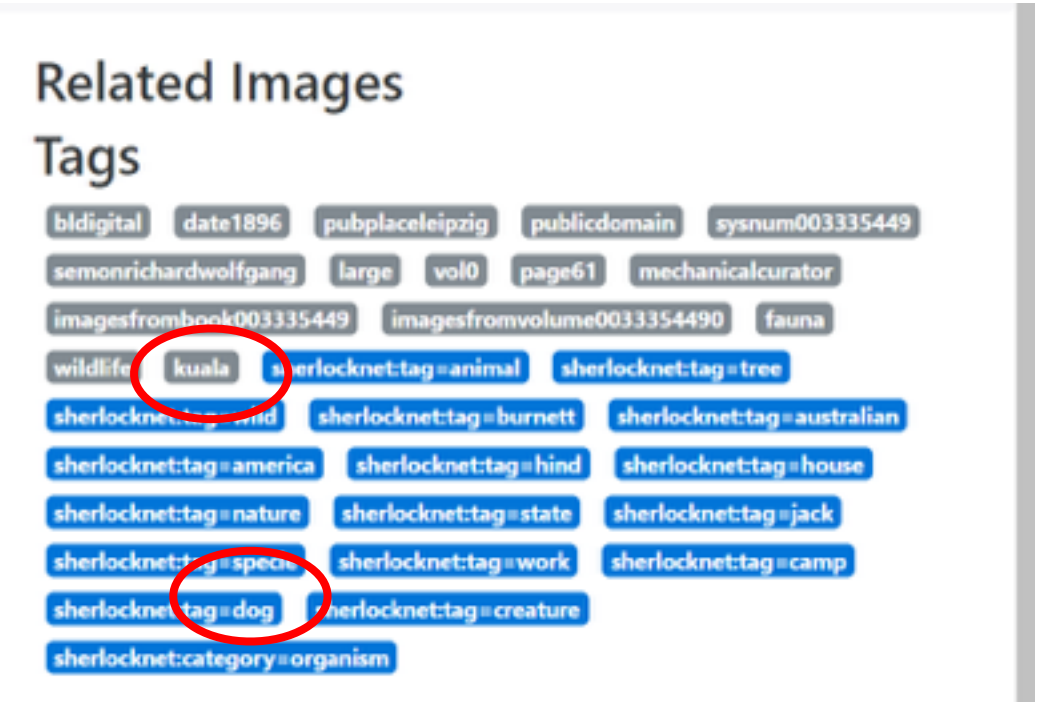
Mendeley

Zotero

EVERNOTE

Storage, preservation and sustainability

- Backups
- Types of storage
- Documentation
- Version control
- Reproducibility



https://enigmatic-tundra-44070.herokuapp.com

Backups

- A hard drive crashes every 15 sec.
- One in 5 computers suffer a fatal hard drive crash during their lifetime.
- 25% of lost data is due to the failure of a portable drive.
- 31% of PC users have lost all of their PC files to events beyond their control.
- 32% of data loss is caused by human error.
- 60% of companies that lose their data close down within 6 months of the disaster.

Also lost!! → https://web.archive.org/web/20160422230428/http://www.kiesoft.com/eb/crashstat.htm

Backups

- Mistakes are guaranteed
- Format impermanence
- Inadequate storage
- Fixit

"I left my USB stick on the train with a variety of personal research projects on it. I had to reconstruct the information trails but the notes from secondary sources were lost and not all of the sources are available in my library meaning I will eventually have to pay out for another train fare and a day's worth of researching lost.

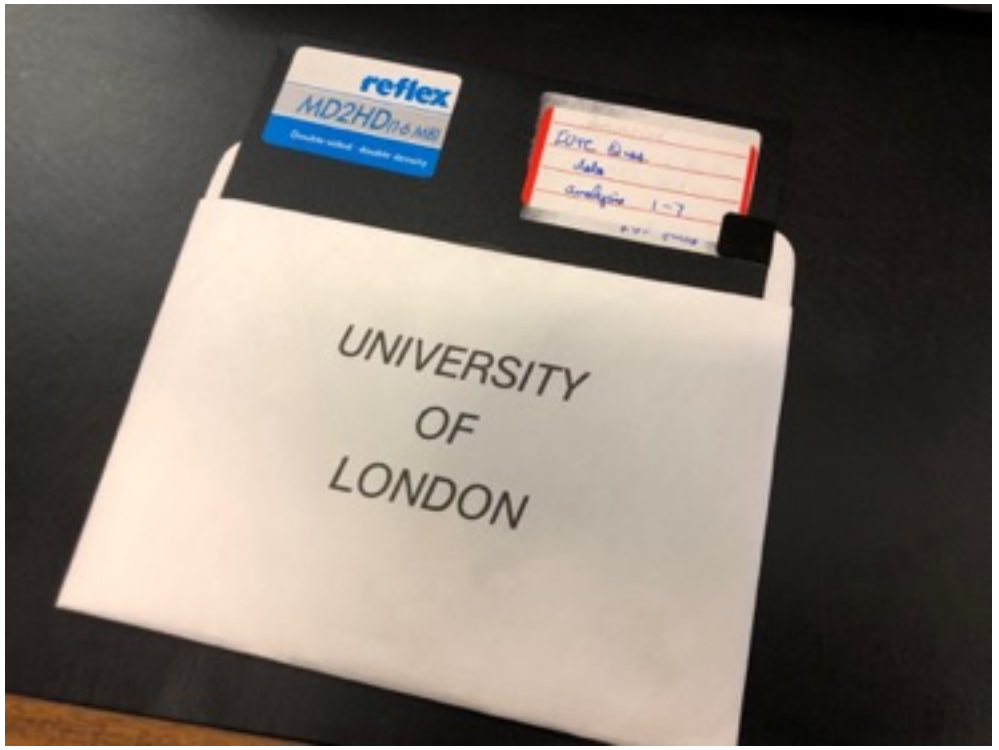
"I took photographs of manuscript pages at The National Archive, but never properly documented what each photograph recorded. I had to go back to the original source to work out which document was which so that I could properly research and reference it.

"I had all my research data on a USB key. While bending over my coffee, I inadvertently dipped the whole USB into the coffee thereby rendering it useless"

- SHARD Project

Backups

- Mistakes are guaranteed
- Format impermanence
- Inadequate storage
- Fixity

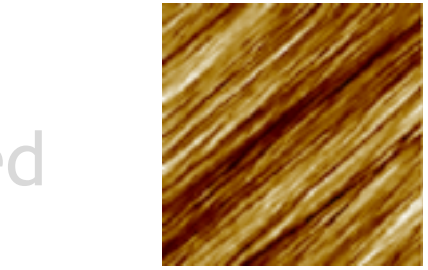


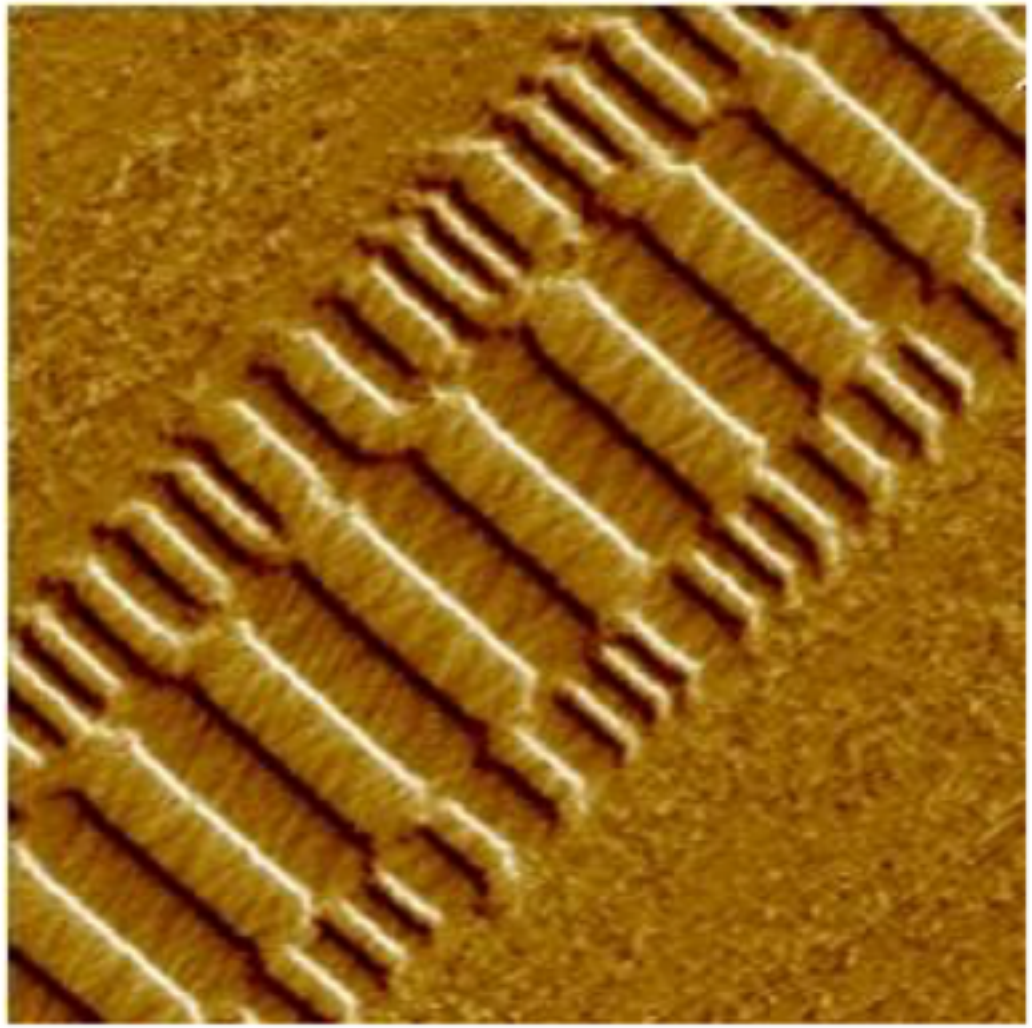
USASCII code chart

$b_7 b_6 b_5$ / Bits					0 0 0	0 0 1	0 1 0	0 1 1
b_4	b_3	b_2	b_1	Row ↓ Column →	0	1	2	3
0	0	0	0	0	NUL	DLE	SP	0
0	0	0	1	1	SOH	DC1	!	1

Backups

- Mistakes are guaranteed
- Format impermanence
- Inadequate storage
- Fixity

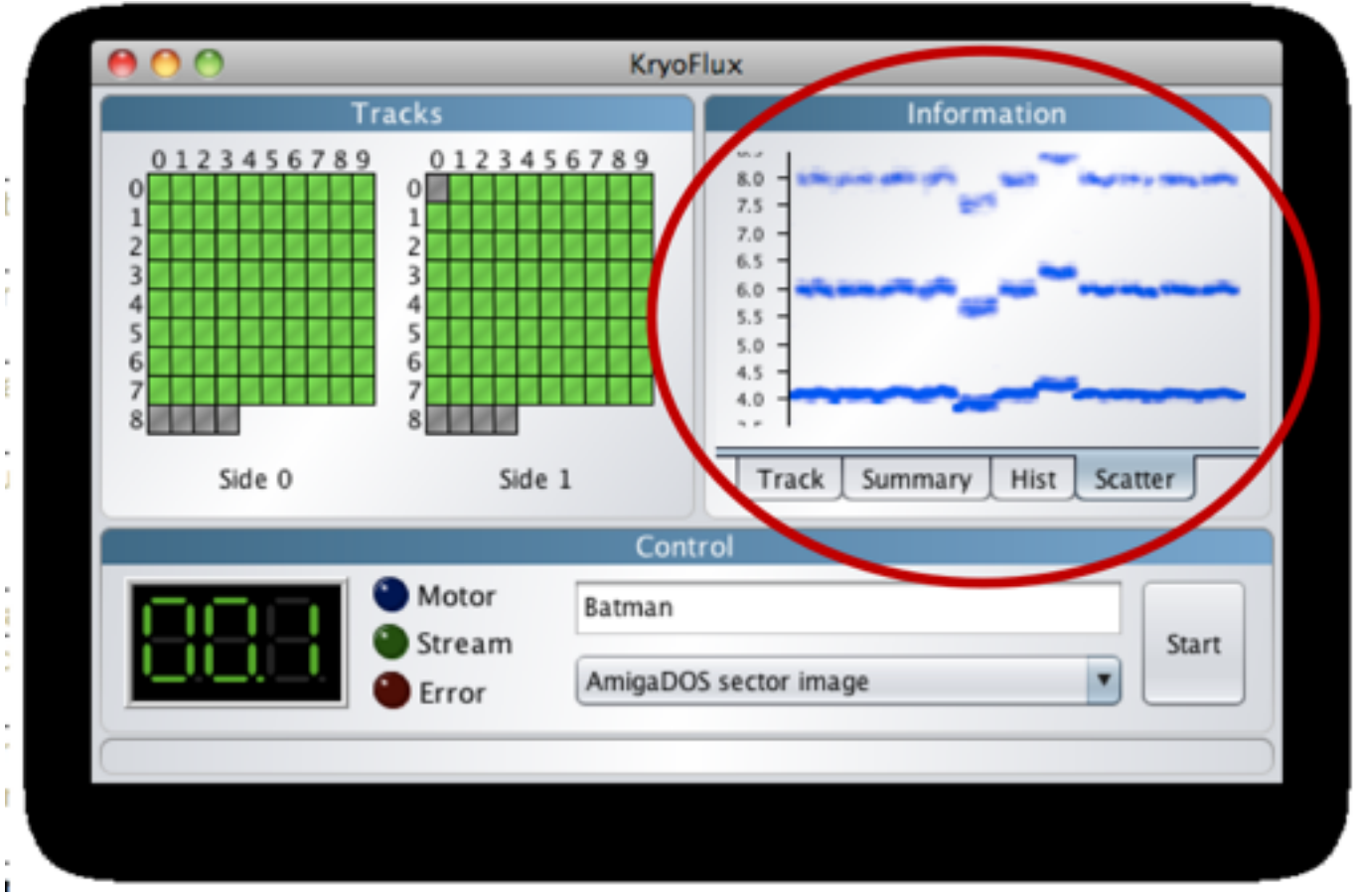




Magnetic Force Microscopy - https://www.engr.sjsu.edu/WofMatE/Mat%27sChar2.htm

Backups

- Mistakes are guaranteed
- Format impermanence
- Inadequate storage
- Fixity/Bit errors



Types of storage

- Internal/external hard drives
- Cloud Storage
- Online repositories
- Network servers
- USB Sticks
- DVDs

“I use Scrivener to organise files and write drafts and notes. The backend of this is a simple folder system on my laptop hard drive that contains all my files. I also use **Zotero** to organise articles and links which unfortunately does not link well with **Scrivener**.

Although I primarily use the hard drive of my laptop to do my research, this automatically syncs to cloud storage (**Onedrive**) whenever I am online. In addition I backup my entire research files on an external hard drive. I have set up a recurring task on my email/calendar system to remind me to do this at the start of each month.”

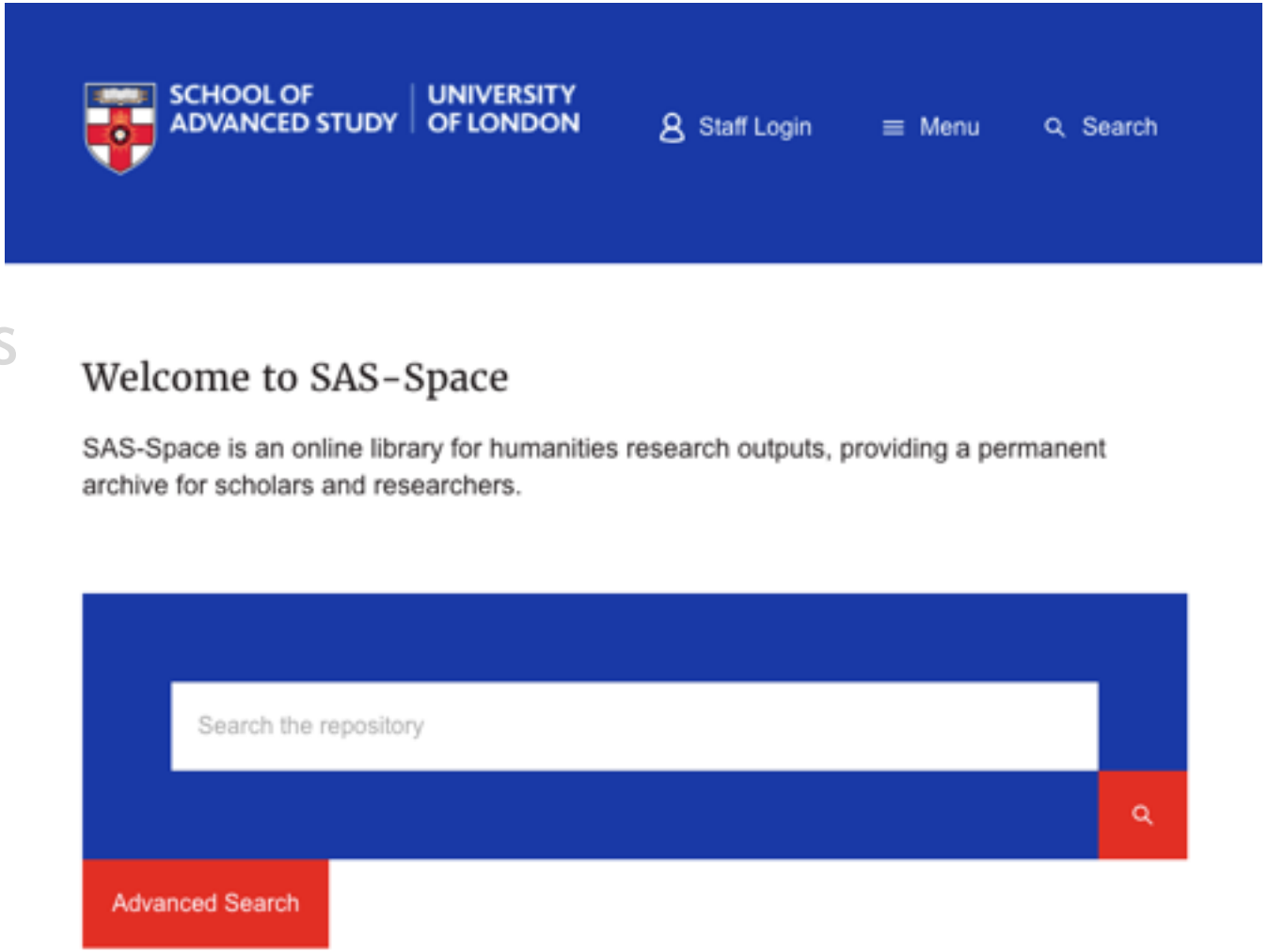
Types of storage

- Internal/external hard drives
- Cloud Storage
- Online repositories
- Network servers
- USB Sticks
- DVDs

Is each type accessible?

Types of storage

- Internal/external hard drives
- Cloud Storage
- Online repositories
- Network servers
- USB Sticks
- DVDs



Types of storage

- Internal/external hard drives
- Cloud Storage
- Online repositories
- Network servers
- USB Sticks
- DVDs

Types of storage

- Internal/external hard driv
- Cloud Storage
- Online repositories
- Network servers
- USB Sticks
- DVDs

Types of storage

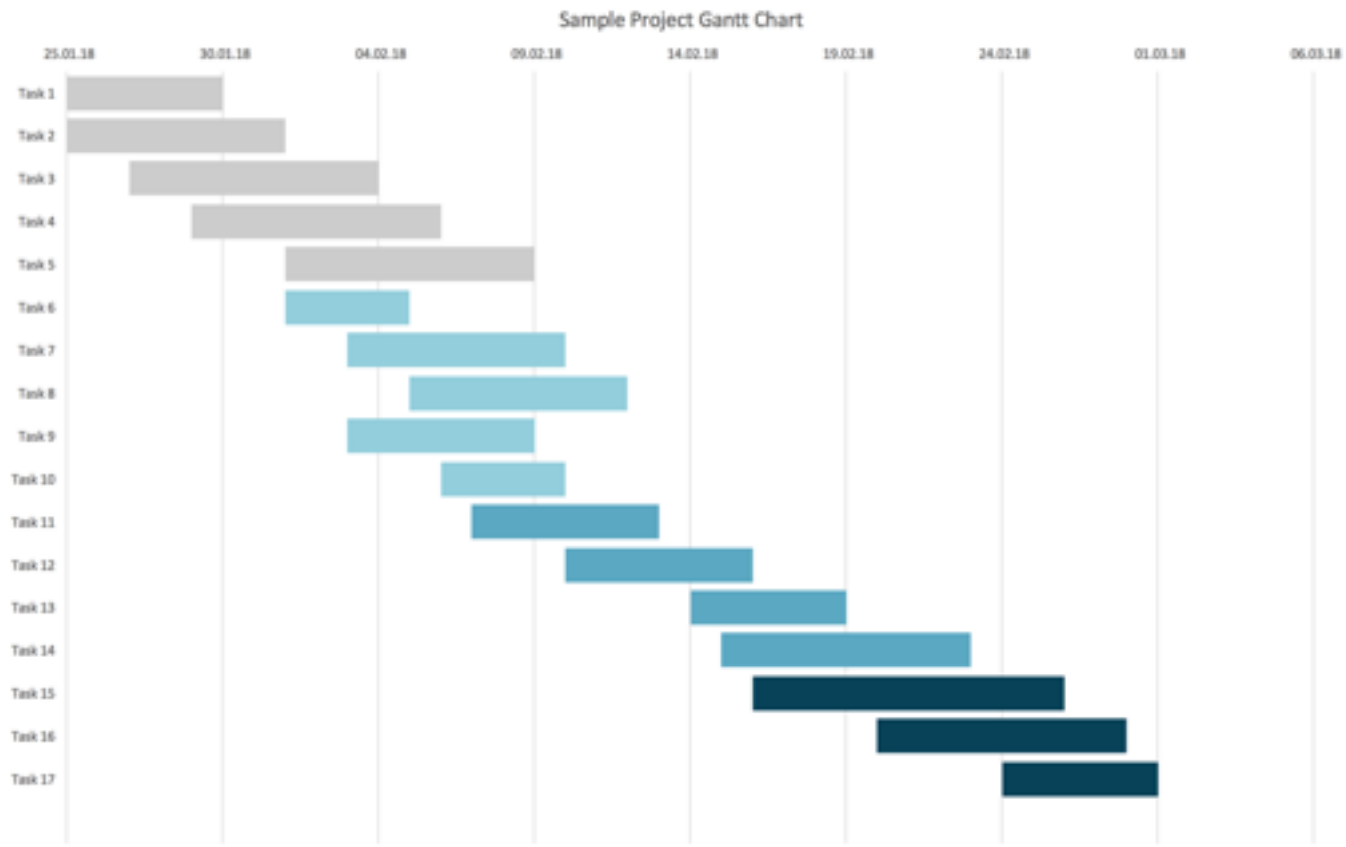
- Internal/external hard drives
- Cloud Storage
- Online repositories
- Network servers
- USB Sticks
- DVDs

Documentation

- Administrative
- Methodological
- Descriptive
- Technical

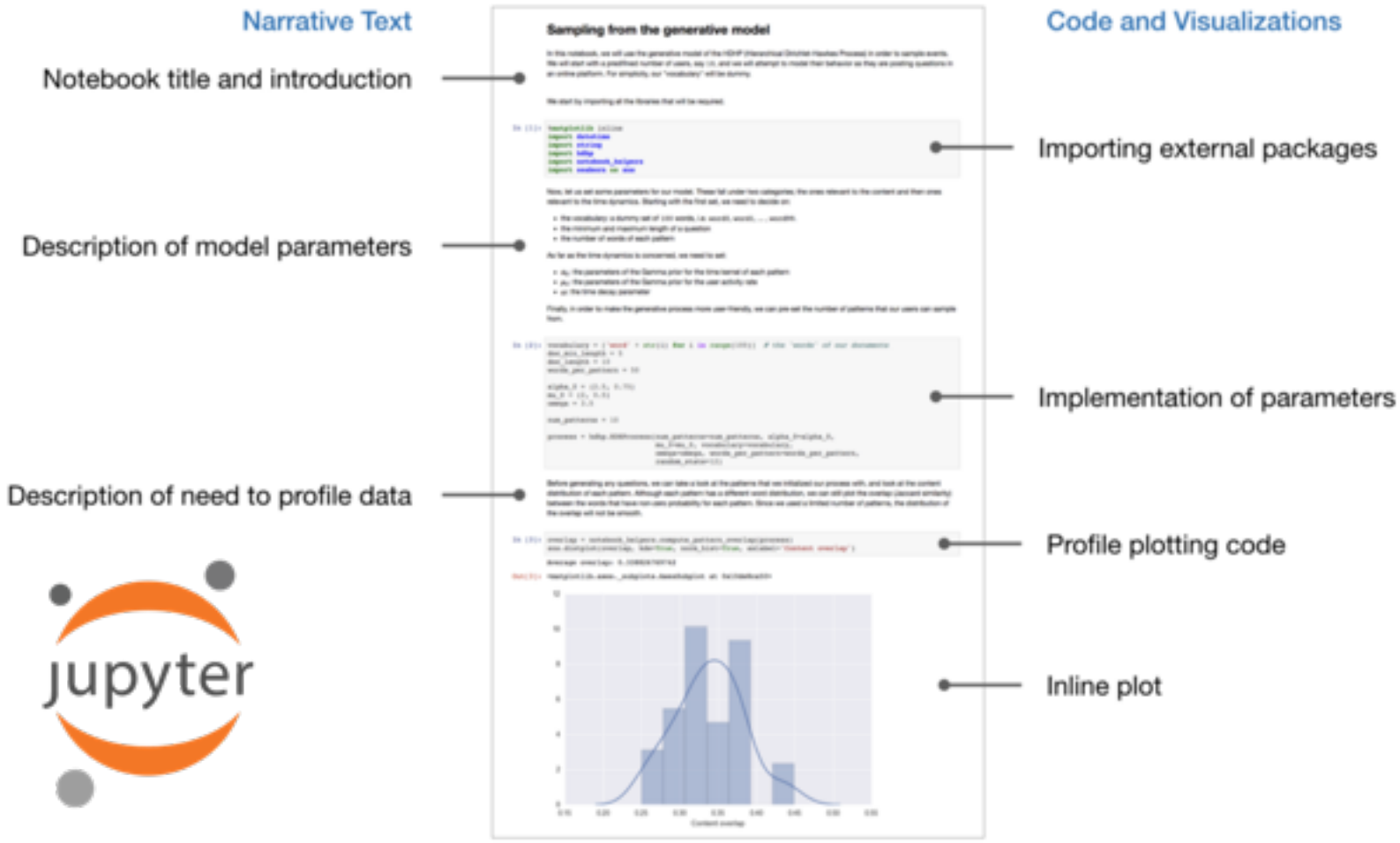
Documentation

- Administrative
- Methodological
- Descriptive
- Technical



Documentatior

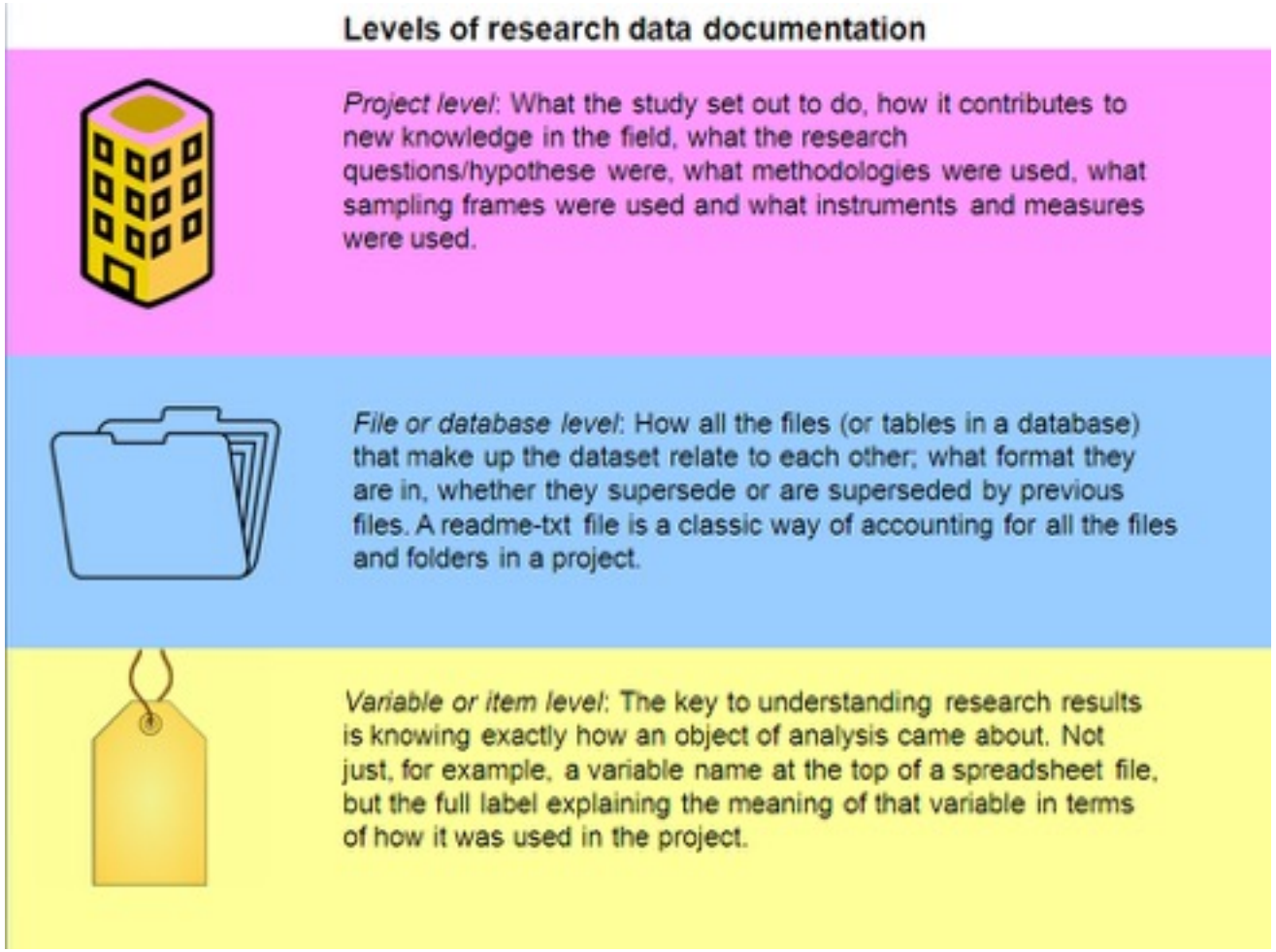
- Administrative
- Methodological
- Descriptive
- Technical



https://jupyter.org

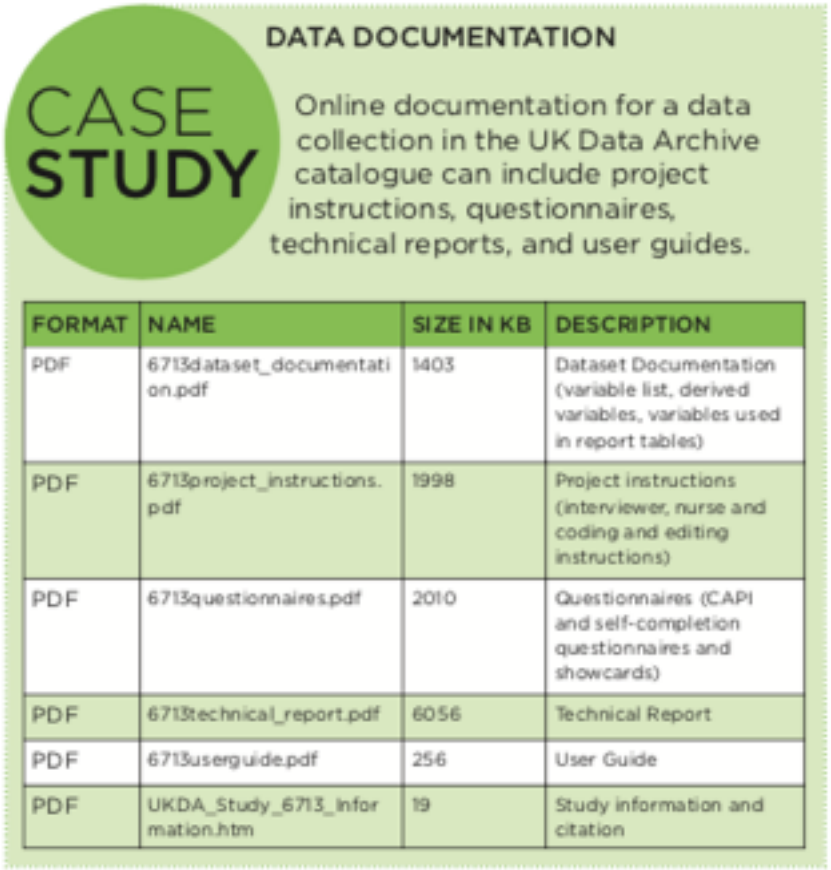
Documentation

- Administrative
- Methodological
- **Descriptive**
- Technical



Documentation

- Administrative
- Methodological
- Descriptive
- Technical



CASE STUDY

DATA DOCUMENTATION

Online documentation for a data collection in the UK Data Archive catalogue can include project instructions, questionnaires, technical reports, and user guides.

FORMAT	NAME	SIZE IN KB	DESCRIPTION
PDF	6713dataset_documentation.pdf	1403	Dataset Documentation (variable list, derived variables, variables used in report tables)
PDF	6713project_instructions.pdf	1998	Project instructions (interviewer, nurse and coding and editing instructions)
PDF	6713questionnaires.pdf	2010	Questionnaires (CAPI and self-completion questionnaires and showcards)
PDF	6713technical_report.pdf	6056	Technical Report
PDF	6713userguide.pdf	256	User Guide
PDF	UKDA_Study_6713_Information.htm	19	Study information and citation

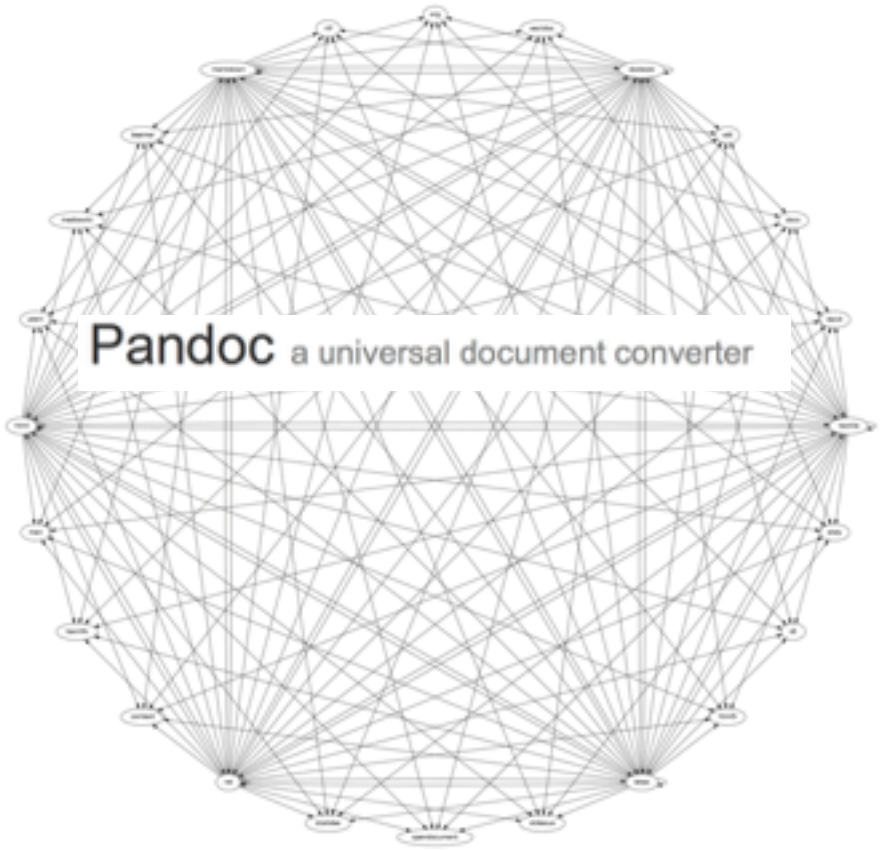
The UK Data Archive, Managing and Sharing Data booklet - https://ukdataservice.ac.uk/media/622417/managingsharing.pdf

Documentation

- Administrative
- Methodological
- Descriptive
- Technical

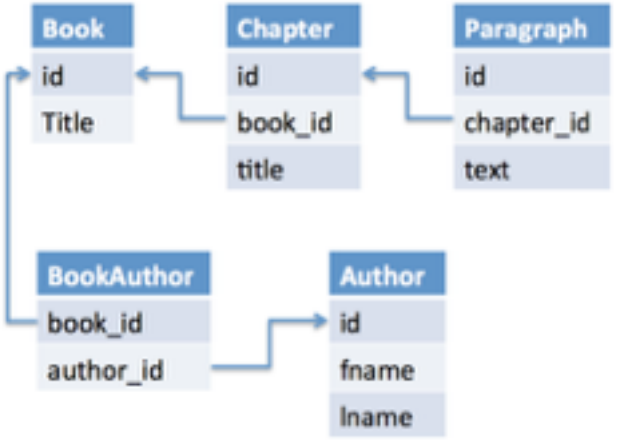
Open or closed Formats?

- XML
- JSON
- TSV
- Word
- Excel
- PDF
- Zip
- Text



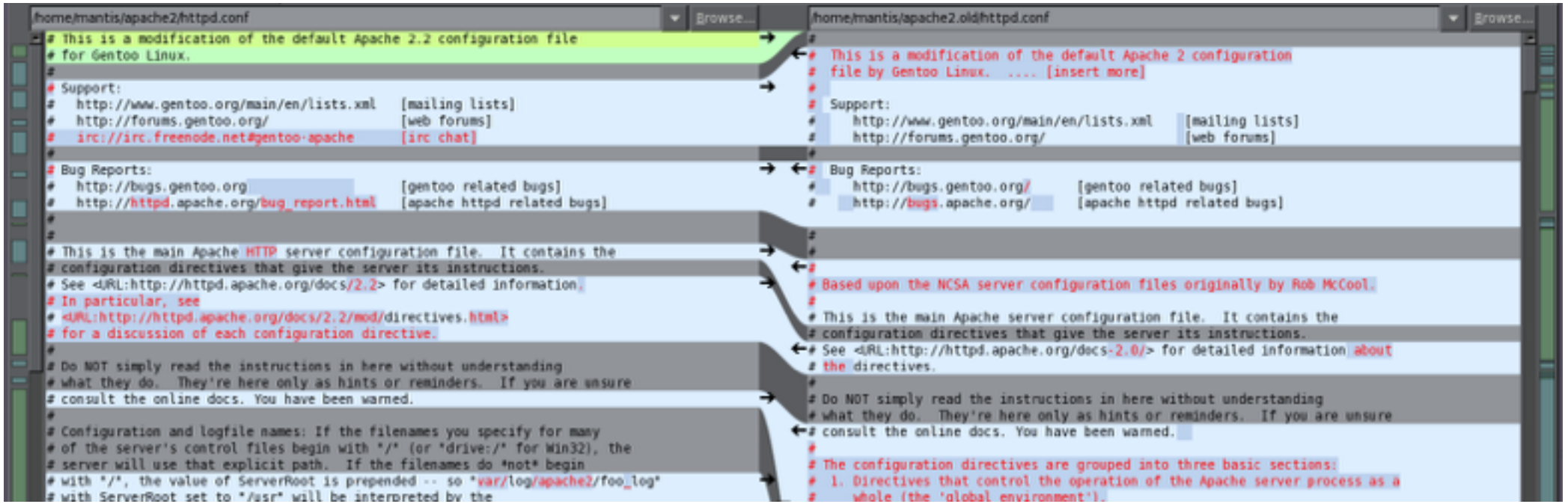
Documentation

- Administrative
- Methodological
- Descriptive
- Technical



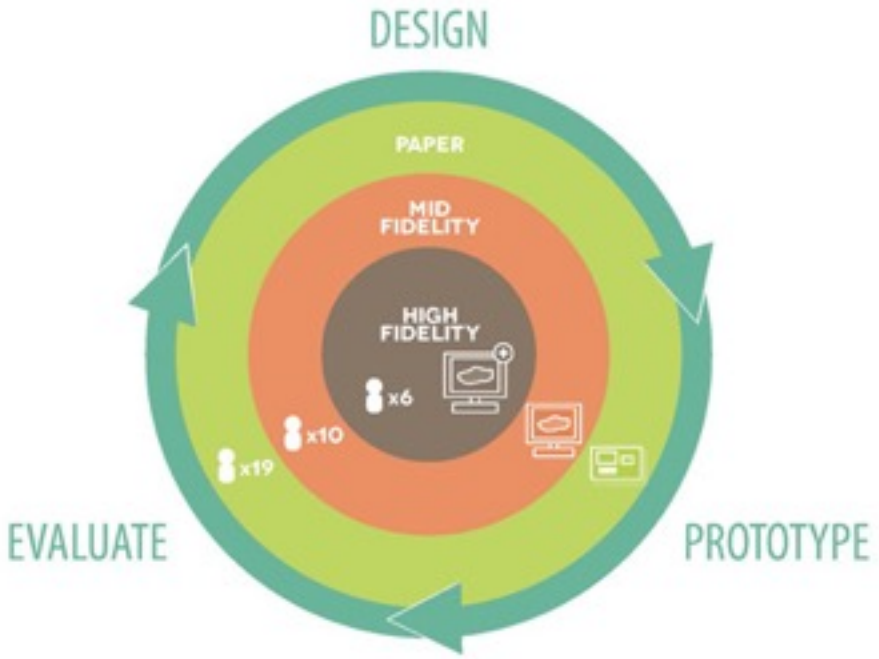
Why do we need version control?

- Record the changes to the document



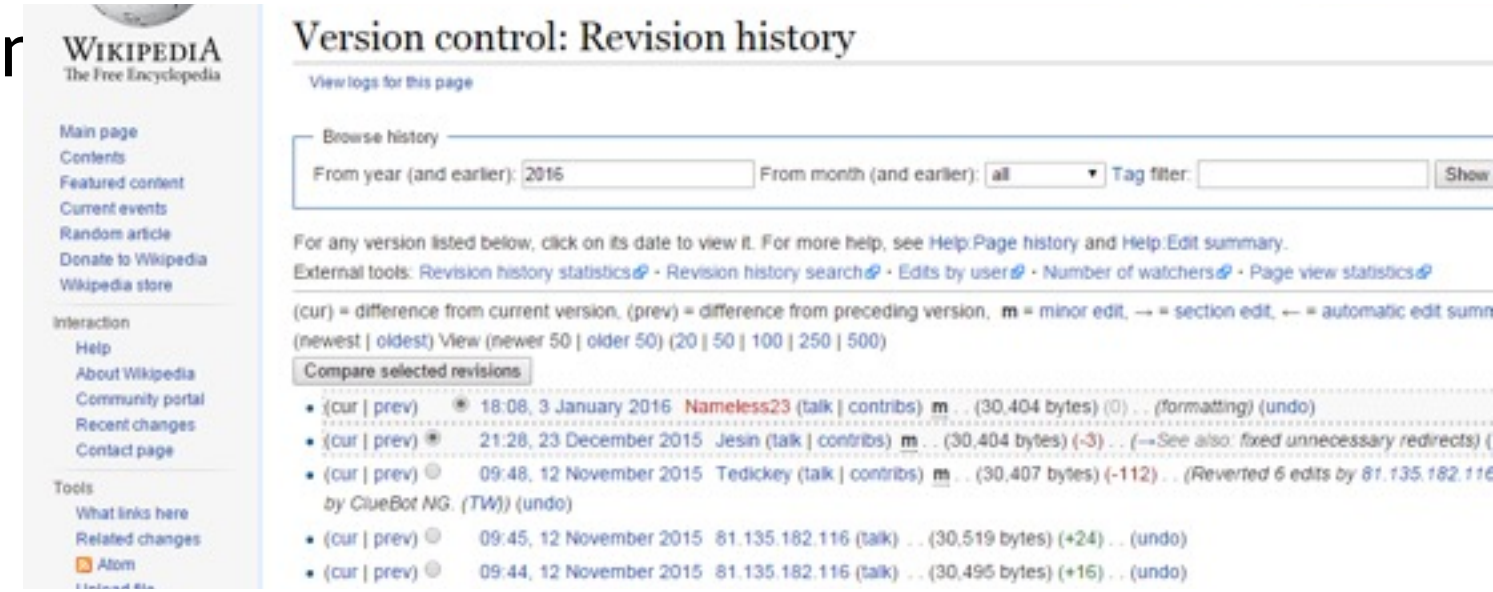
Why do we need version control?

- Record the changes to the documen
- Reflect on past work



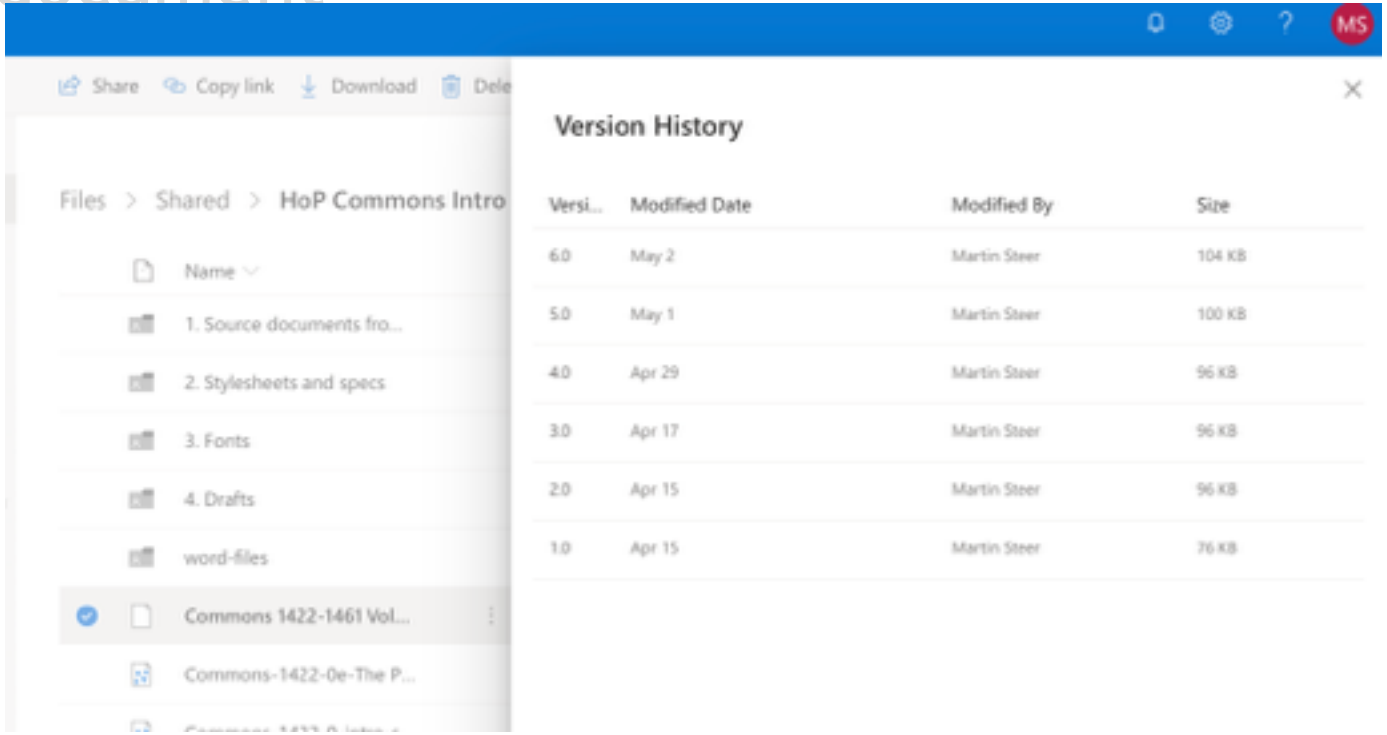
Why do we need version control?

- Record the changes to the document
- Reflect on past work
- Revert to previous versior



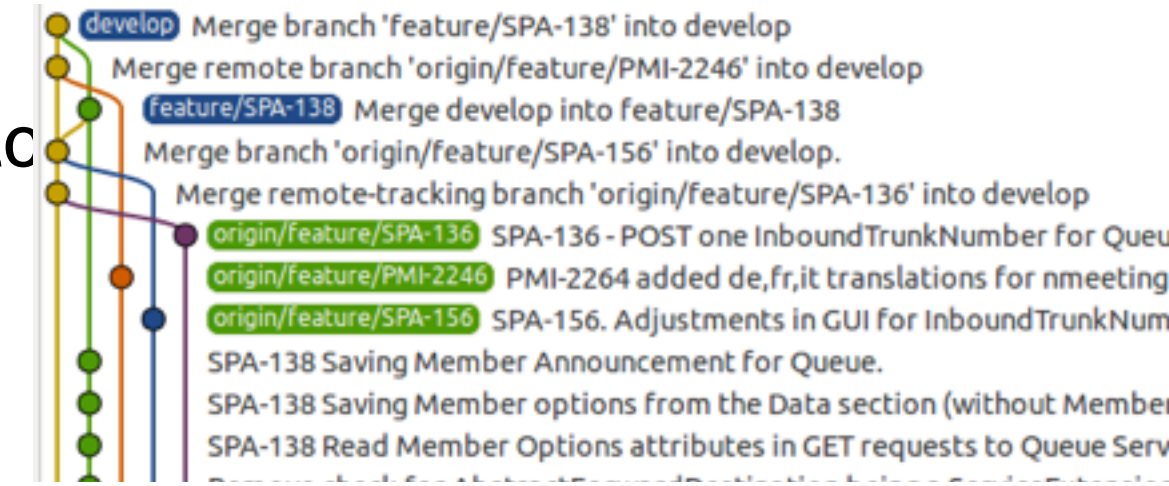
Why do we need version control?

- Record the changes to the document
- Reflect on past work
- Revert to previous versions
- Share with others

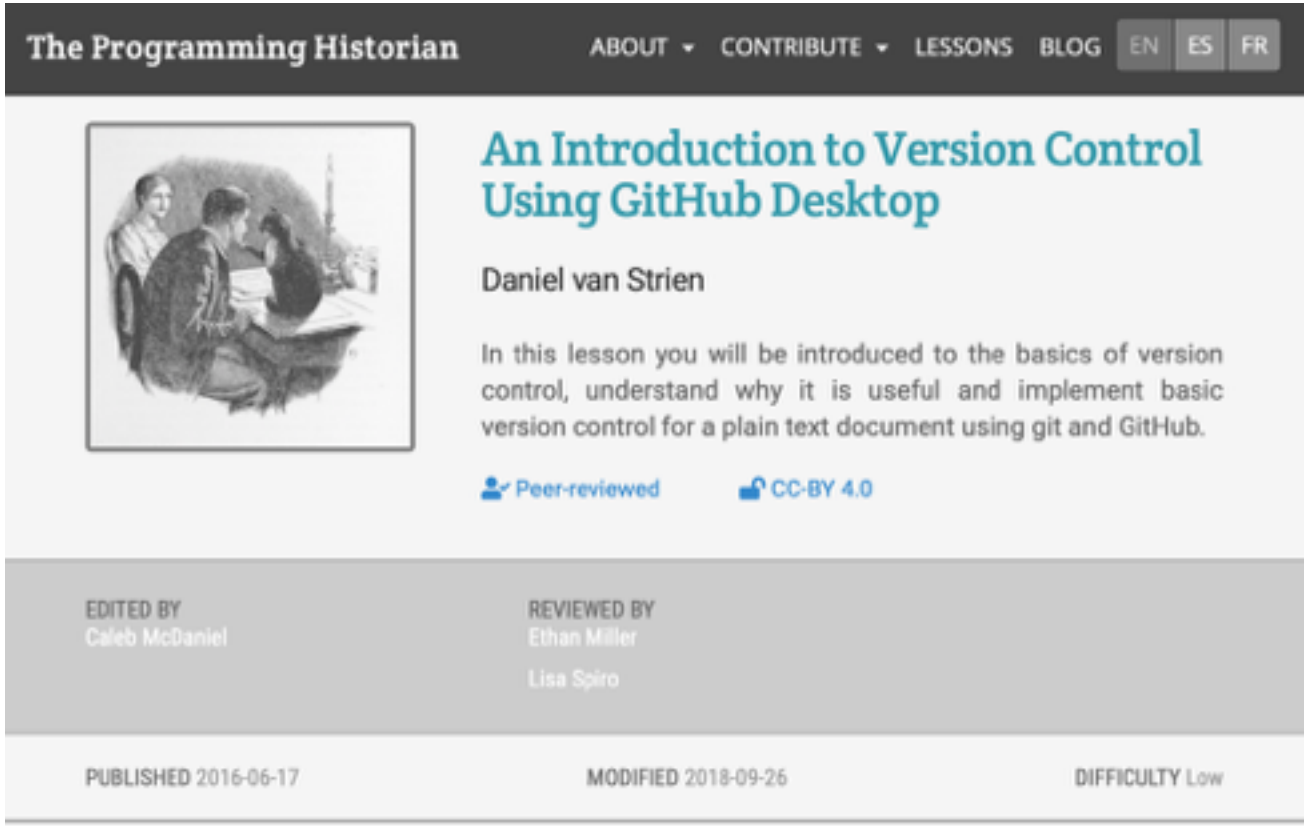


Why do we need version control?

- Record the changes to the document
- Reflect on past work
- Revert to previous versions
- Share with others
- Track important decisions (and acto



Teach yourself GIT!



https://programminghistorian.org/en/lessons/getting-started-with-github-desktop

Reproducibility

- Open data
- Quantitative methods
- Critical inquiry
- Scientific method
- Digital publishing

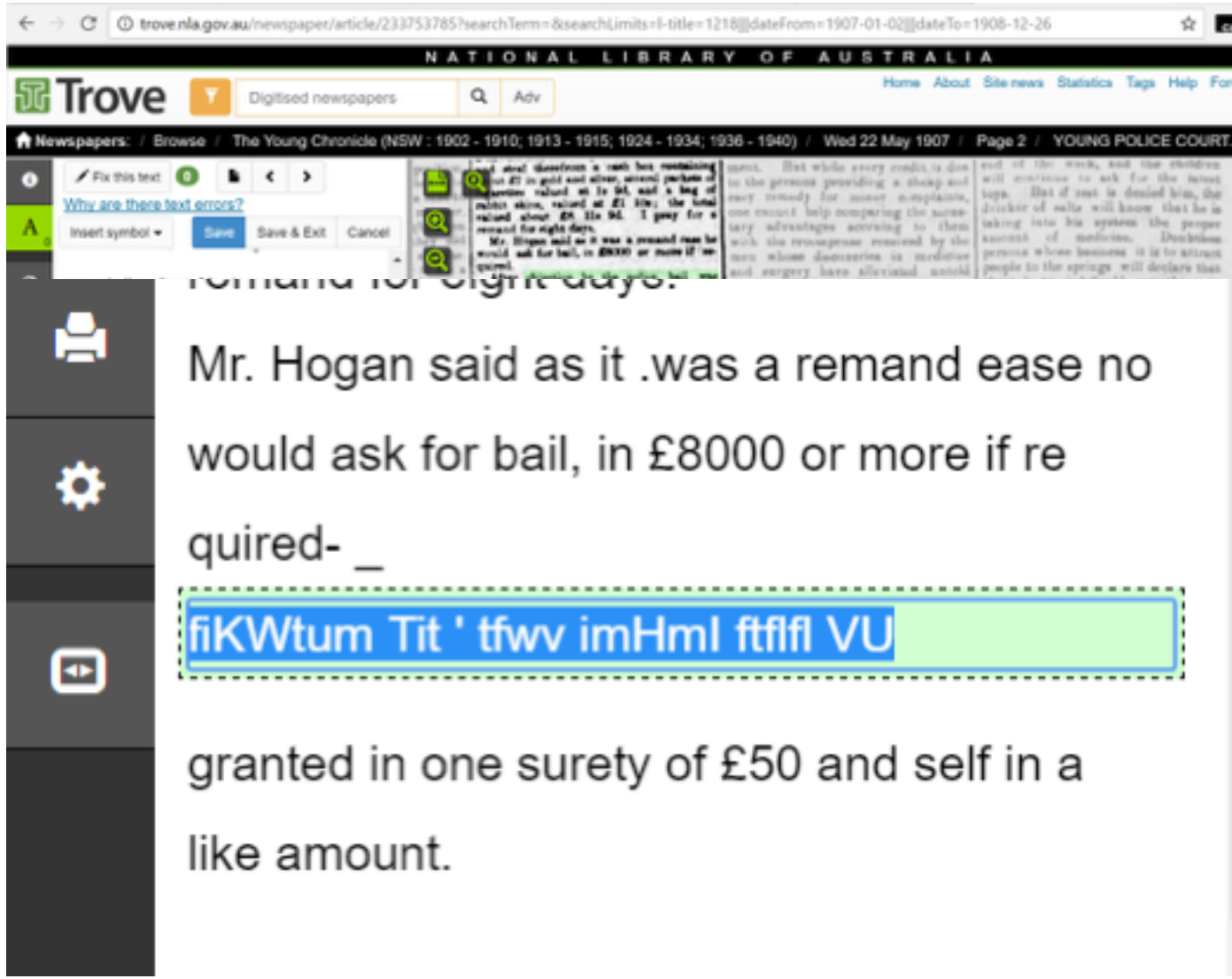
OPEN ACCESS

OPEN DATA

open source initiative

Reproducibility

- Open data
- Quantitative methods
- Critical inquiry
- Scientific method
- Digital publishing



Reproducibility

- Open data
- Quantitative methods
- Critical inquiry
- Scientific method
- Digital publishing

	2010	2011	2012	2013	2014
MacShane, Denis	100.00	25.00	33.33	N/A	N/A
Mactaggart, Fiona	54.55	100.00	47.37	37.50	N/A
Main, Anne	N/A	66.67	N/A	100.00	N/A
Mann, John	61.11	44.44	N/A	N/A	N/A
Maude, Francis	N/A	100.00	N/A	N/A	N/A
May, Theresa	70.83	62.86	53.41	47.58	52.08
McCabe, Steven	33.33	100.00	N/A	0.00	N/A

Figure 1. Extracted sentiment data for MPs on topic of immigration 2010-2014

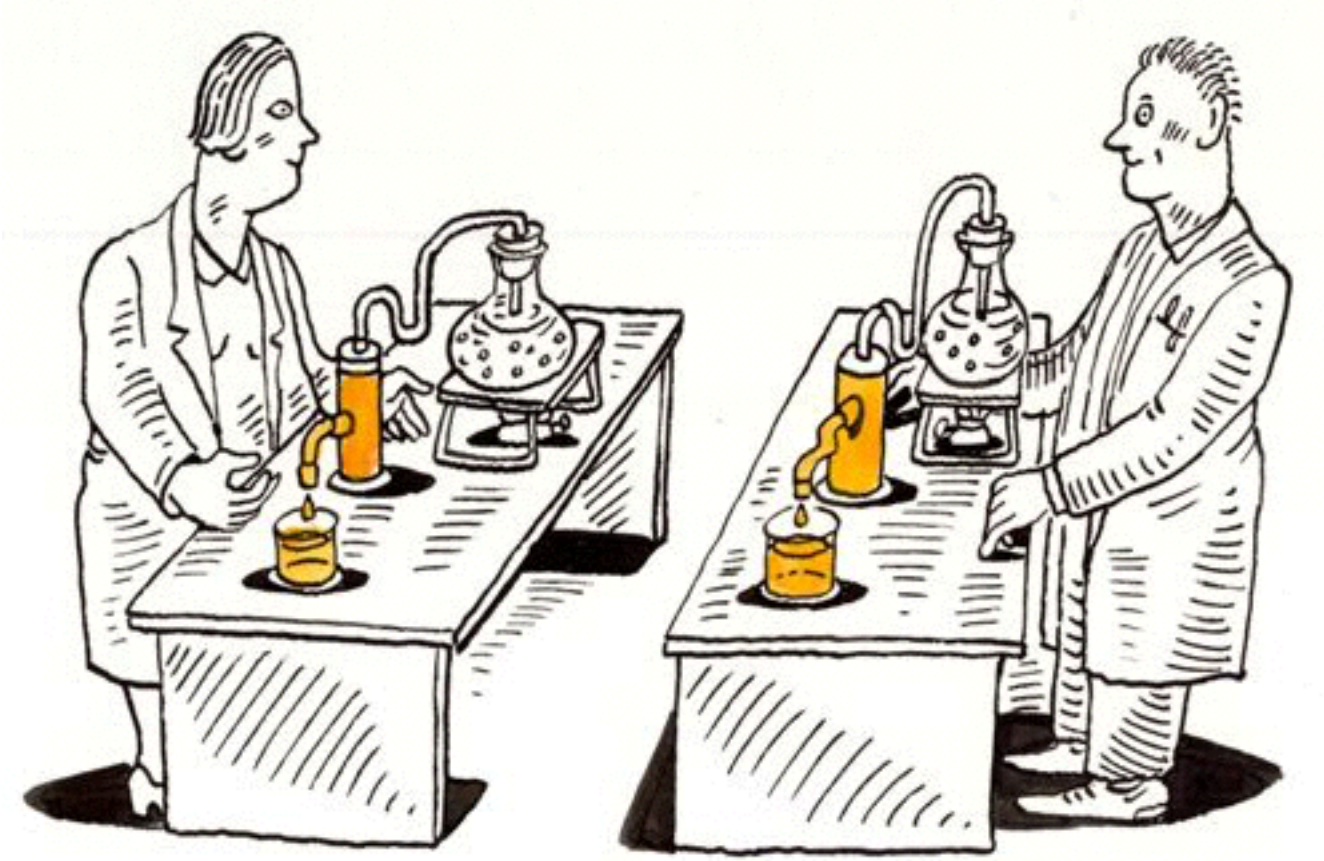
How could we verify this article about sentiment analysis?

Using structured text corpora in Parliamentary Metadata Language for the analysis of legislative proceedings, Richard Gartner, DHQ, http://www.digitalhumanities.org/dhqdev/vol/12/2/000392/000392.html

Reproducibility

- Open data
- Quantitative methods
- Critical inquiry
- Scientific method
- Digital publishing

Replicate vs. Reproduce findings
Scrutable vs. Transparent methods



Reproducibility

- Open data
- Quantitative methods
- Critical inquiry
- Scientific method
- Digital publishing

Original Articles

Confronting the Digital

Or How Academic History Writing Lost the Plot

Tim Hitchcock

Pages 9-23 | Published online: 01 May 2015

Download citation https://doi.org/10.2752/147800413X13515292098070

References Citations Metrics Reprints & Permissions Get access

ABSTRACT

This discussion piece argues that the design and structure of online historical resources and the process of search and discover embodied within them create a series of substantial problems for historians. Algorithm-driven discovery and misleading forms of search, poor OCR, and all the selection biases of a new edition of the Western print archive have changed how we research the past, and the underlying character of the object of study (inherited text). This piece argues that academic historians have largely failed to respond effectively to these challenges and suggests that while they have preserved the form of scholarly good practice, they have ignored important underlying principles.

Keywords: digital humanities, digital history, standards, scholarship, referencing, OCR, search

Reproducibility

- Open data
- Quantitative methods
- Critical inquiry
- Scientific method
- Shift to DATA publishing

Sharing and reuse

Sharing and reuse

- Open access
- Copyrights
- Permission
- Migration
- Citation

Sharing and reuse

- Open access
- Copyrights
- Permission
- Migration
- Citation

Sharing and reuse

- Open access?
- Copyrights
- Permission
- Migration
- Citation

Sharing and reuse

- Open access
- Copyrights
- Permission
- Migration
- Citation

- The biggest problem is often lack of information – the image may be multiple stages removed from any ‘original’

- Creative Commons licensing
  - CC0
  - CC BY
  - CC BY NC
  - CC BY ND

Sharing and reuse

- Open access
- Copyrights
- Permission
- Migration
- Citation

- Permission statement
- Rights metadata
- Licenses

Sharing and reuse

- Open access
- Copyrights
- Permission
- **Migration**
- Citation

ILAS
INSTITUTE OF LATIN AMERICAN STUDIES
SCHOOL OF ADVANCED STUDY
UNIVERSITY OF LONDON

→ SAS-Space

INSTITUTE OF LATIN AMERICAN STUDIES
SCHOOL OF ADVANCED STUDY
UNIVERSITY OF LONDON

Archive was built in 2012/13
Costs £1500/year to host
Managing legacy research data

ILAS
INSTITUTE OF LATIN AMERICAN STUDIES
SCHOOL OF ADVANCED STUDY
UNIVERSITY OF LONDON

1260 records
23GB page scans
1GB metadata

Migrate to another repository
Different data structures
Different taxonomies
Creates another ‘version’

Liberalism in the Americas Digital Archive

Home | Browse by Year | Browse by Themes | Browse by Country/Region | Browse by Author

Login

INSTITUTE OF LATIN AMERICAN STUDIES
SCHOOL OF ADVANCED STUDY
UNIVERSITY OF LONDON

Term mapping
Structural transform
Meaning change

Browse by Themes

Please select a value to browse from the list below.

- Church and State (91)
- Development and Ideas (1)
- Economic Development, Policy and Ideas (9)
- Economic Policy (1)
- Economic Policy, Development and Ideas (190)
- Political Culture (761)
- Race and Ethnicity (18)
- Women and Gender (24)

Browse by Subject

Please select a value to browse from the list below.

- Classics (195)
- Culture, Language & Literature (365)
- Economics (192)
- English (91)
- History (1262)
- Human Rights & Development Studies (352)
- Law (1019)
- Music (1962)
- Philosophy (473)
- Politics (463)
- Sociology & Anthropology (100)

Liberalism in the Americas Digital Archive

Home | Browse by Year | Browse by Themes | Browse by Country/Region | Browse by Author

Login

```
132428    <abstract>Series of lessons concerning political econc
132429    <date>1871</date>
132430    <publisher>Imprenta del Gobierno</publisher>
132431    <place_of_pub>Mexico City</place_of_pub>
132432    <pages>682</pages>
132433    <item_id>A00901</item_id>
132434    <folder>1560_1995</folder>
132435    <bl_cat_no>1560/1995</bl_cat_no>
132436    <themes>
132437      <item>Economic Policy, Development and Ideas</item>
132438    </themes>
132439    <countries>
132440      <item>MEX</item>
132441    </countries>
132442    <keywords_multi>
132443      <item>Property</item>
132444      <item>Industry</item>
132445      <item>Labour</item>
132446      <item>Capital</item>
132447      <item>Commerce</item>
132448      <item>Taxation</item>
132449      <item>Public Revenue</item>
132450      <item>Government Administration</item>
132451      <item>Domestic Commerce</item>
132452      <item>Currency</item>
132453    </keywords_multi>
132454  </eprint>
```

```
12        <collections>
13           <item>liberalism_in_the_americas</item>
14        </collections>
15        <note/>
16        <keywords>Item type: Memorandum;
17  &#xD;Countries: Mexico;
18  &#xD;Themes: Economic Policy, Development and Ideas;
19  &#xD;Other keywords: Property; Industry; Labour; Capi
20  &#xD;</keywords>
21        <referencetext>Project Record ID: A00901;
22  &#xD;Folder: 1560_1995;
23  &#xD;BL Catalogue number: 1560_1995;
24  &#xD;</referencetext>
```

Sharing and reuse

- Open access
- Copyrights
- Intellectual Property
- Permission
- Migration
  - Don’t underestimate data management costs!

Cost 10 days – over a period of 6 months!!