

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

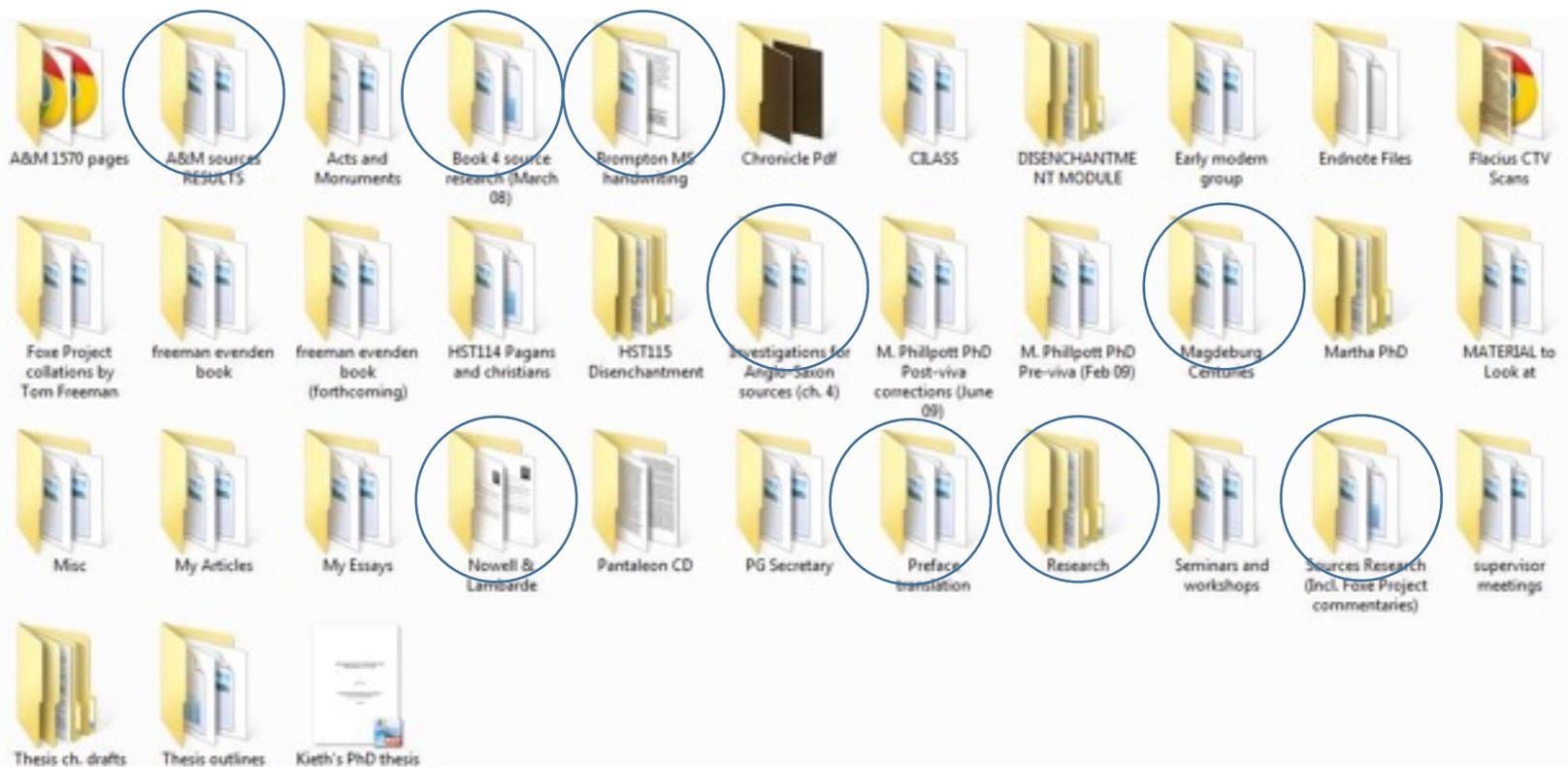


Good data management

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



Created data



My folder system today

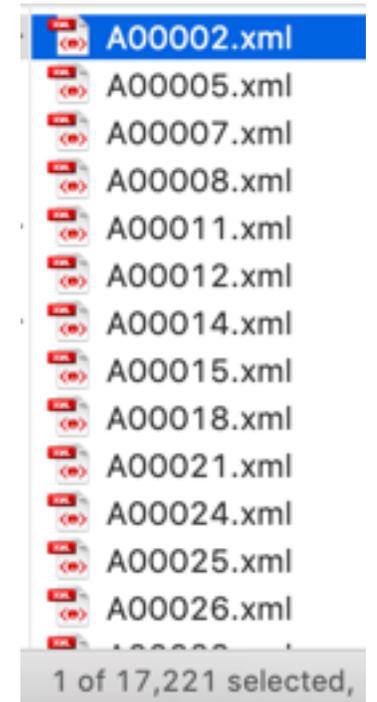
<input type="checkbox"/> Name	Date modified	Type	Size
 01 General	30/10/2015 07:06	File folder	
 02 Research	30/10/2015 07:08	File folder	
 03 Writing	30/10/2015 07:05	File folder	
 04 Presentations	16/07/2015 10:29	File folder	
 05 Articles	30/10/2015 07:07	File folder	
 06 Images	12/05/2015 09:01	File 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:

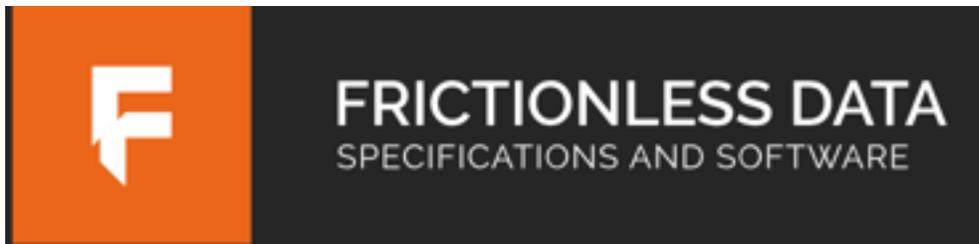
<https://github.com/textcreationpartnership/Texts>



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 as 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 your institution'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

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

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

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

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

Home Public DMPs Funder requirements Help

Welcome

DMPonline helps you to create, review, and share data management plans that meet institutional and funder requirements. It is provided by the Digital Curation Centre (DCC).

Join the growing international community that have adopted DMPonline:



17,622 Users



203 Organisations



23,083 Plans



89 Countries

Some funders mandate the use of DMPonline, while others point to it as a useful option. You can [download funder templates](#) without logging in, but the tool provides tailored guidance and example answers from the DCC and many research organisations. Why not sign up for an account and try it out?

Sign in Create account

* Email

* Password

Forgot password?

Remember email

Sign in

- or -

Sign in with your institutional credentials

Data formats, metadata, coding

Data formats, metadata, coding

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

Data deluge

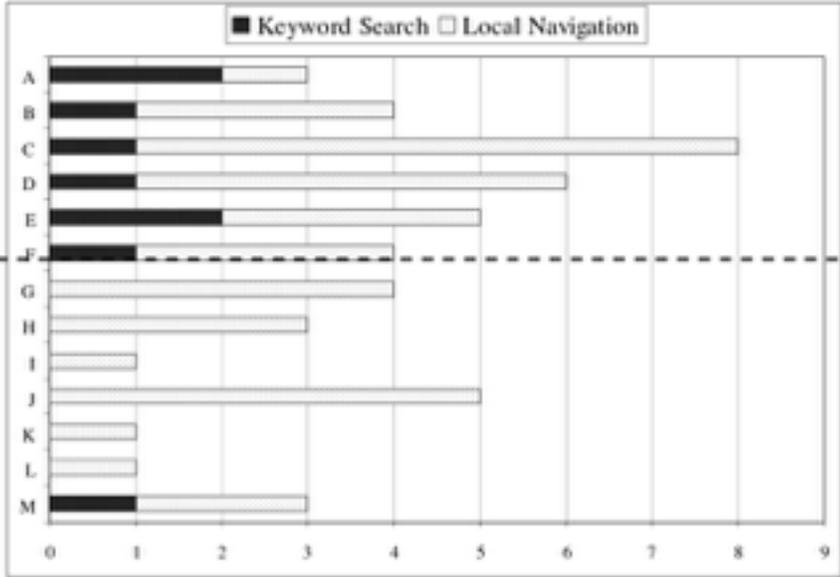


Figure 2: The number of times participants used each search tactic in their files.

Pilers ← → Filers

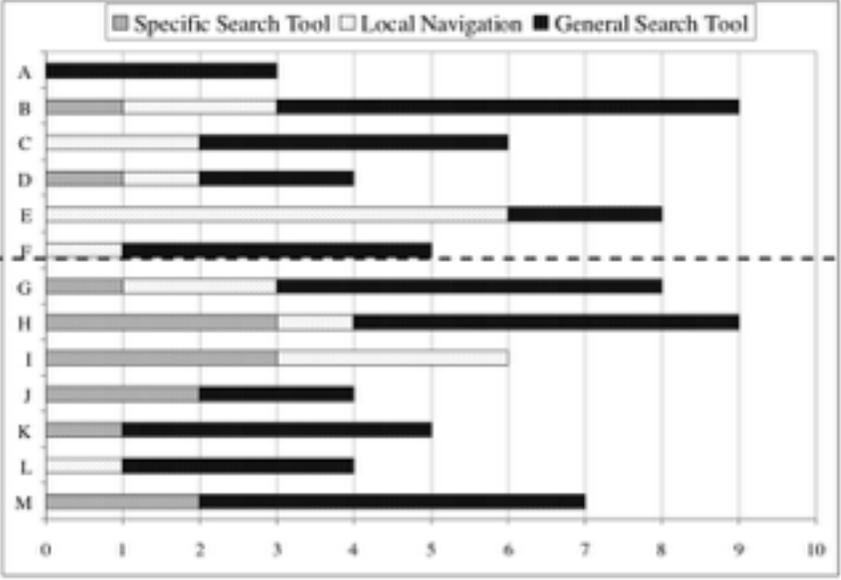


Figure 3: The number of times participants used each tactic on the Web. Note that pilers appear to use specific search tools more often.

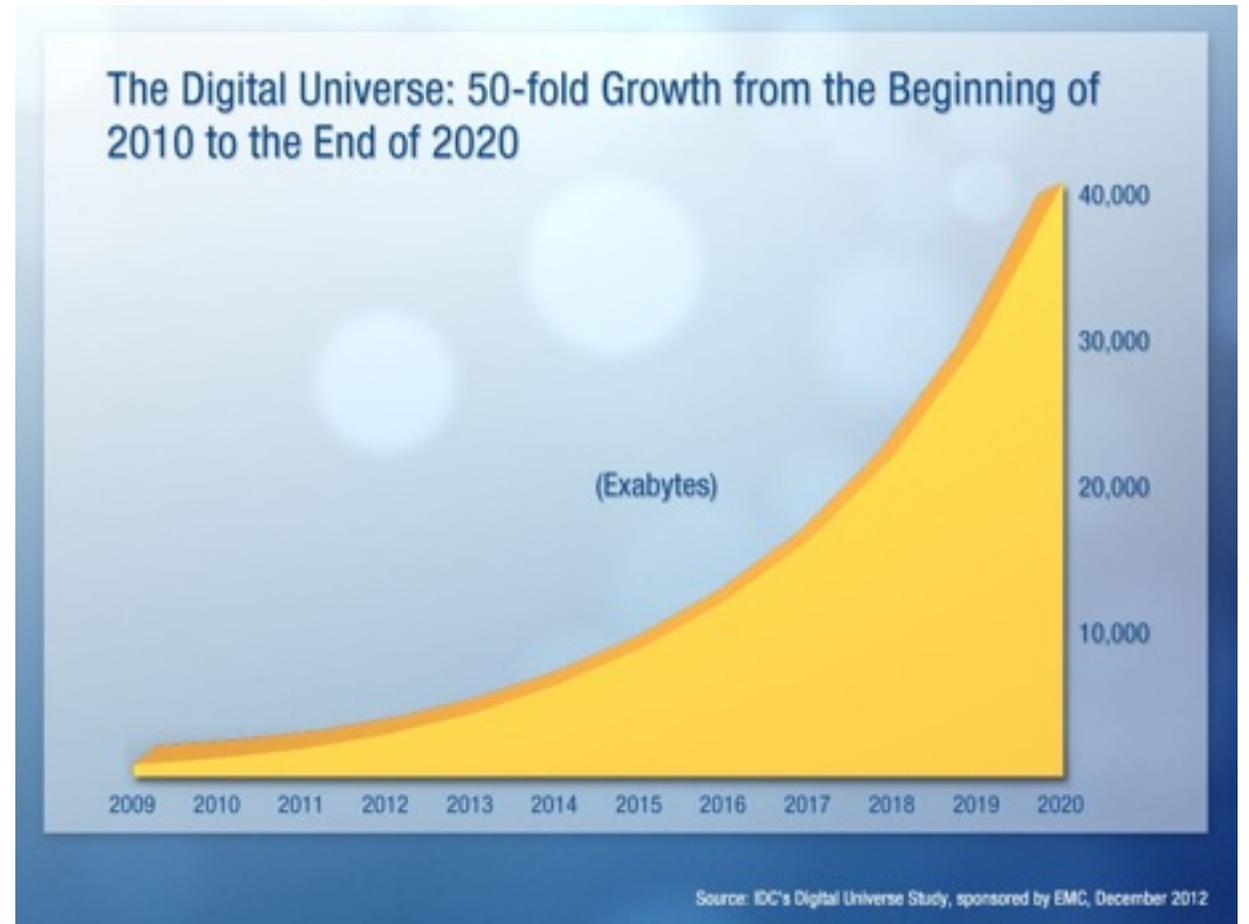
Malone, T. E. (1983). How do people organize their desks? *ACM Transactions on Office Information Systems* 1, 1, 99-112.
 Alvarado, Christine, et al. "Surviving the information explosion: How people find their electronic information." (2003).

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



Search 66,137,655 open access articles



Search 131,500,915 open access articles

CORE

Services About

 **CORE**

The world's largest collection of open access research papers

Search 135,539,113 papers around the world

 We aggregate and enrich open access research papers from around the world
[Read about our data](#)

 We provide seamless access to content and data, through our unique **APIs**
Perfect for **text mining!**

 We create powerful **services** for researchers, universities, and industry

<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

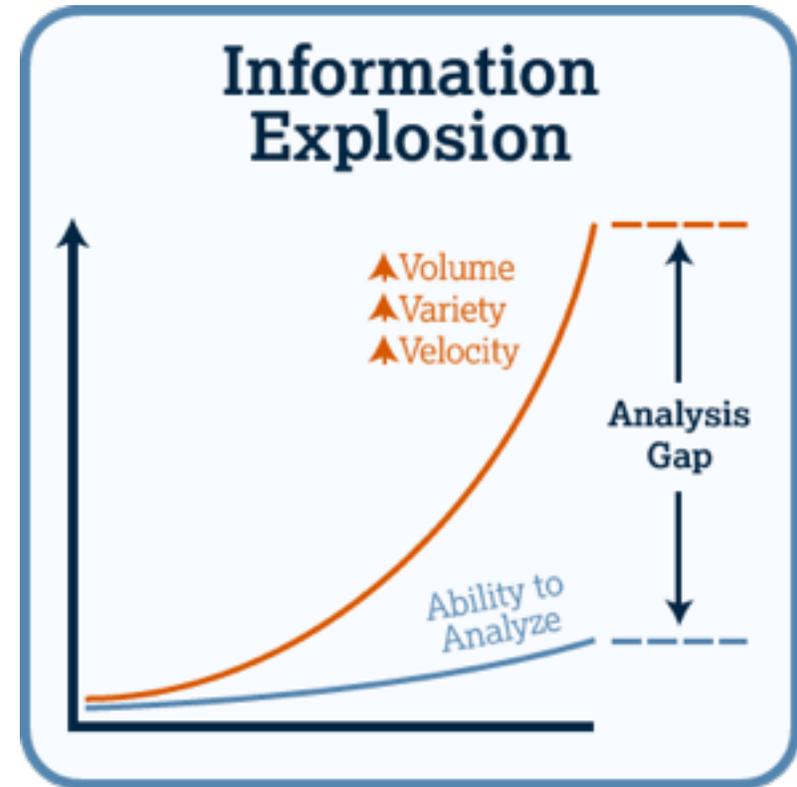


- billions of individual assets (pages, images, videos, pdfs etc.)
- As of 2017, collected approximately 500TB of data.
- Increasing by over roughly 60 – 70 TB a year.

<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)



Data deluge

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

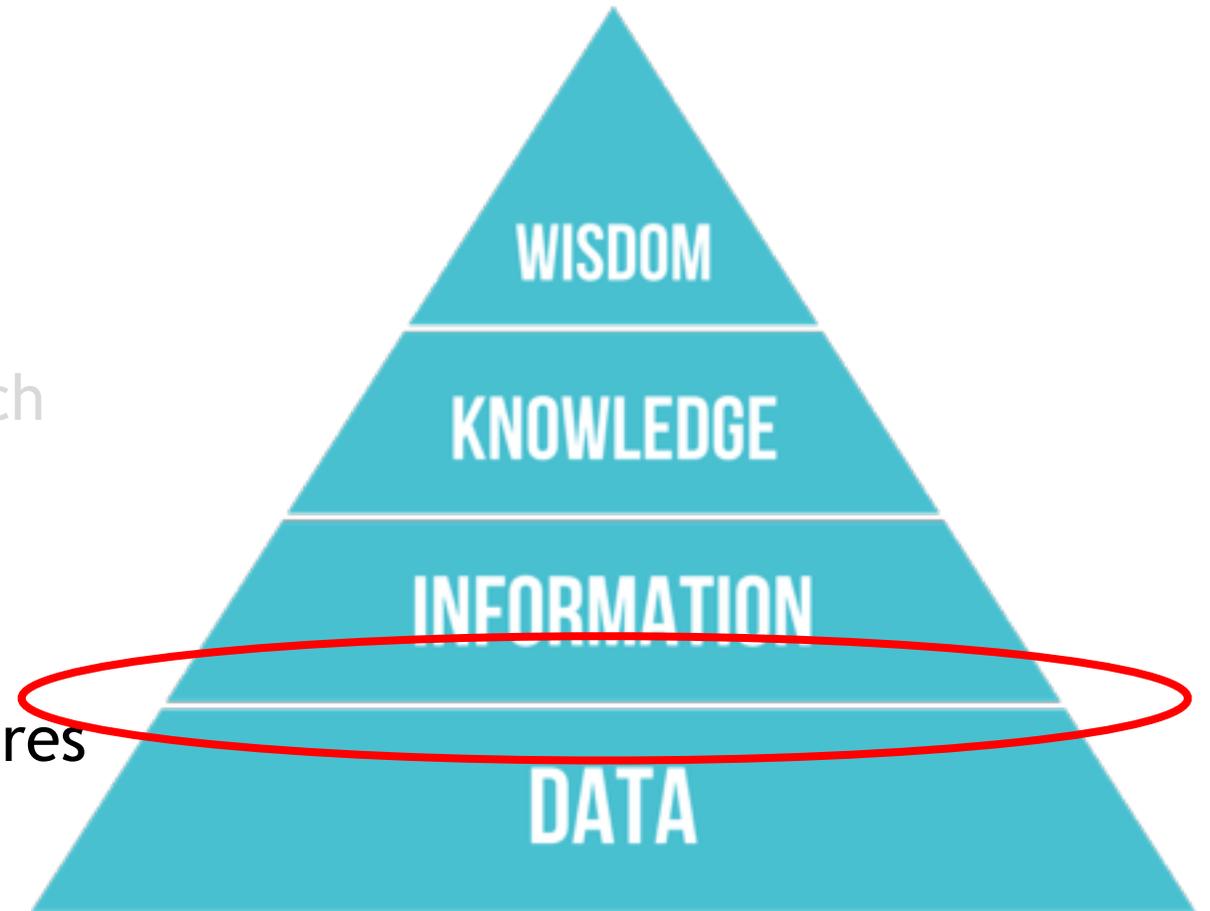
Data deluge

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



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



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

nomenclature
/nə(ʊ)'mɛŋklətʃə, 'nəʊmən, klɛtʃə/ ⓘ

noun
noun: nomenclature

the devising or choosing of names for things, especially in a science or other discipline.
"the Linnean system of zoological nomenclature"

- the body or system of names used in a particular specialist field.
plural noun: nomenclatures
"the students found it hard to decipher the nomenclature of chemical compounds"
- *formal*
the term or terms applied to someone or something.
"customers' was preferred to the original nomenclature 'passengers'"

Origin

The diagram shows the etymology of 'nomenclature'. It starts with 'LATIN' 'calare' (to call) leading to 'LATIN' 'clatura' (calling, summoning). 'LATIN' 'nomen' (name) also leads to 'LATIN' 'nomenclatura'. 'nomenclatura' then leads to 'FRENCH' 'nomenclature' in the early 17th century.

early 17th century: from French, from Latin *nomenclatura*, from *nomen* 'name' + *clatura* 'calling, summoning' (from *calare* 'to call').

Translate nomenclature to

Use over time for: nomenclature

The graph shows the number of mentions of 'nomenclature' over time. The y-axis is labeled 'Mentions' and the x-axis shows years from 1800 to 2010. The line shows a steady increase from 1800, peaking around 1950, and then declining towards 2010.

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.

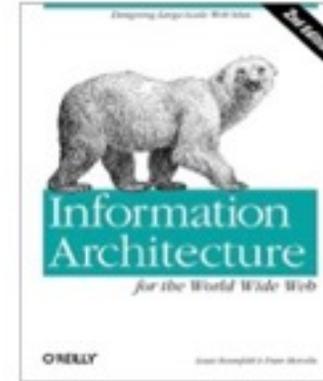
- *Richard Saul Wurman, 1996*

Information Architecture (IA)

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



1998



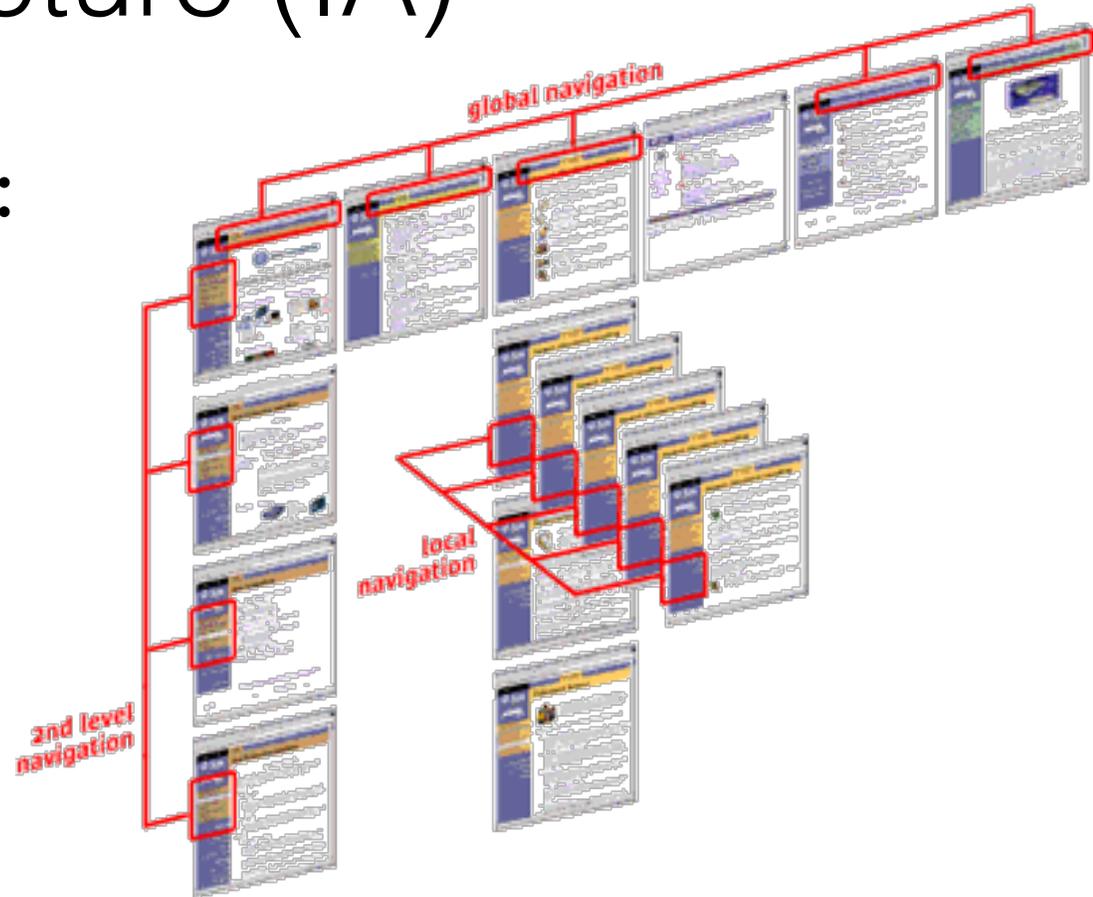
2002



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 implementation of core Frictionless Data libraries in R programming language. You can read more about this in [their grantee profile](#). In this 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, infer a schema, 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

```
library(tidytext)

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

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

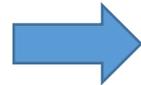
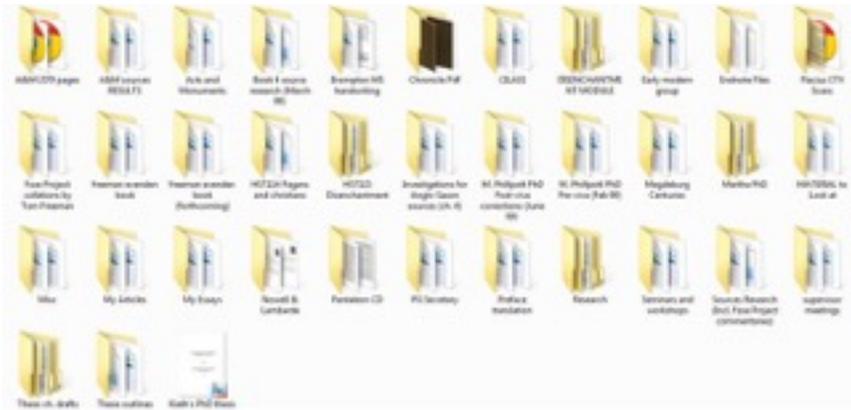
Naming things

- Importance of naming things
 - Long lived
 - Usually hard to change



Naming things

- Importance of naming things
 - Long lived
 - Reduces complexity



<input type="checkbox"/> 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	

Naming things

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

Tidy a list of terms up
and turn into data frame

```
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         2 adult        2
## 3         3 ago          1
## 4         4 alcohol       1
## 5         5 allegedly    1
## 6         6 allen          1
## 7         7 apparently    2
## 8         8 appeared      1
## 9         9 arrested      1
## 10        10 assault       1
## # ... with 302,021 more rows
```

Naming things

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

BRITISH HISTORY ONLINE

<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

BRITISH HISTORY ONLINE

www.british-history.ac.uk/old-new-london/vol4/pp216-226#p33

domain series volume page range paragraph ID

Providing structures

- Hierarchies
- Collations
- Conventions

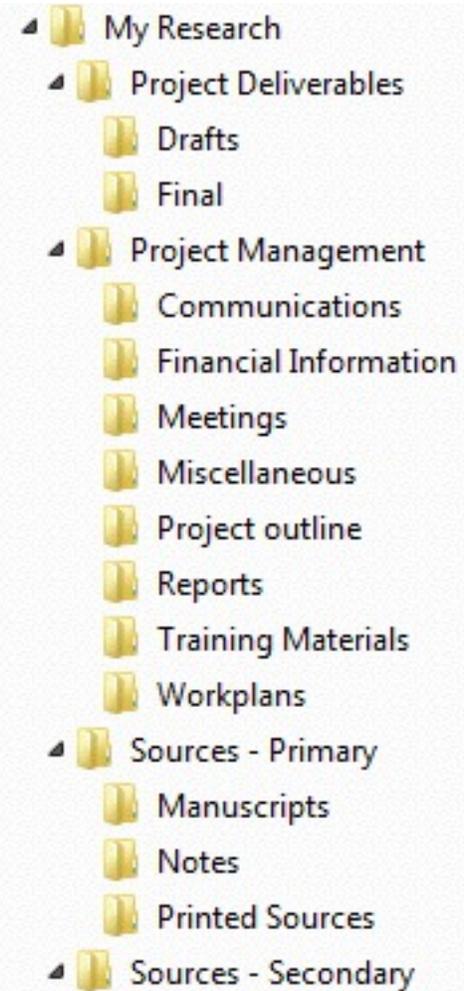
Using structures

- 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 classification
- Subject/Function
 - Type/categories
- Your own vocabulary
- Controlled Vocabulary



INSTITUTE OF
LATIN AMERICAN
STUDIES

SCHOOL OF ADVANCED STUDY
UNIVERSITY OF LONDON

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 \(24\)](#)

Using structures

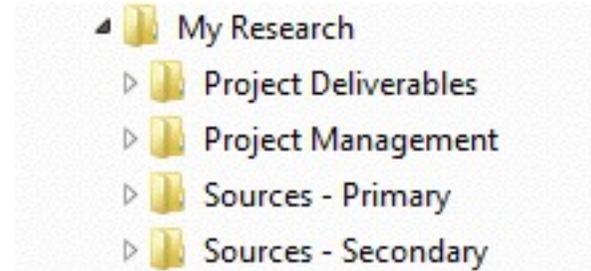
- Hierarchies
- Collations

Using structures

- Hierarchies
- Collations
 - Putting things into order

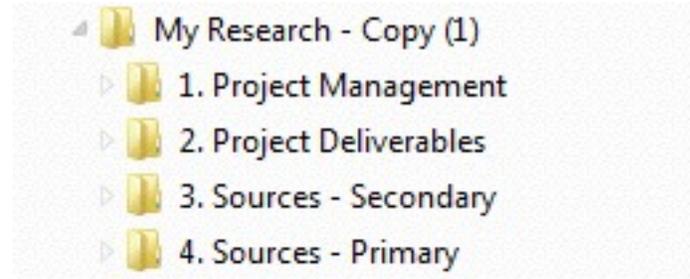
Using structures

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



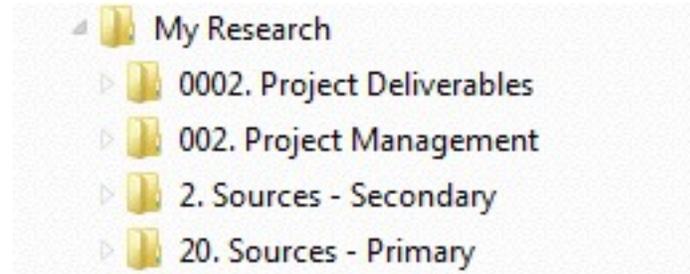
Using structures

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



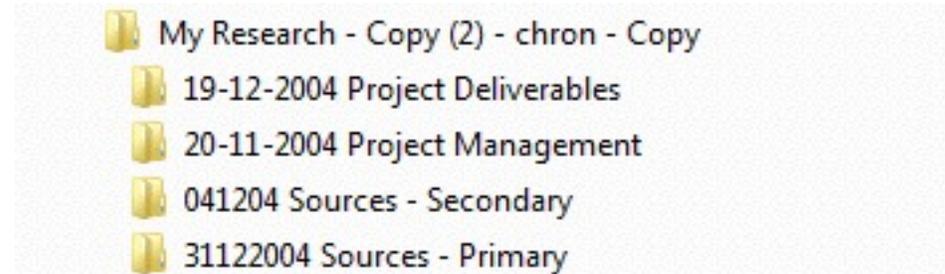
Using structures

- 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

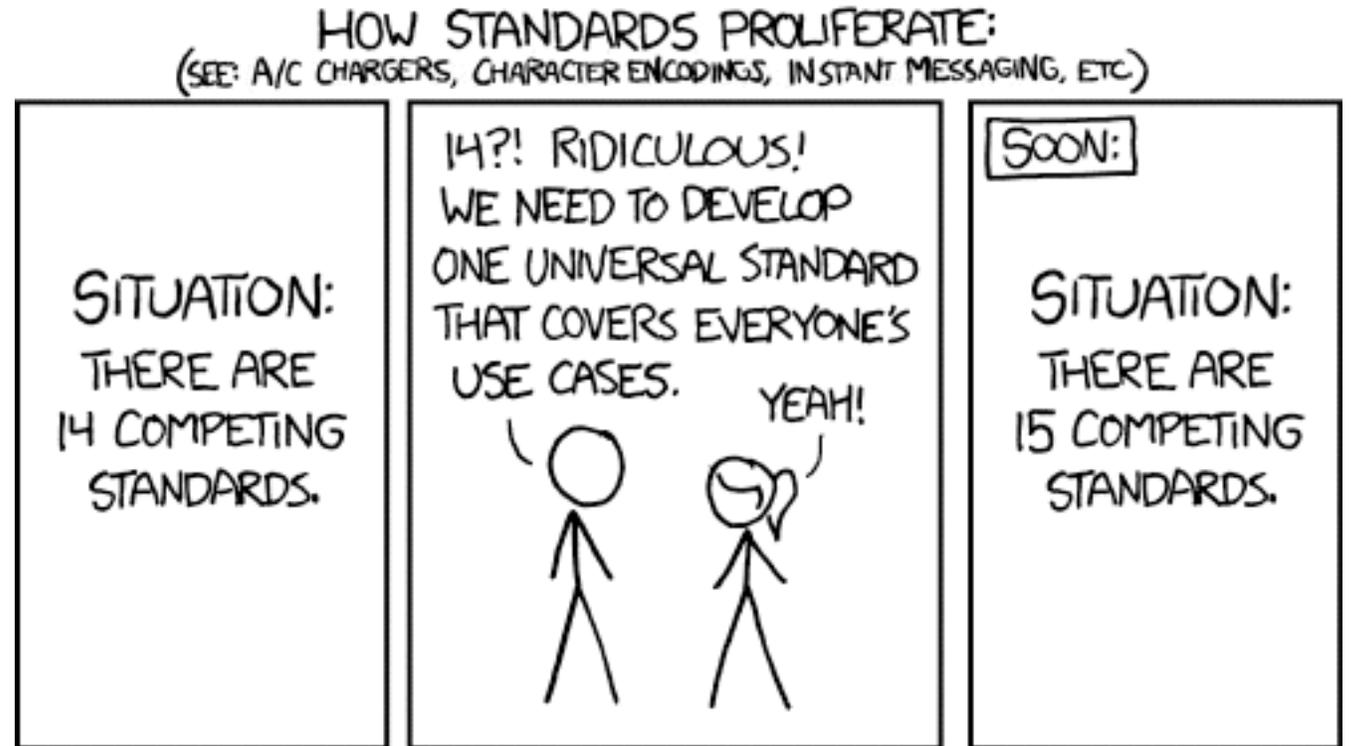
- Hierarchies
- Collations
 - Putting things into order
 - Alphabetical
 - Numerical
 - Chronological

A tidy text dataframe

```
## # A tibble: 50 x 16
##   author      datetimestamp      description heading  id   language origin topics
##   <chr>      <dtm>              <chr>         <chr>  <chr> <chr>  <chr>  <chr>
## 1 <NA>      1987-02-26 15:18:06 ""      COMPUTE... 10    en     Reute... YES
## 2 <NA>      1987-02-26 15:19:15 ""      OHIO MA... 12    en     Reute... YES
## 3 <NA>      1987-02-26 15:49:56 ""      MCLEAN'... 44    en     Reute... YES
## 4 By Cal... 1987-02-26 15:51:17 ""      CHEMLAW... 45    en     Reute... YES
## 5 <NA>      1987-02-26 16:08:33 ""      <COFAB ... 68    en     Reute... YES
## 6 <NA>      1987-02-26 16:32:37 ""      INVESTM... 96    en     Reute... YES
## 7 By Pat... 1987-02-26 16:43:13 ""      AMERICA... 110   en     Reute... YES
## 8 <NA>      1987-02-26 16:59:25 ""      HONG KO... 125   en     Reute... YES
## 9 <NA>      1987-02-26 17:01:28 ""      LIEBERT... 128   en     Reute... YES
## 10 <NA>     1987-02-26 17:08:27 ""      GULF AP... 134   en     Reute... YES
```

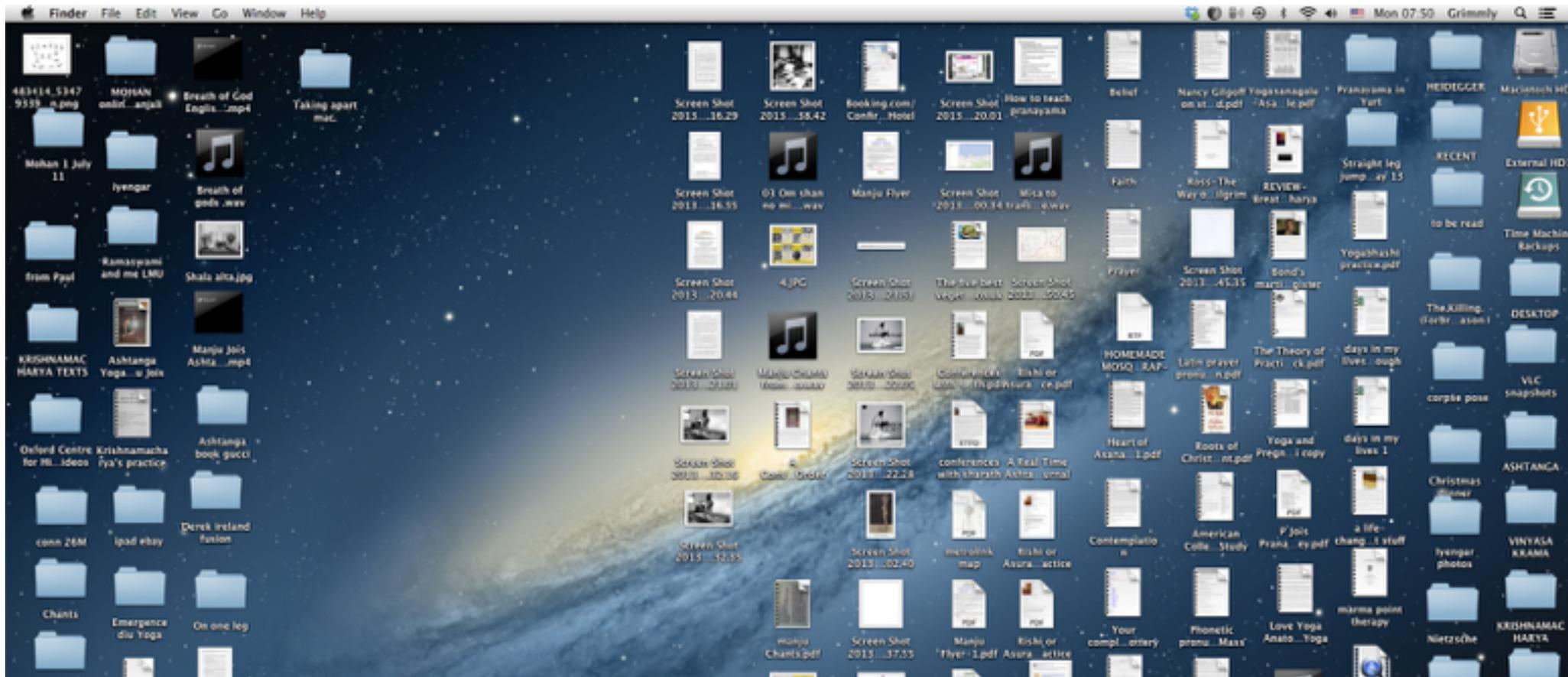
Using structures

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



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

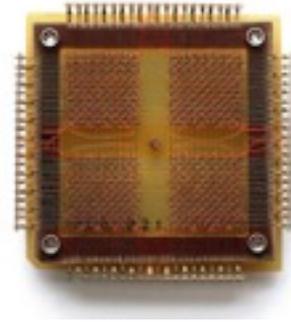
Storage, preservation and sustainability

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



Storage, preservation and sustainability

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



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

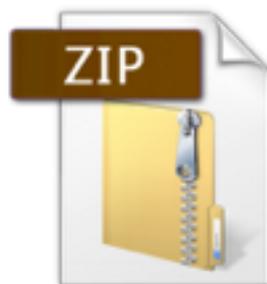
Storage, preservation and sustainability

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



Storage, preservation and sustainability

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



Closed
vs.
Open
standards



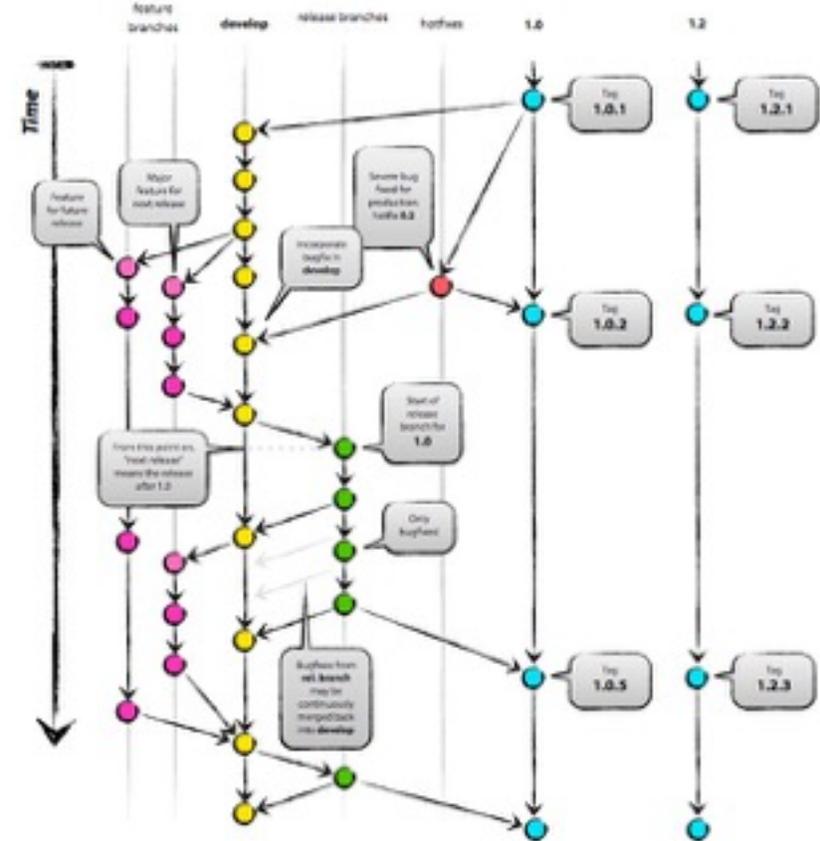
Storage, preservation and sustainability

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

GitHub



Bitbucket



Storage, preservation and sustainability

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



RefWorks



Endnote



Mendeley



Storage, preservation and sustainability

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



Related Images

Tags

- bldigital
- date1896
- pubplaceipzig
- publicdomain
- sysnum003335449
- semonrichardwolfgang
- large
- vol0
- page61
- mechanicalcurator
- imagesfrombook003335449
- imagesfromvolume0033354490
- fauna
- wildlife
- kuala
- sherlocknet:tag=animal
- sherlocknet:tag=tree
- sherlocknet:tag=hind
- sherlocknet:tag=burnett
- sherlocknet:tag=australian
- sherlocknet:tag=america
- sherlocknet:tag=hind
- sherlocknet:tag=house
- sherlocknet:tag=nature
- sherlocknet:tag=state
- sherlocknet:tag=jack
- sherlocknet:tag=specimen
- sherlocknet:tag=work
- sherlocknet:tag=camp
- sherlocknet:tag=dog
- sherlocknet:tag=creature
- sherlocknet:category=organism

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 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.”

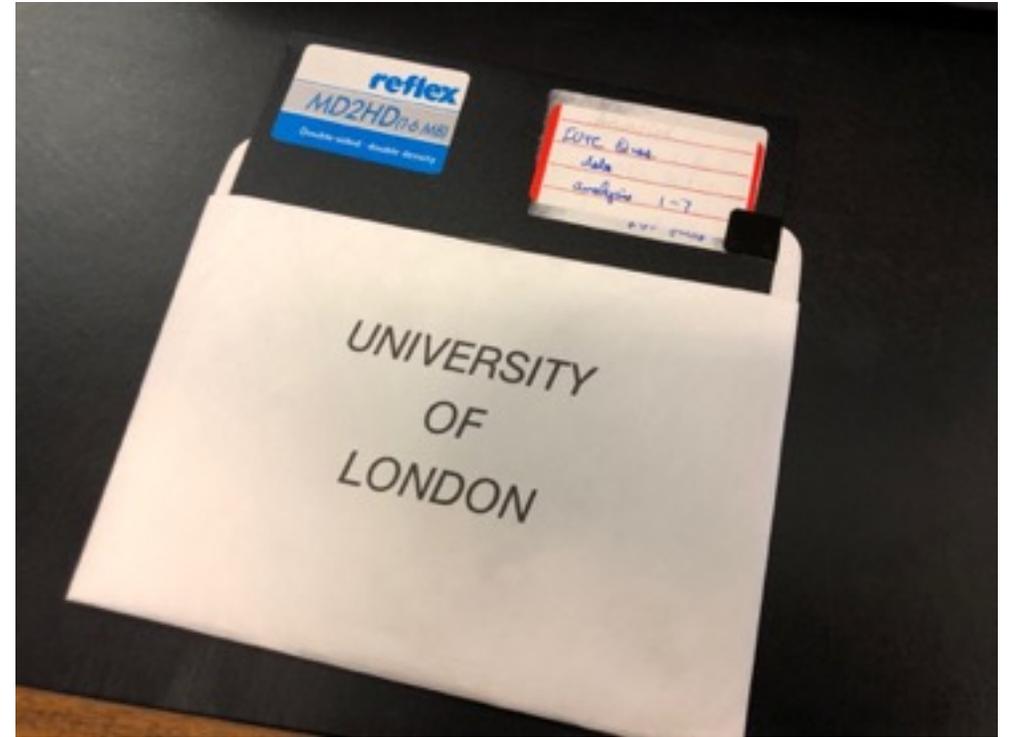
“I left my USB stick on the train with a variety of personal research projects on it. I had to reconstruct the information trails but the notes from secondary sources were lost and not all of the sources are available in my library meaning I will eventually have to pay out for another train fare and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

“I 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

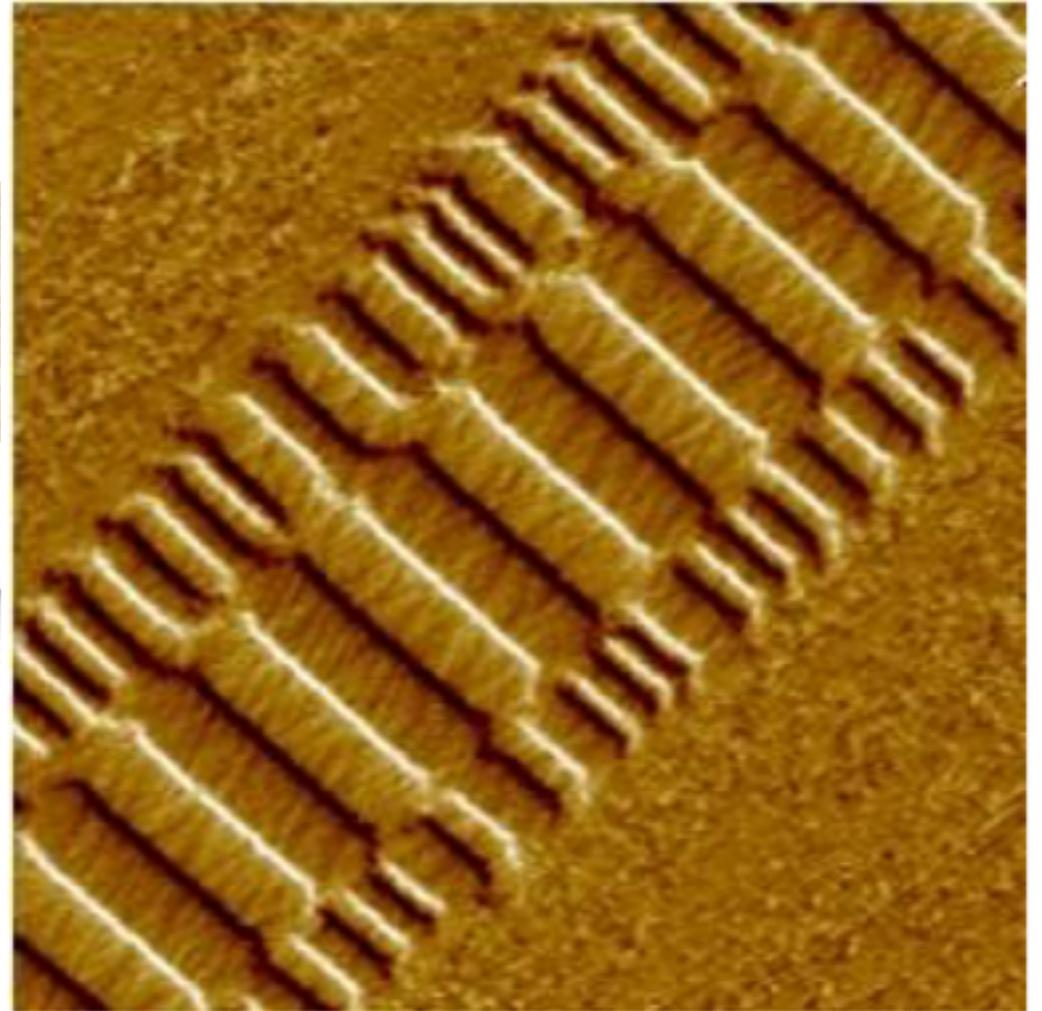
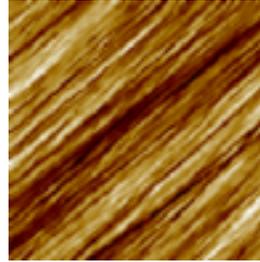


USASCII code chart

Bits					Column	0	1	2	3
b ₇	b ₆	b ₅	b ₄	b ₃	Row	0 0 0	0 0 1	0 1 0	0 1 1
↓	↓	↓	↓	↓	0	0	1	2	3
0	0	0	0	0	0	NUL	DLE	SP	0
0	0	0	1	1	1	SOH	DC1	!	1

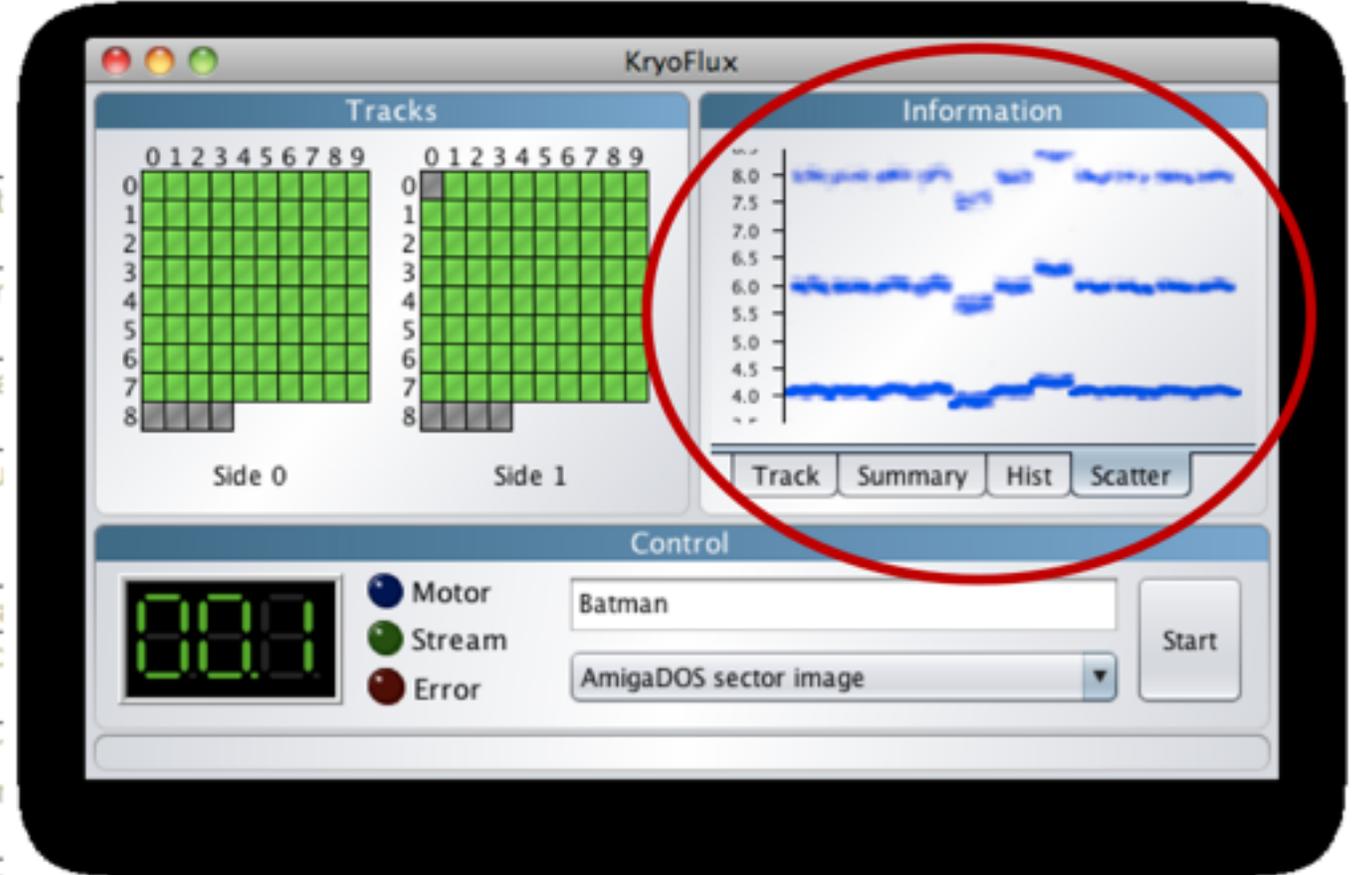
Backups

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



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



Welcome to SAS-Space

SAS-Space is an online library for humanities research outputs, providing a permanent archive for scholars and researchers.



Types of storage

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



Types of storage

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



The image shows two overlapping website screenshots. The left screenshot is from the UK Web Archive (UKWA) website, featuring a search bar and a 'What we do' section. The right screenshot is from the European Open Science Cloud (EOSC) website, featuring a navigation menu and a news article about reports published in November 2018.

UKWA UK WEB ARCHIVE

Home Topics and Themes Save a UI

Search the UK Web Archive

Enter a specific website URL (e.g. www.bl.uk) or any word

Tips/Notes for using the UK Web Archive

What we do

The UK Web Archive (UKWA) collects millions of websites each year, preserving them for future generations. The UKWA is a partnership of the six [UK Legal Deposit Libraries](#).

Use this site to discover old or obsolete versions of UK websites, search the websites and browse websites curated on different topics and themes.

European Open Science Cloud (EOSC)

Home > Research and Innovation > Strategy > Goals of research and innovation policy > Open Science

This is a cloud for research data in Europe. Background, policy information, events and publications related to the EOSC

Home Open Access European Open Science Cloud Open Science Policy Platform

The reports "Prompting an EOSC in practice" and "Turning FAIR into reality" have been published

29 November 2018

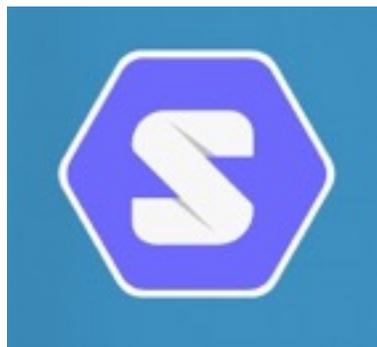
In the perspective of the launch of the European Open Science Cloud (EOSC) implementation phase 2018-2020, two important reports are being published by the Commission that constitute major sources of strategic orientations and concrete actions for the new EOSC governance structure:

- [Prompting an EOSC in practice](#)
Report of the Commission 2nd High Level Expert Group on the European Open Science Cloud (EOSC 2nd HLEG)
- [Turning FAIR into reality](#)
Report of the Commission FAIR Data Expert Group (FAIR Data EG)

WayBackMachine

Types of storage

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

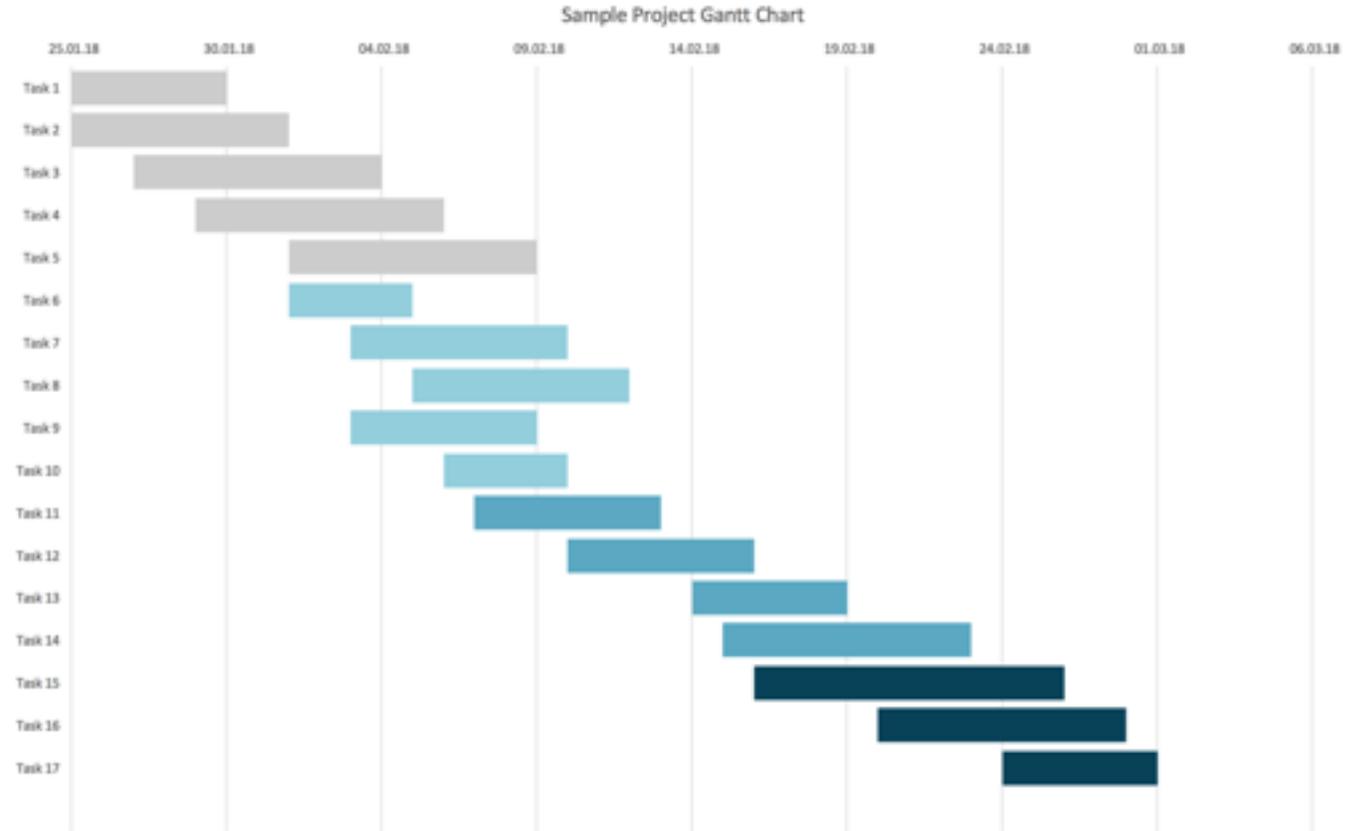


Documentation

- Administrative
- Methodological
- Descriptive
- Technical

Documentation

- Administrative
- Methodological
- Descriptive
- Technical



Documentation

- Administrative
- Methodological
- Descriptive
- Technical

Narrative Text

Notebook title and introduction

Description of model parameters

Description of need to profile data



Code and Visualizations

Importing external packages

Implementation of parameters

Profile plotting code

Inline plot

Sampling from the generative model

In this notebook, we will use the generative model of the ICDP International Directed Network Process in order to sample events. We will start with a predefined number of users, say 10, and we will attempt to model their behavior as they are posting questions in an online platform. For simplicity, our "network" will be during:

We start by importing all the libraries that will be required:

```
In [1]: %matplotlib inline
import numpy as np
import pandas as pd
import networkx as nx
import random
import time
```

Next, let us set some parameters for our model. These fall under two categories, the ones relevant to the content and their ones relevant to the time dynamics. Starting with the first set, we need to decide on:

- the vocabulary or dummy set of 100 words, i.e. words, words, ... words
- the minimum and maximum length of a question
- the number of words of each pattern

As for the time dynamics to concerned, we need to set:

- α , the parameters of the Gamma prior for the time interval of each pattern
- β , the parameters of the Gamma prior for the user activity rate
- γ , the time decay parameter

Finally, in order to make the generative process more user-friendly, we can pre-set the number of patterns that our users can sample from:

```
In [2]: vocabulary = ['word' + str(i) for i in range(100)] # list 'words' of our dictionary
min_len_questions = 5
max_questions = 10
words_per_pattern = 10
alpha_1 = (0.5, 0.75)
alpha_2 = (0.5, 0.5)
beta = 1.0
gamma = 0.5
max_patterns = 10
process = nxp.icdp_network_process(vocabulary, alpha_1=alpha_1,
                                 alpha_2=alpha_2,
                                 beta=beta, words_per_pattern=words_per_pattern,
                                 min_len_questions, max_questions)
```

Before generating any questions, we can take a look at the patterns that we initialized our process with, and look at the content distribution of each pattern. Although each pattern has a different word distribution, we can still plot the average content weighting between the words that each user uses probably for each pattern. Since we used a fixed number of patterns, the distribution of the average will not be smooth:

```
In [3]: word_tag = networkx.average_neighbor_degree(process)
min_avg_content_weight = min(word_tag.values())
max_avg_content_weight = max(word_tag.values())
average_content_weight = (min_avg_content_weight + max_avg_content_weight) / 2
```

Plotting the average content weighting for the initialization

Documentation

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

Levels of research data documentation



Project level: What the study set out to do, how it contributes to new knowledge in the field, what the research questions/hypotheses were, what methodologies were used, what sampling frames were used and what instruments and measures were used.



File or database level: How all the files (or tables in a database) that make up the dataset relate to each other; what format they are in, whether they supersede or are superseded by previous files. A readme-txt file is a classic way of accounting for all the files and folders in a project.



Variable or item level: The key to understanding research results is knowing exactly how an object of analysis came about. Not just, for example, a variable name at the top of a spreadsheet file, but the full label explaining the meaning of that variable in terms of how it was used in the project.

Documentation

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



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

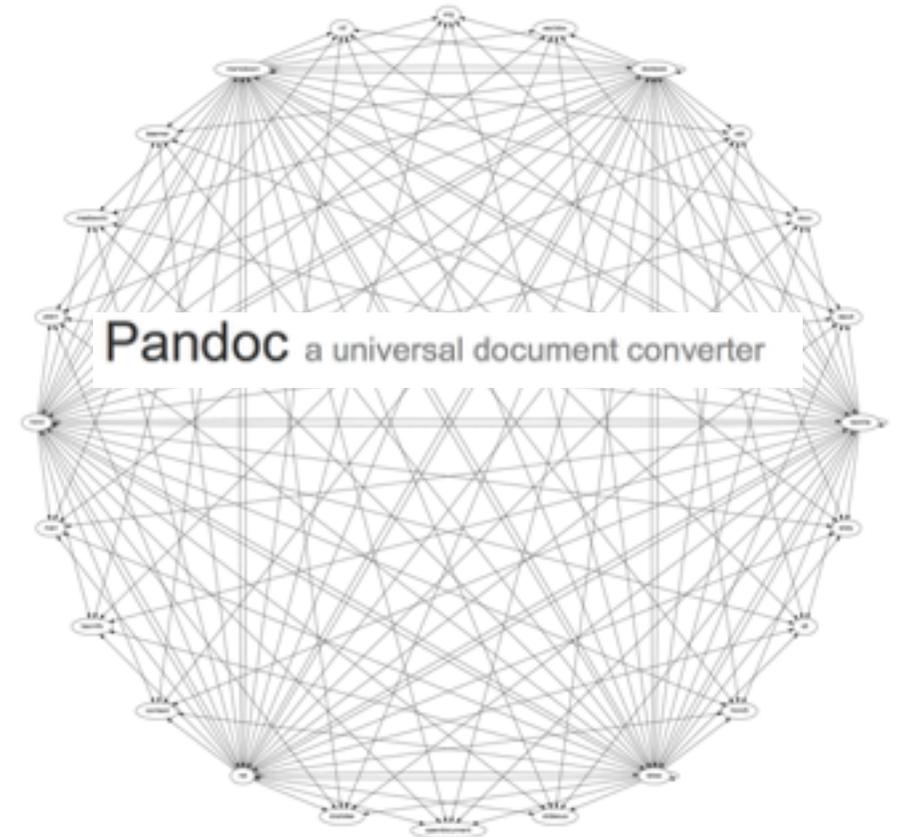
The UK Data Archive, [Managing and Sharing Data booklet](https://ukdataservice.ac.uk/media/622417/managingsharing.pdf) - <https://ukdataservice.ac.uk/media/622417/managingsharing.pdf>

Documentation

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

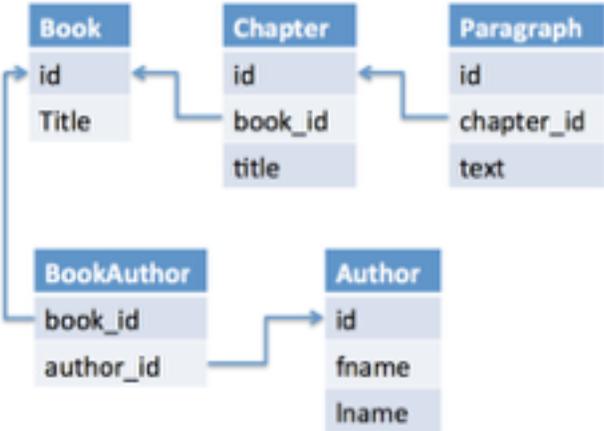
Open or closed
Formats?

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



Documentation

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



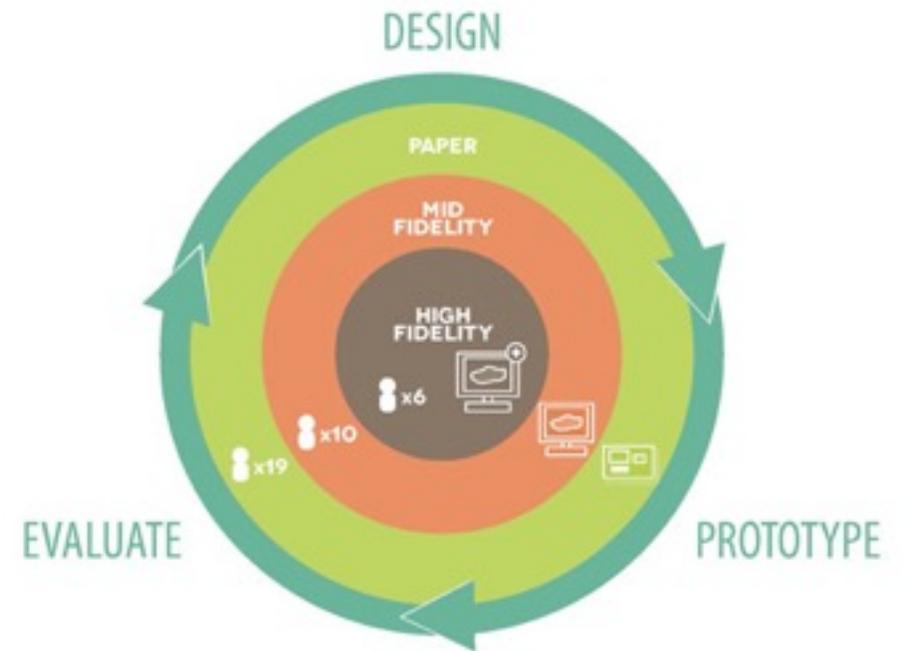
Why do we need version control?

- Record the changes to the document

```
home/mantis/apache2/httpd.conf | home/mantis/apache2.old/httpd.conf
# This is a modification of the default Apache 2.2 configuration file | # This is a modification of the default Apache 2 configuration
# for Gentoo Linux. | # file by Gentoo Linux. .... [insert more]
# | #
# Support: | # Support:
# http://www.gentoo.org/main/en/lists.xml [mailing lists] | # http://www.gentoo.org/main/en/lists.xml [mailing lists]
# http://forums.gentoo.org/ [web forums] | # http://forums.gentoo.org/ [web forums]
# irc://irc.freenode.net#gentoo-apache [irc chat] | #
# | #
# Bug Reports: | # Bug Reports:
# http://bugs.gentoo.org [gentoo related bugs] | # http://bugs.gentoo.org/ [gentoo related bugs]
# http://httpd.apache.org/bug_report.html [apache httpd related bugs] | # http://bugs.apache.org/ [apache httpd related bugs]
# | #
# This is the main Apache HTTP server configuration file. It contains the | #
# configuration directives that give the server its instructions. | #
# See <URL:http://httpd.apache.org/docs/2.2> for detailed information. | #
# In particular, see | # Based upon the NCSA server configuration files originally by Rob McCool.
# <URL:http://httpd.apache.org/docs/2.2/mod/directives.html> | #
# for a discussion of each configuration directive. | #
# This is the main Apache server configuration file. It contains the
# configuration directives that give the server its instructions.
# See <URL:http://httpd.apache.org/docs-2.0> for detailed information about
# the directives.
# Do NOT simply read the instructions in here without understanding | #
# what they do. They're here only as hints or reminders. If you are unsure | #
# consult the online docs. You have been warned. | #
# | #
# Configuration and logfile names: If the filenames you specify for many | #
# of the server's control files begin with '/' (or 'drive:/' for Win32), the | #
# server will use that explicit path. If the filenames do *not* begin | #
# with '/', the value of ServerRoot is prepended -- so "var/log/apache2/foo_log" | #
# with ServerRoot set to "/usr" will be interpreted by the | #
# The configuration directives are grouped into three basic sections:
# 1. Directives that control the operation of the Apache server process as a
# whole (the 'global environment').
```

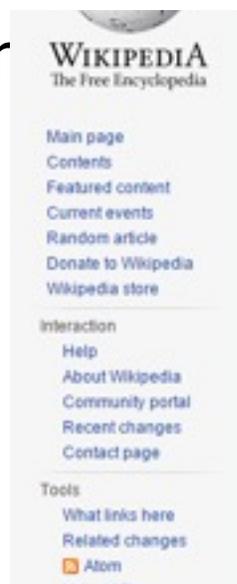
Why do we need version control?

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



Why do we need version control?

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



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Atom

Version control: Revision history

View logs for this page

Browse history

From year (and earlier): 2016

From month (and earlier): all

Tag filter:

Show

For any version listed below, click on its date to view it. For more help, see [Help:Page history](#) and [Help:Edit summary](#).

External tools: [Revision history statistics](#) · [Revision history search](#) · [Edits by user](#) · [Number of watchers](#) · [Page view statistics](#)

(cur) = difference from current version, (prev) = difference from preceding version, m = minor edit, → = section edit, ← = automatic edit summary

(newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- [\(cur | prev\)](#)  18:08, 3 January 2016 [Nameless23](#) ([talk](#) | [contribs](#)) [m](#) . . (30,404 bytes) (0) . . [\(formatting\)](#) ([undo](#))
- [\(cur | prev\)](#)  21:28, 23 December 2015 [Jesin](#) ([talk](#) | [contribs](#)) [m](#) . . (30,404 bytes) (-3) . . [\(→See also: fixed unnecessary redirects\)](#) ([undo](#))
- [\(cur | prev\)](#)  09:48, 12 November 2015 [Tedickey](#) ([talk](#) | [contribs](#)) [m](#) . . (30,407 bytes) (-112) . . [\(Reverted 6 edits by 81.135.182.116 by ClueBot NG \(TW\)\)](#) ([undo](#))
- [\(cur | prev\)](#)  09:45, 12 November 2015 [81.135.182.116](#) ([talk](#)) . . (30,519 bytes) (+24) . . ([undo](#))
- [\(cur | prev\)](#)  09:44, 12 November 2015 [81.135.182.116](#) ([talk](#)) . . (30,495 bytes) (+16) . . ([undo](#))

Why do we need version control?



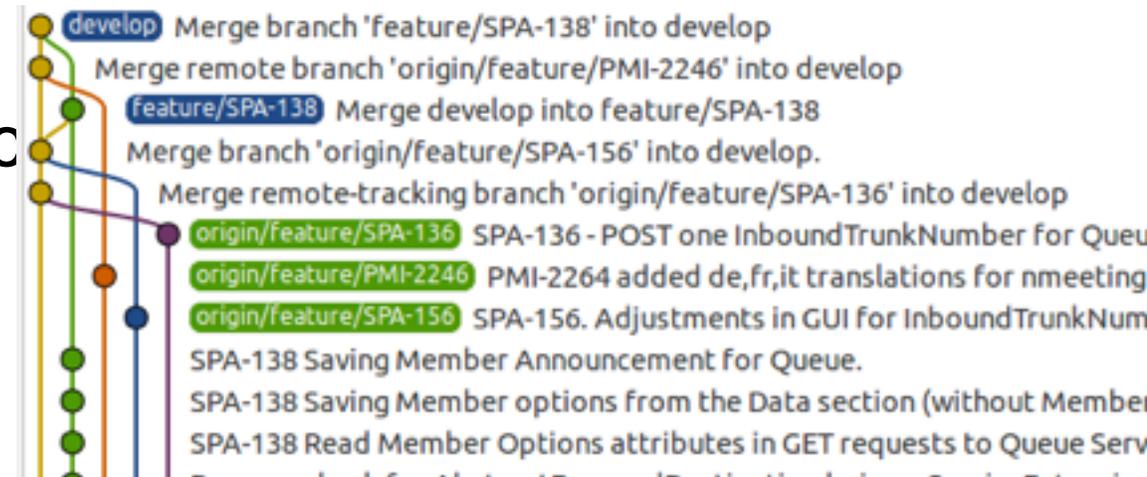
- Record the changes to the document
- Reflect on past work
- Revert to previous versions
- **Share with others**

A screenshot of the OneDrive web interface showing the "Version History" for a document. The interface includes a blue header bar with navigation icons and a red "MS" logo. Below the header, there are action buttons for "Share", "Copy link", "Download", and "Delete". The main content area is divided into two panels: a file list on the left and a "Version History" table on the right. The file list shows a folder structure with "Commons 1422-1461 Vol..." selected. The version history table lists six versions, all created by "Martin Steer", with dates ranging from April 15 to May 2 and sizes between 76 KB and 104 KB.

Versi...	Modified Date	Modified By	Size
6.0	May 2	Martin Steer	104 KB
5.0	May 1	Martin Steer	100 KB
4.0	Apr 29	Martin Steer	96 KB
3.0	Apr 17	Martin Steer	96 KB
2.0	Apr 15	Martin Steer	96 KB
1.0	Apr 15	Martin Steer	76 KB

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 actions)



Teach yourself GIT!

The Programming Historian

ABOUT ▾ CONTRIBUTE ▾ LESSONS BLOG EN ES FR



An Introduction to Version Control Using GitHub Desktop

Daniel van Strien

In this lesson you will be introduced to the basics of version control, understand why it is useful and implement basic version control for a plain text document using git and GitHub.

 Peer-reviewed  CC-BY 4.0

EDITED BY
Caleb McDaniel

REVIEWED BY
Ethan Miller
Lisa Spiro

PUBLISHED 2016-06-17

MODIFIED 2018-09-26

DIFFICULTY Low

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

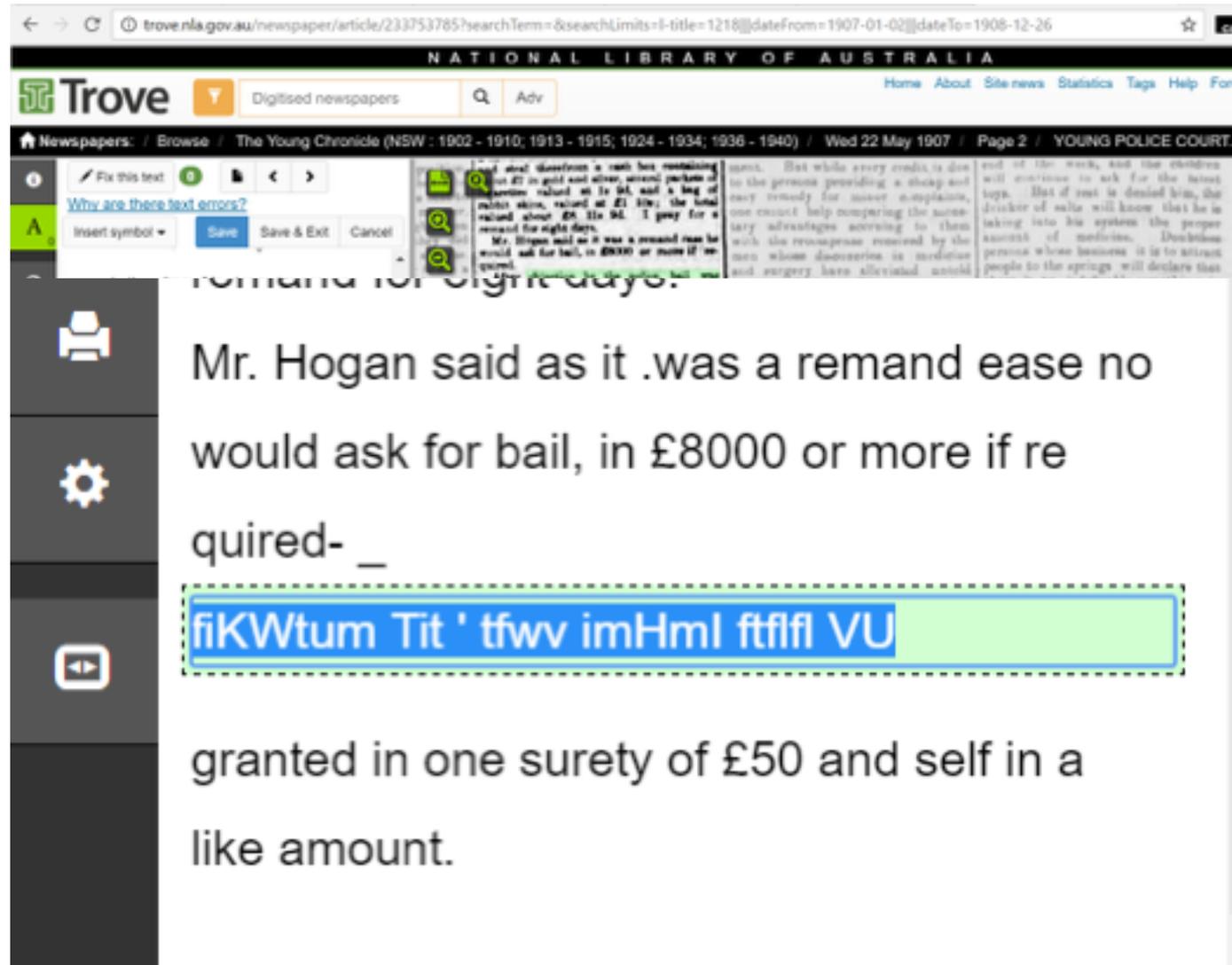
Reproducibility

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



Reproducibility

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



The screenshot shows a web browser window displaying a newspaper article from Trove. The URL is trove.nla.gov.au/newspaper/article/233753785?searchTerm=&searchLimits=f-title=1218&dateFrom=1907-01-02&dateTo=1908-12-26. The page title is "NATIONAL LIBRARY OF AUSTRALIA" and the article is from "The Young Chronicle (NSW : 1902 - 1910; 1913 - 1915; 1924 - 1934; 1936 - 1940) / Wed 22 May 1907 / Page 2 / YOUNG POLICE COURT".

The article text is partially obscured by a vertical toolbar on the left and a text correction dialog box at the top. The visible text includes:

...remains for eight days.

Mr. Hogan said as it .was a remand ease no
would ask for bail, in £8000 or more if re
quired- _

fiKWtum Tit ' tfwv imHml ftflfl VU

granted in one surety of £50 and self in a
like amount.

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

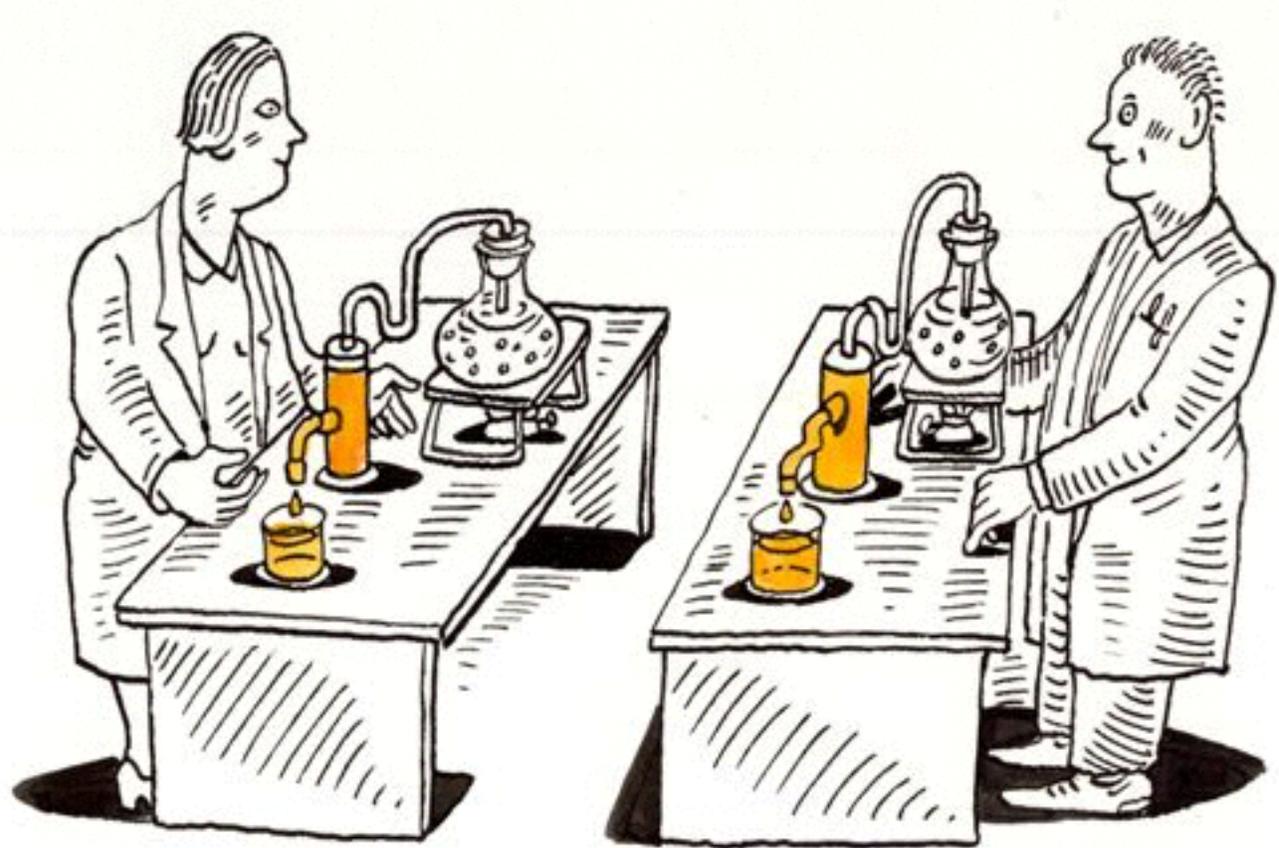
How could we verify this article about sentiment analysis?

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

 data.europa.eu



**EU Open
Data Portal**

The [European Union Open Data Portal](#) is your single point of access to open data produced by EU institutions and bodies.



**EUROPEAN
DATA PORTAL**

The [European Data Portal](#) harvests the metadata of Public Sector Information available on public data portals across European countries.



**Persistent
URIs**

[Resources with persistent URIs](#) of the EU institutions and bodies.



**EU Web
archive**

The [EU web archive](#) contains the websites of, mainly, EU institutions and agencies. Most of these sites are hosted on the europa.eu domain.

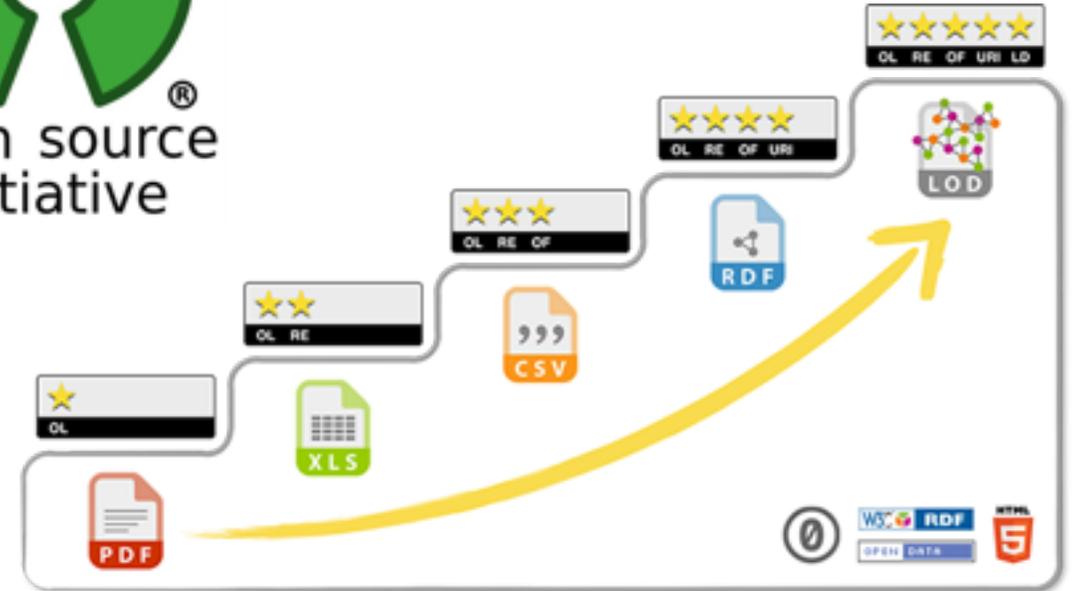
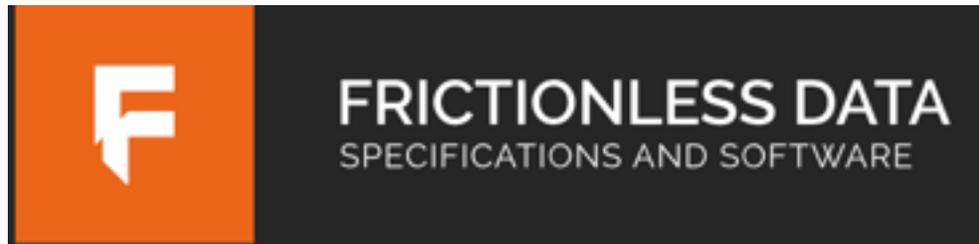
Sharing and reuse

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



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



SAS-Space



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



Welcome to SAS-Space

SAS-Space is an online library for humanities research outputs.

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



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



Welcome to SAS-Space

SAS-Space is an online library for humanities research outputs.

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 Subject

Please select a value to browse from the list below.

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)

- [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!



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



SAS-Space

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