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
Example
Gathered data
Created data
My folder system today

<table>
<thead>
<tr>
<th>Name</th>
<th>Date modified</th>
<th>Type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 General</td>
<td>30/10/2015 07:06</td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>02 Research</td>
<td>30/10/2015 07:08</td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>03 Writing</td>
<td>30/10/2015 07:05</td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>04 Presentations</td>
<td>16/07/2015 10:29</td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>05 Articles</td>
<td>30/10/2015 07:07</td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>06 Images</td>
<td>12/05/2015 09:01</td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>07 Research Plans</td>
<td>06/07/2015 07:29</td>
<td>File folder</td>
<td></td>
</tr>
</tbody>
</table>
EEBO-TCP corpus structure

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

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
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

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

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

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

Data deluge

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

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
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
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!

Information Architecture (IA)

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

The tidy text format

```r
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

Tidy a list of terms up and turn into data frame

```r
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

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
Using structures

- Hierarchies
  - Simple form of classification
  - 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
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
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
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.

Backups

• **Mistakes are guaranteed**
• **Format impermanence**
• **Inadequate storage**
• **Fixity**

"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

"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."
Backups

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

https://jupyter.org
Documentation

- Administrative
- Methodological
- Descriptive
- Technical
Documentation

- Administrative
- Methodological
- Descriptive
- Technical

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 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
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 actors)
Teach yourself GIT!

Reproducibility

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

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

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

How could we verify this article about sentiment analysis?

Reproducibility

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

Replicate vs. Reproduce findings
Scrubtable vs. Transparent methods
Reproducibility

• Open data
• Quantitative methods
• Critical inquiry
• Scientific method
• Digital publishing
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
Archive was built in 2012/13
Costs £1500/year to host
Managing legacy research data
1260 records
23GB page scans
1GB metadata

Migrate to another repository
Different data structures
Different taxonomies
Creates another ‘version’
Term mapping
Structural transform
Meaning change
Series of lessons concerning political econ.

Publisher: Imprenta del Gobierno

Place of pub: Mexico City

Pages: 682

Item ID: A00901

Folder: 1560_1995

BL catalogue no: 1560/1995

Themes: Economic Policy, Development and Ideas

Countries: MEX

Keywords: Item type: Memorandum;
Countries: Mexico;
Themes: Economic Policy, Development and Ideas;
Other keywords: Property; Industry; Labour; Capital

Project Record ID: A00901;
Folder: 1560_1995;
BL Catalogue number: 1560_1995;
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!!