GHENT
UNIVERSITY

FACULTY OF ARTS AND PHILOSOPHY

FACULTY LIBRARY OF ARTS & PHILOSOPHY

IF ONLY WE'D KNOWN: COLLECTING RESEARCH DATA

Katrien Deroo - LW Research Day

GHENT UNIVERSITY

WHAT IS DATA MANAGEMENT ABOUT?

– Spreadsheet bursting at the seams	Data model
– Database reconstruction	Documentation
– Missing USB drive	Storage
– Stuck in a research tool	Tool criticism

GHENT UNIVERSITY

RESEARCH DATA

Research data management
Managing data during your project
Data organisation
Complex data
Documentation
Storage and back up
Tools and tool criticism
Data management planning
Data sharing
Data management at Ghent University

GHENT UNIVERSITY

RESEARCH DATA TYPES

Content type	Textual, numerical, multimedia…
Data format/object	Spreadsheets/tabular data, notes, databases, marked up tekst, images, audiovisual recordings…
Mode of data collection	Experimental, observatoinal, derived/compiled data…
Primary vs secondary	Original data created in context of research project vs. reuse of existing data
Digital vs non-digital	Digital-born/digitised vs analogue data
Level of processing	Raw, processed, analysed data

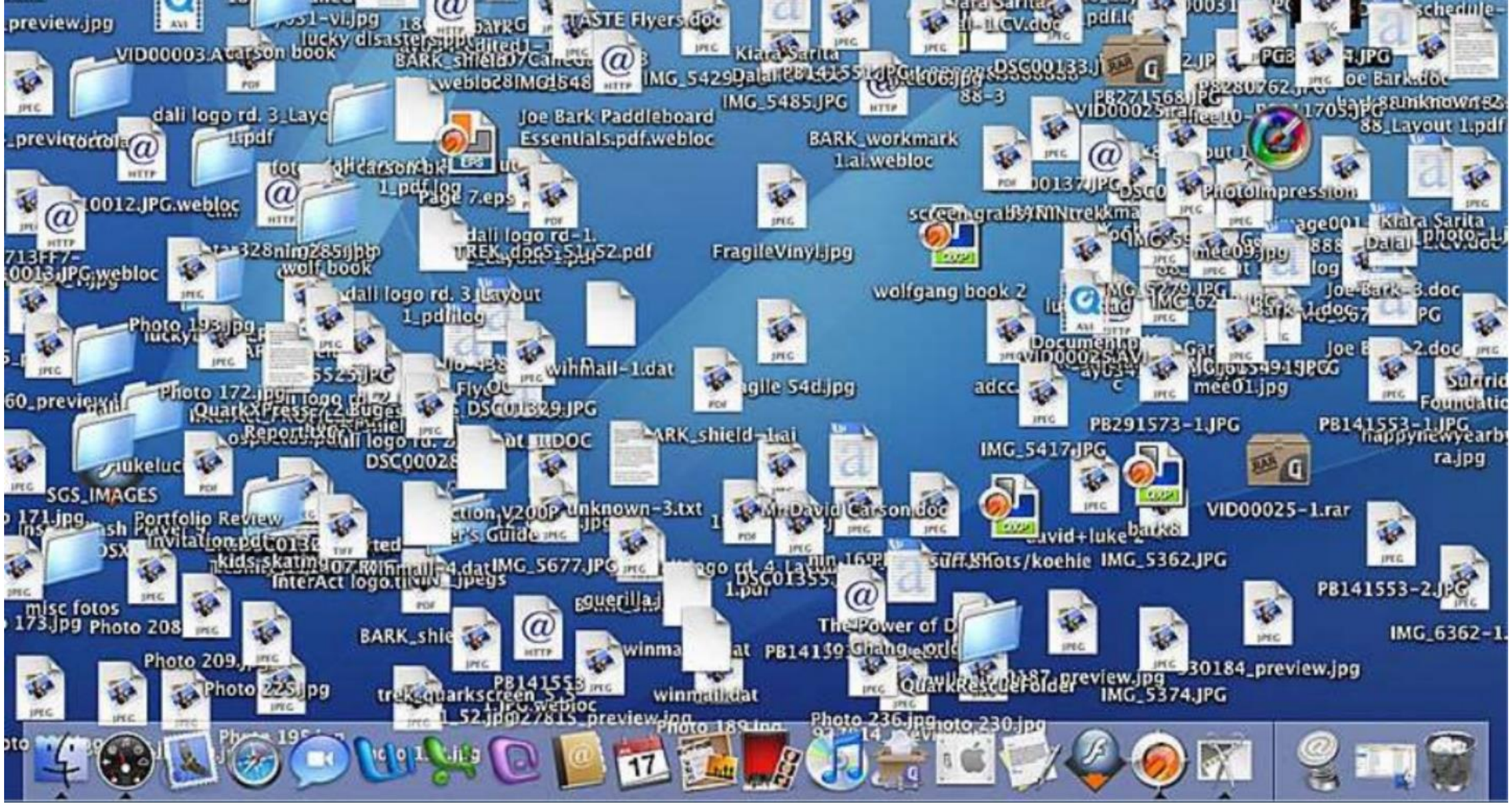
GHENT UNIVERSITY

HUMANITIES DATA

Humanities researchers increasingly create and use digital data

- Data proliferation: navigating is the challenge
- Digital data are susceptible to loss
- Dealing with digital data requires a different skill set

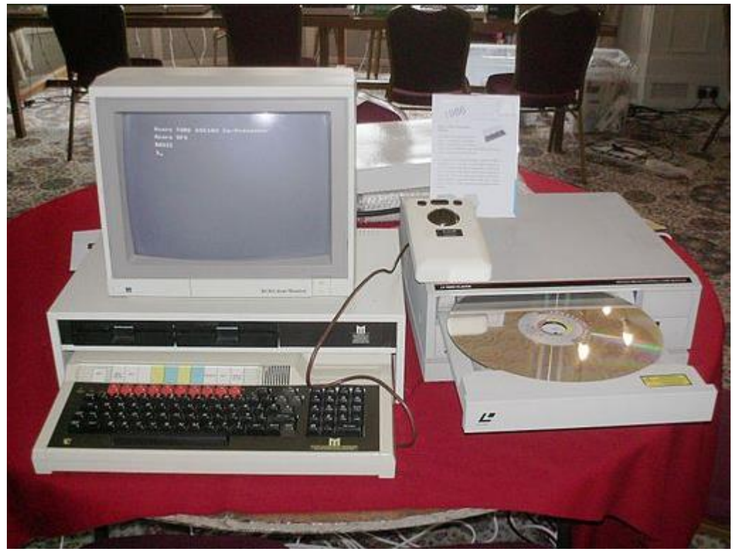
GHENT UNIVERSITY



GHENT
UNIVERSITY

DISASTROUS DATA LOSS

software/hardware failure, malicious attack, theft, natural disaster, human error…



Risks over time:

- data files no longer readable
- data no longer understandable

GHENT UNIVERSITY

Research data

RESEARCH DATA MANAGEMENT

Managing data during your project
Data organisation
Complex data
Documentation
Storage and back up
Tools and tool criticism
Data management planning
Data sharing
Data management at Ghent University

GHENT UNIVERSITY

“(...) The compilation of many small practices

that make your data

easier to understand,

less likely to be lost,

and **more likely to be useable**

during a project or ten years later.” (Briney 2015)

GHENT UNIVERSITY

DATA MANAGEMENT COVERS MANY THINGS

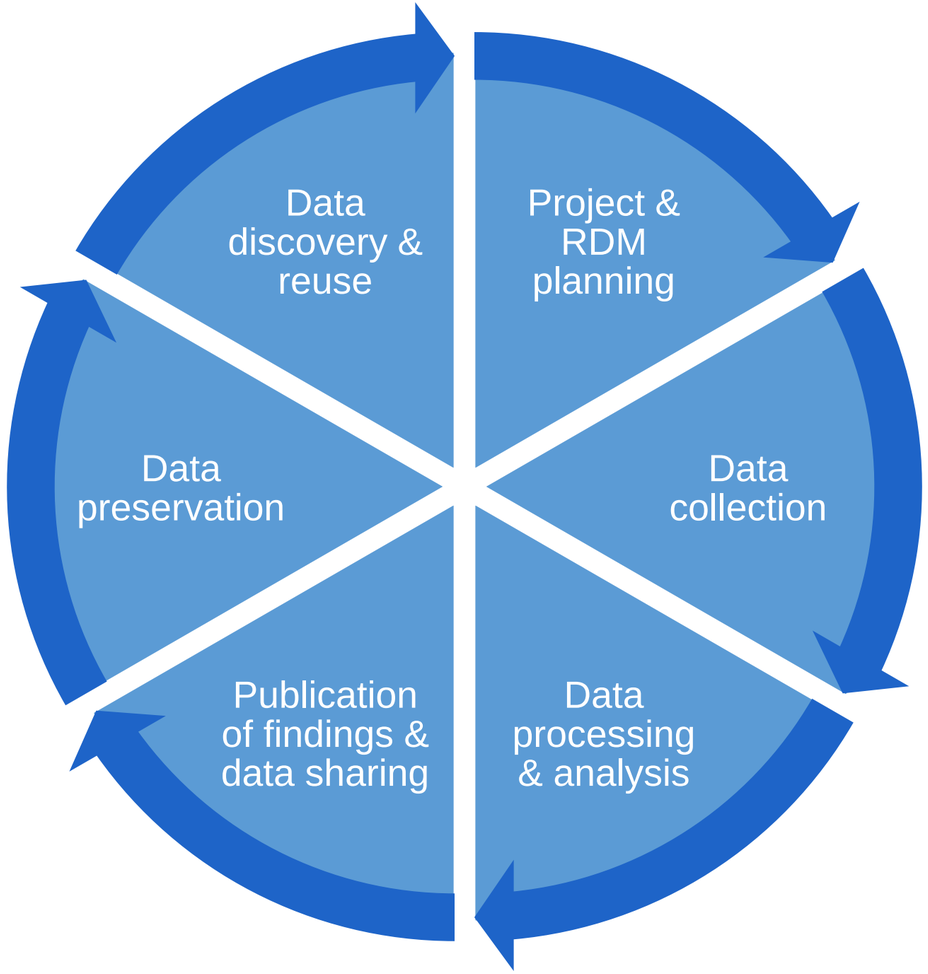
- Organising
- Storage
- Backing up
- Structuring
- Choosing technology
- Preservation
- Versioning
- Documenting
- Sharing
- Curation
- Security
- …

GHENT UNIVERSITY

TRADITIONAL RESEARCH PROCESS

GHENT UNIVERSITY

RESEARCH DATA LIFECYCLE



GHENT UNIVERSITY

Research data
Research data management

MANAGING DATA DURING YOUR PROJECT

Data organisation
Complex data
Documentation
Storage and back up
Tools and tool criticism
Data management planning
Data sharing
Data management at Ghent University

GHENT UNIVERSITY

DATA MANAGEMENT PLANNING

= translating your research questions to **pragmatic questions**

- What do I want to achieve?
- **How will I get there?**

GHENT UNIVERSITY

RESEARCH PROJECT

I want to study the concept of modernity in a corpus consisting of newspapers, literary texts and correspondence

- How will you **consult** these documents?
  Scans - pictures - pdfs - …
- How will you **process** these documents?
  Metadata - content (full text) - annotations
- How will I analyse my data?
  Close reading - NER - complex queries

GHENT UNIVERSITY

RESEARCH PROJECT

I want to study the concept of modernity in a corpus consisting of newspapers, literary texts and correspondence

- Where will the texts come from? Digital archives? Databases?
- Have they been scanned already? Are you going to scan them yourself? What scanner will you use? What resolution do you need? How will you merge all the pages into pdfs? Will your scanning software do that for you? Does this cause loss in quality?
- What information will you collect about each document? Is this information available already somewhere? Where will you keep it?
- Is the full text available? Is the full text transcription of good quality or are there many errors? Is it important that you have flawless OCR or do you need to add corrections manually? How much time do you need for correcting the OCR?
- At what level has the text been split up? (e.g. journal: article level? entire journal? do you need this information?)
- What methods do you want to use? What data format do they require? What level of enrichment is necessary? (e.g. name/place tagging, sentiment tagging, txt vs csv…)

RESEARCH PROJECT

Ideally, you have an idea of how you will tackle these questions **before you start a project**, so you can think about **requirements and timing**

GHENT UNIVERSITY

Research data
Research data management
Managing data during your project

DATA ORGANISATION

Complex data
Documentation
Storage and back up
Tools and tool criticism
Data management planning
Data sharing
Data management at Ghent University

GHENT UNIVERSITY

ORGANISATION

Archival sources

- Document **facsimile**: picture, scan (PDF/JPG), txt-file…
- Document **description**
- **Annotations**
- **Links** to other sources in your corpus

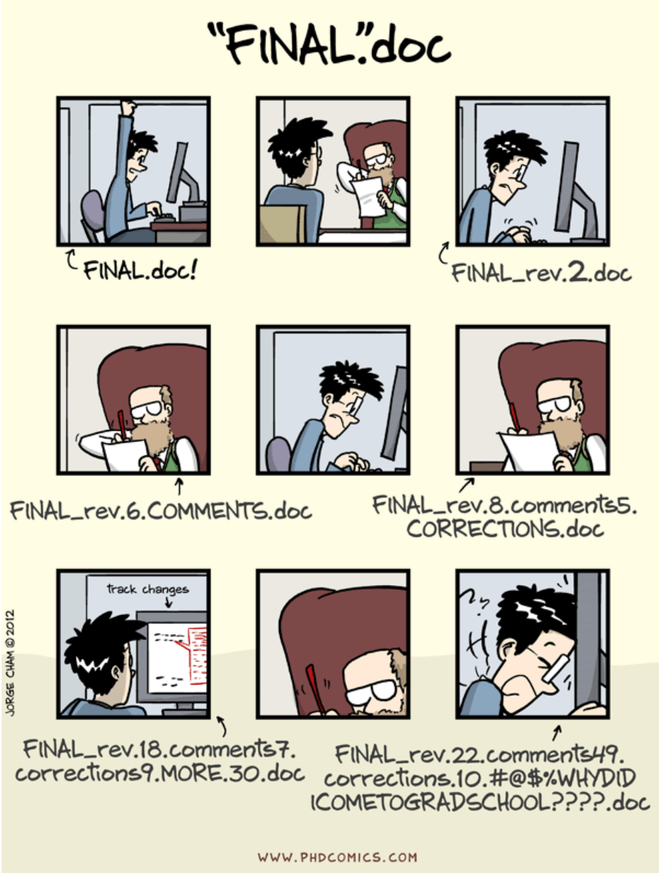
Big pile of data: how do you navigate everything efficiently?

GHENT UNIVERSITY

ORGANISATION - STRATEGIES

File & folder structure

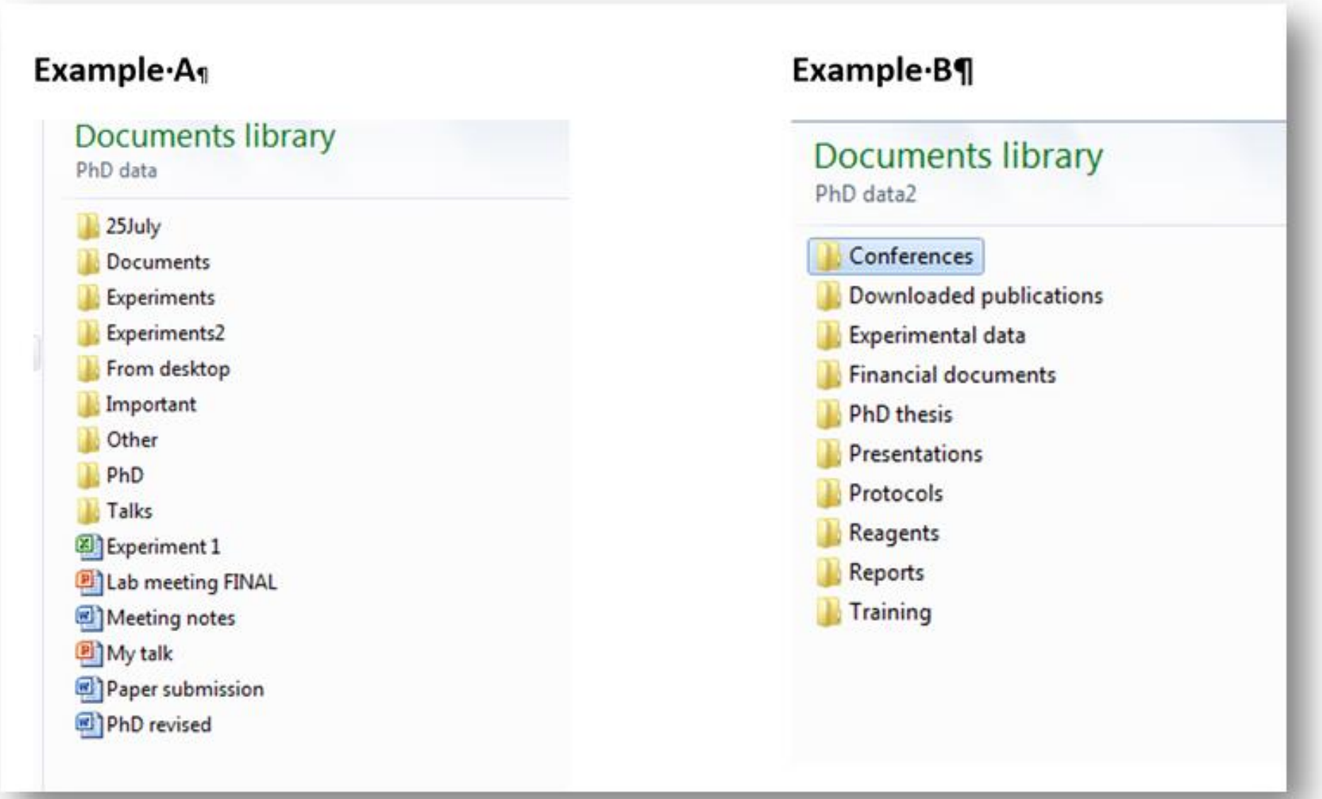
- Where do I keep everything?
- Do I know where all my files are?



Copyright Jorge Cham, PHD Comics:
http://www.phdcomics.com/comics/archive.php?comicid=1531

GHENT UNIVERSITY

ORGANISATION



From Template Research Data Management workshop for STEM researchers (https://zenodo.org/record/239090#.WMAS5RDvBKY)

GHENT UNIVERSITY

2_475854.zip
03.itsnotaboutthetoolsitsbaboutthedata....
4_475854.zip
25_476142.zip
230.txt
875_05122016.zip
ManagingResearchData_02052014.pdf
ManagingResearchData_DRAFT.txt
ManagingResearchData_FINAL.txt
RDM_Meetup_2016.md

20140502_ManagingResearchData_slides.pdf
20140502_ManagingResearchData_talk_DRAFT.txt
20140502_ManagingResearchData_talk_FINAL.txt
20150114.itsnotaboutthetoolsitsbaboutthedata.pdf
20160814_RDM_Meetup.md
20170105_ZenodoRDMTemplate.zip
20170124_RDMDoctoralSchool_Materials.zip
20170308_WorkshopDDH.zip
20170405_DSDH_notes.txt

ORGANISATION

- File structures
- File naming
- File versioning

Use a good **hierarchy**, this enables sorting, be **consistent** when placing files in folders, try to be **transparent**

https://libraries.mit.edu/data-management/files/2014/05/FileOrg_20160121.pdf

GHENT UNIVERSITY

Research data
Research data management
Managing data during your project
Data organisation

COMPLEX DATA

Documentation
Storage and back up
Tools and tool criticism
Data management planning
Data sharing
Data management at Ghent University

GHENT UNIVERSITY

COMPLEX DATA

Spreadsheets can help you keep track of documents and metadata.

+ Easy to work with (but: easy to get lost in)
+ Widely used

- No “complex queries”
- No closed terms
- No relations between entities

GHENT UNIVERSITY

COMPLEX DATA

Databases can help you keep track of documents and metadata *and* allow relations between data, as well as complex queries

+ More opportunities for querying and structuring your data (*"I want to see all **works** written between **1932 - 1940** by **author A** or **author B,** featuring **theme 1**")*
+ Forces you to create a **data model** beforehand

- Longer set up time

GHENT UNIVERSITY

COMPLEX DATA

Choosing between these data organisation systems:

- How complex is my **data structure**?
- How do I want to **query** my data?
- **How many people** will be doing data input?

But also: (how) do I want to publish the data I collect?

GHENT UNIVERSITY

COMPLEX DATA

Spreadsheet			**Database**
Author	*Born*	*Died*	*Person*
Woolf, Virginia	25/01/1882	28/03/1941	Last Name: Woolf
Woolf, Virginia (Adeline Virginia Stephen)	25/01/1882	28/03/1941	First Name: Virginia
Woolfe, Virginia	25/01/1882	28/03/1941	Born: 25/01/1882
Stephen, Virginia	25/01/1882	28/03/**1914**	Died: 28/03/1882

GHENT UNIVERSITY

COMPLEX DATA

Spreadsheet		**Database**
Text	*Themes*	*Theme tags*
Mrs. Dalloway	Mental illness (shell shock, depression, PTSD), ...	Anxiety Depression PTSD Mental illness
To the Lighthouse	Mental illness (depression, anxiety), ...	

GHENT UNIVERSITY

Research data
Research data management
Managing data during your project
Data organisation
Complex data

DOCUMENTATION

Storage and back up
Tools and tool criticism
Data management planning
Data sharing
Data management at Ghent University

GHENT UNIVERSITY

DOCUMENTATION

- All the information you need to pick up where you left off, **no matter how long it's been**.
- All the implicit and explicit information **someone else** (a new colleague, your supervisor, someone who will use your data) needs to explore your data

GHENT UNIVERSITY

DOCUMENTATION

Types of documentation:

- **Status** (“Did I correct the OCR-transcription of this document already?”)
- **Decisions** (“When transcribing egodocuments, spelling errors made by the author will be annotated with the following code: “)
- **Issues** (“This name might be a pseudonym for person X”)

GHENT UNIVERSITY

DOCUMENTATION

- Imagine someone else will be collaborating with you: what do they need to know?
- **‘Implicit’** knowledge is often forgotten
- Write down decisions you make when encountering **exceptions** / unusual use cases
- When done consistently, **immense timesaver**

GHENT UNIVERSITY

DOCUMENTATION

Documentation

- Dates
- Frequently Asked Questions
- Headwords
- How to View Syriac Text on Syriaca.org
- Language and Script Identifiers
- Place Types
- Relations
- Religious Communities
- Technical Terminology
- TEI Tag usage in Gazetteer
- Syriaca.org API Documentation
- URI Policies
- Using TEI to Catalog Syriac Manuscripts

http://syriaca.org/exist/apps/srophe/documentation/index.html

GHENT UNIVERSITY

METADATA

= form of documentation

- **Rich description** of what every file in your project contains
- Very useful for **navigating** your data, easily **queryable** because the data are **structured**
- **What information do I need to have about each data file in my collection?**

GHENT UNIVERSITY

METADATA STANDARDS

e.g. TEI, Dublin core…

- **Why use standards?**
- A lot of thought was put into these models (e.g. http://jtei.revues.org/1433)
- **Tested** on a lot of different use cases: scenarios for exceptions
- Your data set can be **easily processed** by others
- **User community** to help you with possible issues

http://www.dcc.ac.uk/resources/metadata-standards/list
http://rd-alliance.github.io/metadata-directory/

GHENT UNIVERSITY

METADATA STANDARDS

Basic goals for a metadata format

Physical format

- No limits on record or field size
- Support part/whole relationships
- Fully extensible; can add new elements as needed
- Support linking between entities (e.g. FRBR work/work relationships)
- Enable Unicode everywhere
- Support versioning
- Format definition is in a standard machine-actionable encoding

Metadata definitions

- Each data point (piece of information) exists only once in each description
- All entity descriptions are coded identically wherever they appear
- Lists of values maintained outside the format standard
- Support internationalization of input and output displays
- Feasible integration of local elements and values
- Coding fully defines data elements, not order
- No ambiguous ("X or Y") elements
- Display is data-driven where possible

https://github.com/kcoyle/MARC21/blob/master/goals.md

GHENT UNIVERSITY

Research data
Research data management
Managing data during your project
Data organisation
Complex data
Documentation

STORAGE AND BACK UP

Tools and tool criticism
Data management planning
Data sharing
Data management at Ghent University

GHENT UNIVERSITY

STORAGE AND BACK UP (DURING RESEARCH)

How will active data be stored & backed up in the short term?

- how many copies?
- storage media (e.g. hard drive, university network drive, cloud...) & locations (local/offsite)
- backup strategy (e.g. what, who, how often, full/incremental, automatic?)

GHENT UNIVERSITY

STORAGE AND BACK UP (DURING RESEARCH)

How will data be kept secure?

- security risks (e.g. in terms of unauthorised access, editing, destruction...)
- security measures (physical/network/computer system & file security)

GHENT UNIVERSITY

STORAGE AND BACK UP (DURING RESEARCH)

- Use Ghent University network drives (shares and H-drive) whenever possible
- Don’t rely on cloud storage or external hard drives only

Talk to the **faculty IT department** about your storage needs!

https://www.ugent.be/intranet/nl/op-het-werk/ict/informatieveiligheid/overzicht.htm

GHENT UNIVERSITY

STORAGE DEVICE LIFESPANS

Media	Estimated Lifespan
Magnetic data (tapes)	Up to 10 years
Nintendo cartridge	10-20 years
Floppy disk	10-20 years
CDs and DVDs	5-10 unrecorded, 2-5 recorded
Blu-Ray	Not certain, probably over 2-5 recorded
M-Disc	1,000 years (theoretically)
Hard disk	3-5 years
Flash storage	5-10 years or more (depends on write cycles)

Managing and maintaining storage devices is a lot of work

https://www.storagecraft.com/blog/data-storage-lifespan/

GHENT UNIVERSITY

Research data
Research data management
Managing data during your project
Data organisation
Complex data
Documentation
Storage and back up

TOOLS AND TOOL CRITICISM

Data management planning
Data sharing
Data management at Ghent University

GHENT UNIVERSITY

TOOLS AND TOOL CRITICISM

This Ancient Laptop Is The Only Key To The Most Valuable Supercars On The Planet

Máté Petrány
4/28/16 3:30pm · Filed to: MCLAREN F1

598 66

Photo: Patrick Gosling

http://jalopnik.com/this-ancient-laptop-is-the-only-key-to-the-most-valuabl-1773662267

GHENT UNIVERSITY

TOOLS AND TOOL CRITICISM

- **Proprietary** file formats: to be used in certain (commercial) software only, e.g. xlsx, fmp12...
- **Open** file formats: readable by any program, e.g. txt, json, csv...

Proprietary formats can impose risks on the long-term accessibility of your data!

TOOLS AND TOOL CRITICISM

- How does this tool influence my **data structure**?
  - Build a data model **before** using the tool
- How does this tool influence my **methodology**?
  - “Black box” effect
- If I put my data in the tool, **how will it come out**?
  - Proprietary file? Conversion to open formats?
- If I put my data into a tool, do I need a **front end** (e.g. for a website)?

GHENT UNIVERSITY

Research data
Research data management
Managing data during your project
Data organisation
Complex data
Documentation
Storage and back up
Tools and tool criticism

DATA MANAGEMENT PLANNING

Data sharing
Data management at Ghent University

GHENT UNIVERSITY

WHY PLAN?

Planning takes time upfront, but…

- saves a lot of time and problems later on
- helps you consider range of RDM issues involved
- makes expectations, tasks & procedures explicit
- leads to more informed decisions about data

GHENT UNIVERSITY

HOW?

Write a **Data Management Plan (DMP)**

"*(...) plans typically state what data will be created and how, and outline the plans for sharing and preservation, noting what is appropriate given the nature of the data and any restrictions that may need to be applied*." (Digital Curation Centre website)

GHENT UNIVERSITY

DATA MANAGEMENT PLAN (DMP)

- a formal document specifying how data will be handled during and after a project
- a “living” document
- increasingly required by research funders/institutions
- good practice even if not required!

GHENT UNIVERSITY

DATA MANAGEMENT PLAN (DMP)

Common topics

- Data collection & organisation
- Data documentation
- Ethical & legal issues
- Data storage & backup (during research)
- Data preservation (after research)
- Data sharing
- Responsibilities & resources

GHENT UNIVERSITY

DATA MANAGEMENT PLAN (DMP)

- use an online planning tool (dmponline.be)
- check applicable data policies
- have a look at example DMPs (https://osf.io/mcr4e/)
- keep in mind - writing a DMP is just the first step!

GHENT UNIVERSITY

Research data
Research data management
Managing data during your project
Data organisation
Complex data
Documentation
Storage and back up
Tools and tool criticism
Data management planning

DATA SHARING

Data management at Ghent University

GHENT UNIVERSITY

DATA SHARING

- Providing access to research data and documentation beyond your own team, in order to allow reuse
- Value your data as an important part of research output

GHENT UNIVERSITY

DATA SHARING

Does not necessarily mean making (all of) your data **open!**

- open = *“anyone can freely access, use, modify and share for any purpose”* (The Open Definition)
- *“as open as possible, as closed as necessary”* (European code of conduct for research integrity)

GHENT UNIVERSITY

DATA SHARING

- Possible to share data under more restricted access & use conditions, e.g.
  - only with certain (types of) users (registered, approved...)
  - only for certain types of use
  - only through secure access mechanisms
  - only after an embargo period

GHENT UNIVERSITY

Research data
Research data management
Managing data during your project
Data organisation
Complex data
Documentation
Storage and back up
Tools and tool criticism
Data management planning

DATA MANAGEMENT AT GHENT UNIVERSITY

GHENT UNIVERSITY

GHENT UNIVERSITY

- Central Research Department: **Data Management working group**
- University Library: **research data officer (Myriam Mertens) + DMPOnline.be**
- Central IT Department: **information security policy**
- Data management policy (june 2016)

GHENT UNIVERSITY

FACULTY LIBRARY

- One on one advice for your research project:
  - Building a data model
  - Writing a data management plan
- Information sessions and training
- Faculty guidelines for dealing with data

E-mail me: katrien.deroo@ugent.be

GHENT UNIVERSITY

FACULTY OF ARTS
AND PHILOSOPHY

Katrien Deroo

Data management support

FACULTY LIBRARY OF ARTS AND PHILOSOPHY

E katrien.deroo@ugent.be
T +32 9 331 33 88

GHENT
UNIVERSITY