Data Information Literacy

GEO 802 Fall 2021

Gary Seitz, MA

Anna C. Véron, Dr. sc. nat.

Course Schedule

Day 1		
1. Introduction	Anna	09:00-10:00
Break		
2. Discovery & Acquisition	Gary	10:15-11:30
Lunch Break		
3. Creating Data	Anna	13:00-13:45
4. Organizing Data	Gary	13:45-14:45
Break		
5. Data Types & Formats	Gary	15:00-16:00

Course Schedule

Day 2		
6. Data Documentation & Metadata	Anna	09:00-09:45
7. Storage, Backup, Security & Preservation	Anna	09:45-10:30
Break		
8. Data Sharing, Reusing & Citation	Gary	10:45-11:45
Lunch Break		
9. Ethics & Copyright	Gary	13:00-13:45
10. Data Management Planning	Anna	13:45:14:45
Break		
Exercise		15:00-16:00

Your course goals

- **You'll be able to apply efficient research data management techniques during your Master / PhD research project (and during your further career).**
- We'll give you a «buffet» of knowledge and tools for data management.

- Only you can decide and pick what you need for your research!

To be handed in

- Solved Exercises
- **Your own DMP** (for your research project)
- Create a folder structure and file naming convention and upload your files (exercises and DMP) to SWITCHdrive.

Lesson 1: Introduction

→ **What is Research Data?**

- **The Importance of Data Management**
- **The Data Lifecycle**

Why Data Management?

Data Loss

https://www.flickr.com/photos/quinnanya/3239528185/in/gallery-wlef70-72157633022909105/

Blick am Abend, 25.10.2018

Exercise 1.2: Reproducibility crisis

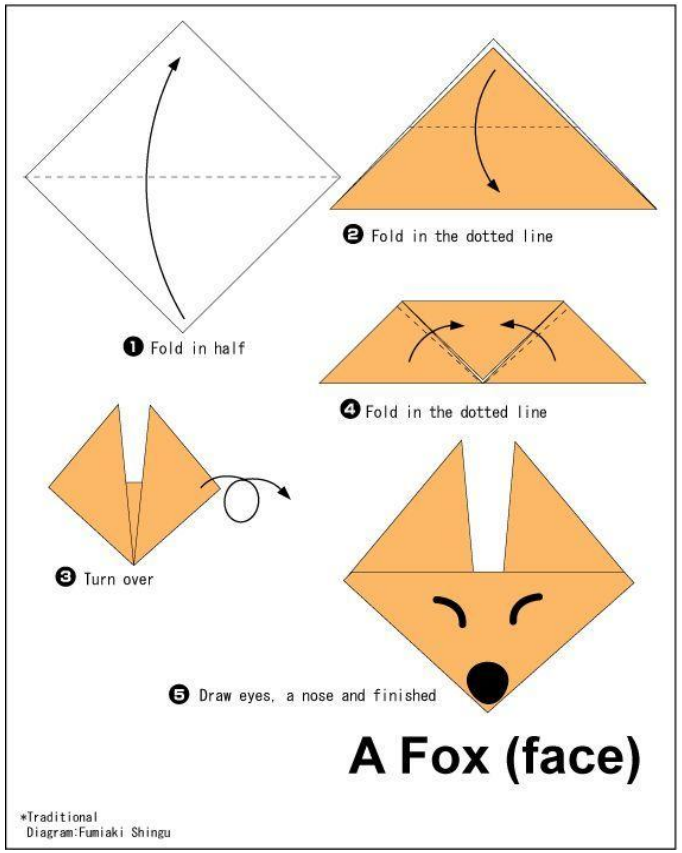
Please follow the instructions

1. Put the paper with the blank side up in front of you
2. Fold in half
3. Fold in dotted line
4. Fold the outer edges up
5. Turn over

Exercise 1.2: Reproducibility crisis

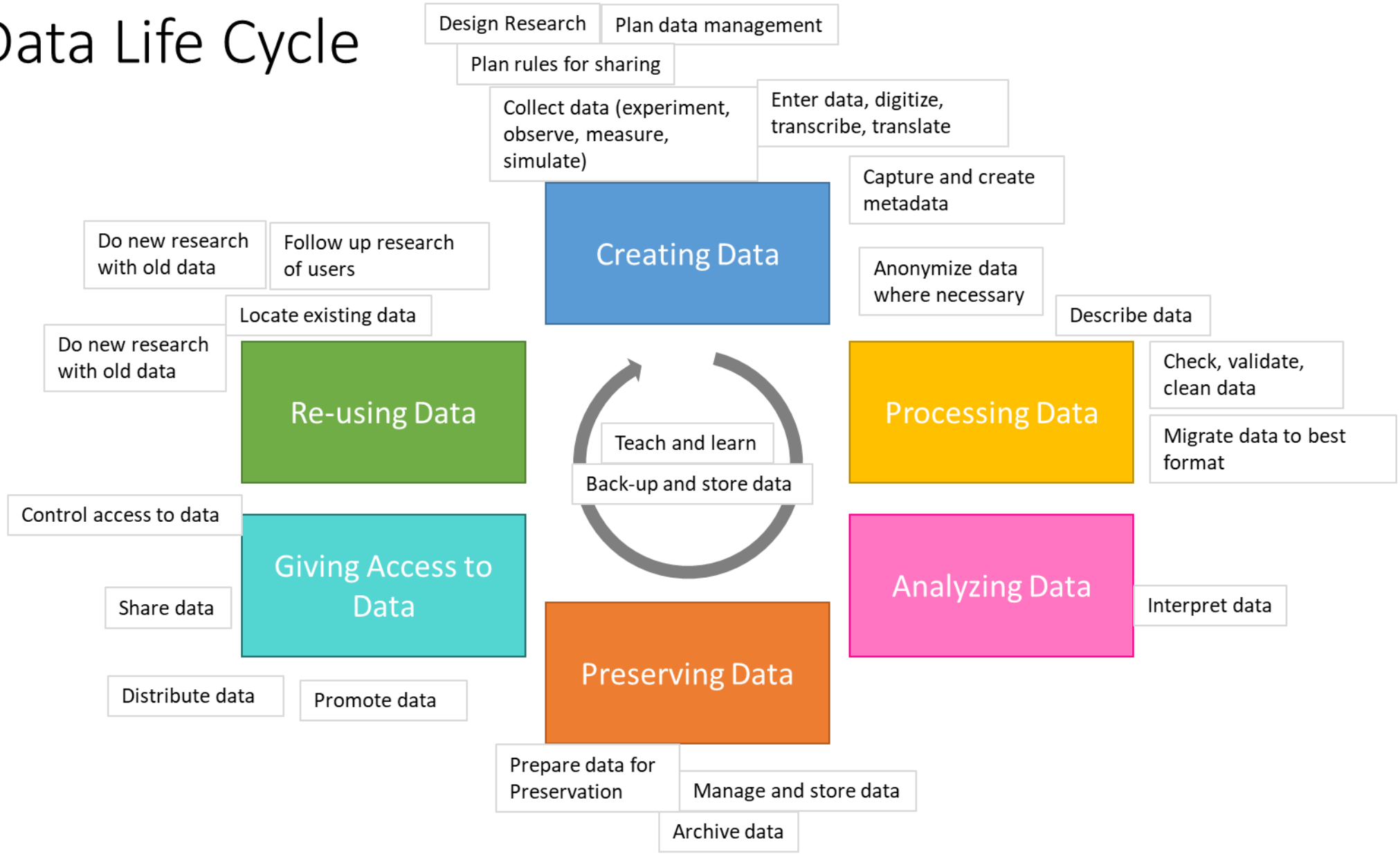
Please follow the instructions

1. Put the paper with the blank side up in front of you
2. Fold in half
3. Fold in dotted line
4. Fold the outer edges up
5. Turn over



Incomplete instructions are not reproducible!

Data Life Cycle



Summary of Lesson 1

Data deluge: **The amount of data created every year is increasing exponentially**

Improper data management can be **costly**

Data management allows you to **find, access, understand, integrate and re-use data.**

If data are:
- ✓ Well-organized
- ✓ Documented
- ✓ Preserved
- ✓ Accessible
- ✓ Verified to accuracy and validity

The benefits are:
- ✓ High quality data
- ✓ Easy to share and re-use
- ✓ Citation & credibility for the researcher
- ✓ Saving costs

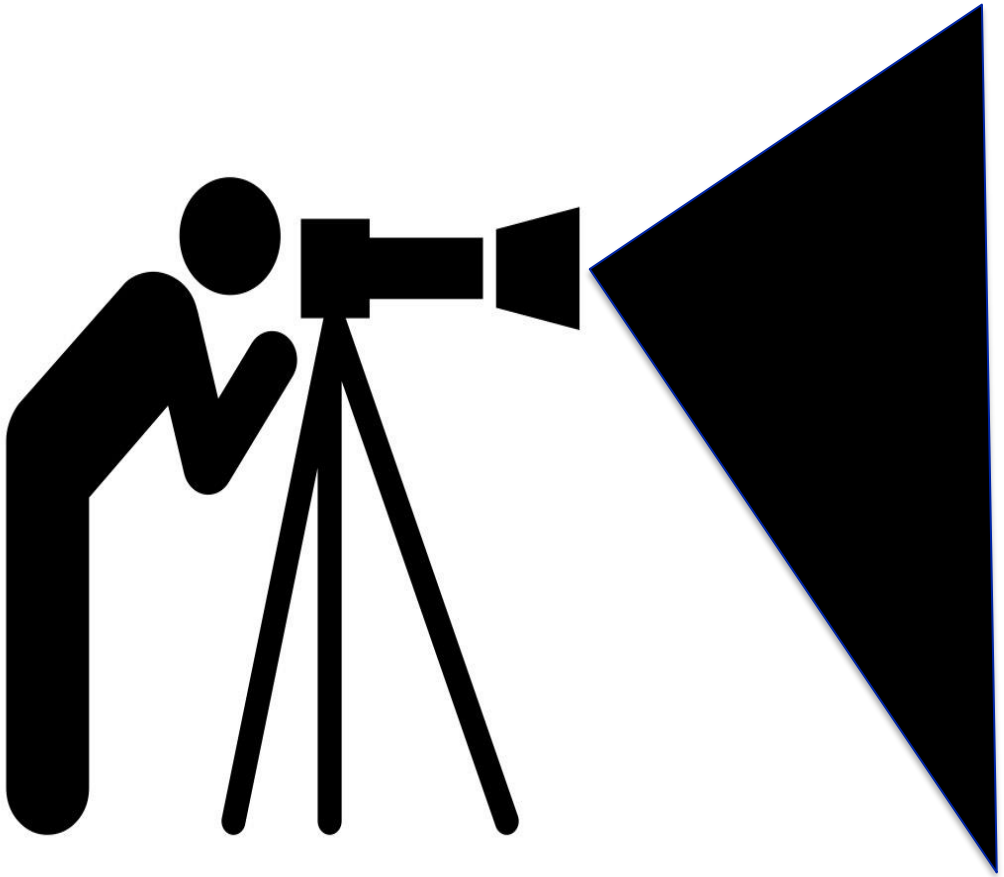
Discovery and Acquisition of Data

GEO 802, Data Information Literacy

Fall 2021 – Lecture 2

Gary Seitz, MA

Lesson 2 Outline



Luis Prado from The Noun Project

Data Repositories

Discipline-related repositories

Portals for data publication

Open Data from organizations

Re3data:Registry of Research Data Repositories

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Home | Search | Browse | Suggest | FAQ | About | Schema | API | Contact | Legal notice / Impressum

Filter

- Subjects
- Content Types
- Countries
- AID systems
- API
- Certificates
- Data access
- Data access restrictions
- Database access
- Database access restrictions
- Database licenses
- Data licenses
- Data upload
- Data upload restrictions
- Enhanced publication
- Institution responsibility type
- Institution type

Search... Search

Toogle short help

← Previous 1 2 3 4 5 6 7 ... 67 Next →

Sort by

Found 1674 result(s)

ICSU World Data System

International Council for Science World Data System

Subject(s): Humanities and Social Sciences | Life Sciences | Natural Sciences | Engineering Sciences

Content type(s): Standard office documents | Images | Scientific and statistical data formats | Raw data | Plain text | Archived data | Structured text

Country: Japan | International

The Prototype Data Portal allows to retrieve Data from World Data System (WDS) members. WDS ensures the long-term stewardship and provision of quality-assessed data and data services to the international science community and other stakeholders

Exercise 2.1

Check Registry of Research Data Repositories

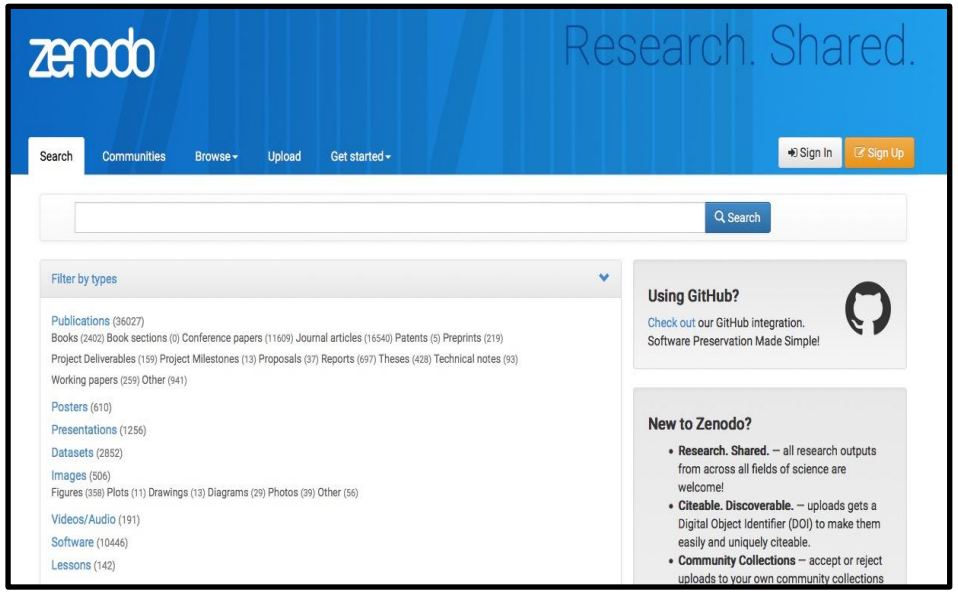
www.re3data.org:

– Can you find data repositories in your field?

– List 3 data repositories, where you think you could find data for your thesis.

Data repositories

Zenodo



- A research data repository. It was created by OpenAIRE and CERN to provide a place for researchers to deposit datasets
- Has integration with GitHub to make code hosted in GitHub citable
- Provides secure archiving and referability, including digital object identifiers (DOIs)
- Easy access
- Disadvantage: No curation, no quality control

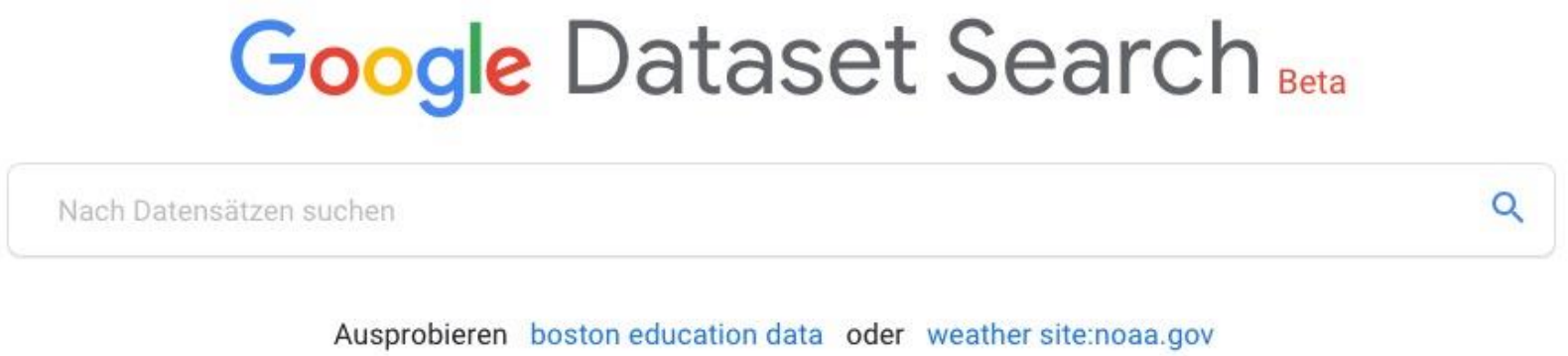
Discipline-related repositories

– GIS and Geography

- GeoCommons.com GIS file repository and finding tool
- Federal Geographic Data Committee - Provides access to the National Spatial Data Infrastructure (NSDI) Clearing House Network and the geodata.gov portal
- https://inspire-geoportal.ec.europa.eu/: The INSPIRE Geoportal is the central European access point to the data provided by EU Member States and several EFTA countries under the INSPIRE Directive.
- Geoportal, Geodaten aus Deutschland https://www.geoportal.de/
- Geodatenkatalog : https://wiki.gdi-de.org/display/gdk
- OpenTopography: OpenTopography facilitates community access to high-resolution, Earth science-oriented, topography data, and related tools and resources

Data Search Engine

– Google Dataset Search



How well does the Google Search work, after your knowledge and experiences with the data repositories you have looked at?

Data papers & data journals

Earth System Science Data

Earth System Science Data
The Data Publishing Journal

| Contact |

Home
Online Library ESSD
- Recent Final Revised Papers
- Volumes and Issues
- Special Issues
- Full Text Search
- Title and Author Search

Online Library ESSDD
Alerts & RSS Feeds
General Information
Submission
Review
Production
Subscription
Comment on a Paper

Earth Syst. Sci. Data, 4, 47-73, 2012
www.earth-syst-sci-data.net/4/47/2012/
doi:10.5194/essd-4-47-2012

Article | Related Articles

Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates

Y.-W. Luo[1], S. C. Doney[1], L. A. Anderson[2], M. Benavides[3], I. Berman-Frank[4], A. Bode[5], S. Bonnet[6], K. H. Boström[7], D. Böttjer[8], D. G. Capone[9], E. J. Carpenter[10], Y. L. Chen[11], M. J. Church[8], J. E. Dore[12], L. I. Falcón[13], A. Fernández[14], R. A. Foster[15], K. Furuya[16], F. Gómez[17], K. Gundersen[18], A. M. Hynes[19,*], D. M. Karl[8], S. Kitajima[16], R. J. Langlois[20], J. LaRoche[20], R. M. Letelier[21], E. Marañón[14], D. J. McGillicuddy Jr.[2], P. H. Moisander[22,], C. M. Moore[23], B. Mouriño-Carballido[14], M. R. Mulholland[24], J. A. Needoba[25], K. M. Orcutt[18], A. J. Poulton[26], E. Rahav[4], P. Raimbault[6], A. P. Rees[27], L. Riemann[28], T. Shiozaki[16], A. Subramaniam[29], T. Tyrrell[23], K. A. Turk-Kubo[22], M. Varela[5], T. A. Villareal[30], E. A. Webb[9], A. E. White[21], J. Wu[31], and J. P. Zehr[22]**

[1]Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA
[2]Department of Applied Ocean Science and Engineering, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA
[3]Instituto de Oceanografía y Cambio Global, Universidad de Las Palmas de Gran Canaria, 35017, Las Palmas de Gran Canaria, Spain
[4]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 52900, Israel
[5]Instituto Español de Oceanografía, Centro Oceanográfico de A Coruña, 15080 A Coruña, Spain
[6]IRD-INSU-CNRS, Laboratoire d'Océanographie Physique et Biogéochimique, UMR 6535, Centre d'Océanologie de Marseille, Aix Marseille Université, France
[7]Department of Natural Sciences, Linnaeus University, 39182 Kalmar, Sweden
[8]School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, Hawaii 96822, USA
[9]Department of Biological Sciences and Wrigley Institute for Environmental Studies, University of Southern California, Los Angeles, California 90089, USA
[10]Romberg Tiburon Center, San Francisco State University, Tiburon, California 94920, USA
[11]Department of Marine Biotechnology and Resources, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
[12]Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, MT 59717, USA
[13]Laboratorio de Ecología Bacteriana, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico

Search ESSD
Search
Full Text

Special Issue
MAREDAT - Towards a world atlas of marine plankton functi...

Final Revised Paper
Supplement (79 KB)

Citation
BibTeX
EndNote

Discussion Paper
Published on 2012-02-13

Share

Creating Data

GEO 802 Fall 2021, Data Information Literacy

Anna C. Véron, Dr. sc. nat.

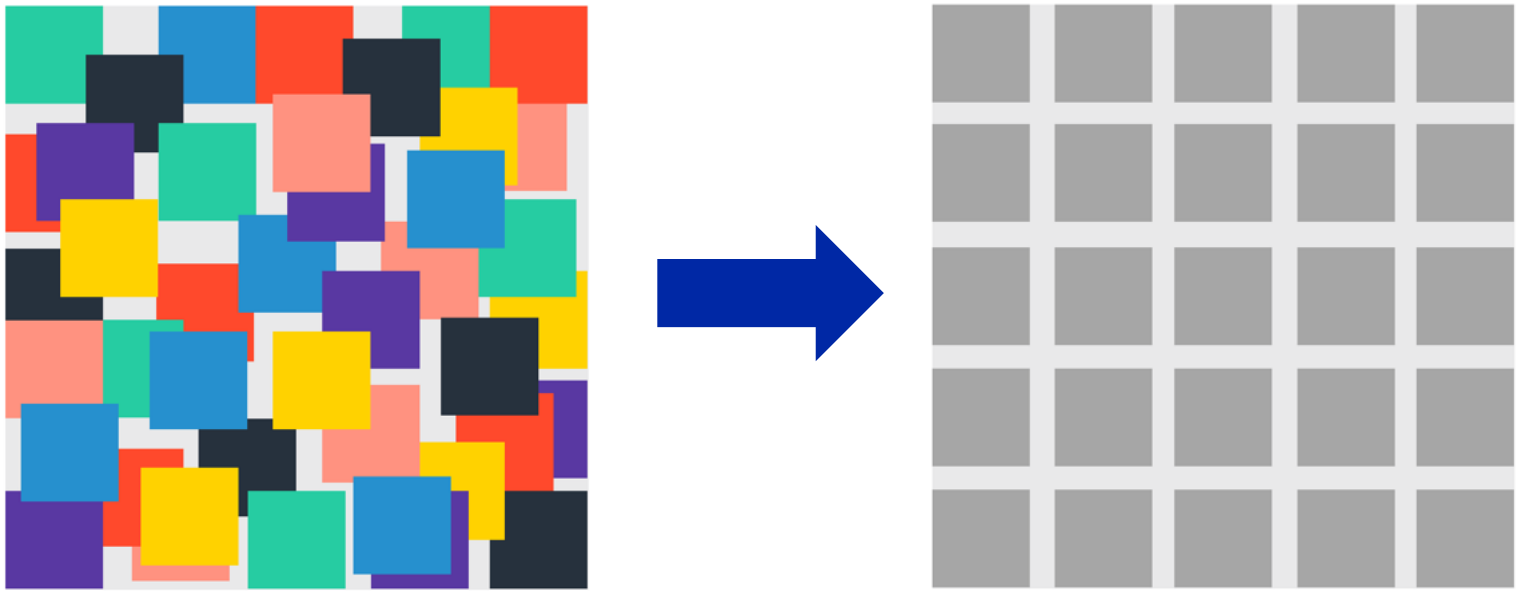
Lesson 3: Data Entry and Manipulation

→ **How to structure your data: Best practices**

- **Quality of research data**
- **Data entry tools**
- **Databases**
- **Data Analysis**

Structured vs. Unstructured data

Whenever possible, create structured data!



Devin Pickell, G2 Learning Hub, Structured vs. Unstructured Data – What's the Difference?
https://learn.g2.com/structured-vs-unstructured-data; accessed Aug 26th 2020

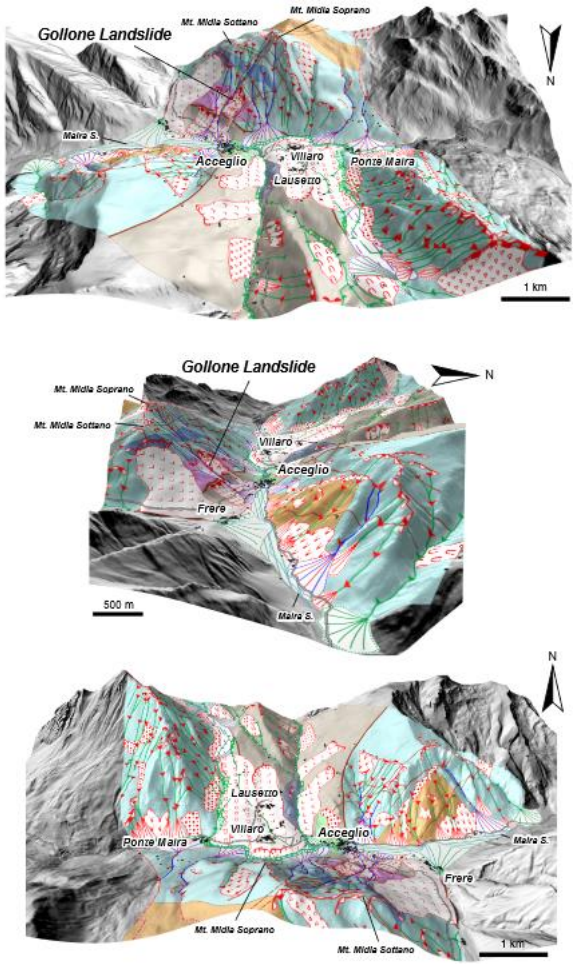
Structured vs. Unstructured data

Unstructured data

- No pre-defined data model
- Difficult to search
- Not «machine-readable», but can be analyzed with text mining, data mining and AI techniques (time-consuming)
- More than 80% of data generated in the world

Sources of Unstructured Data:

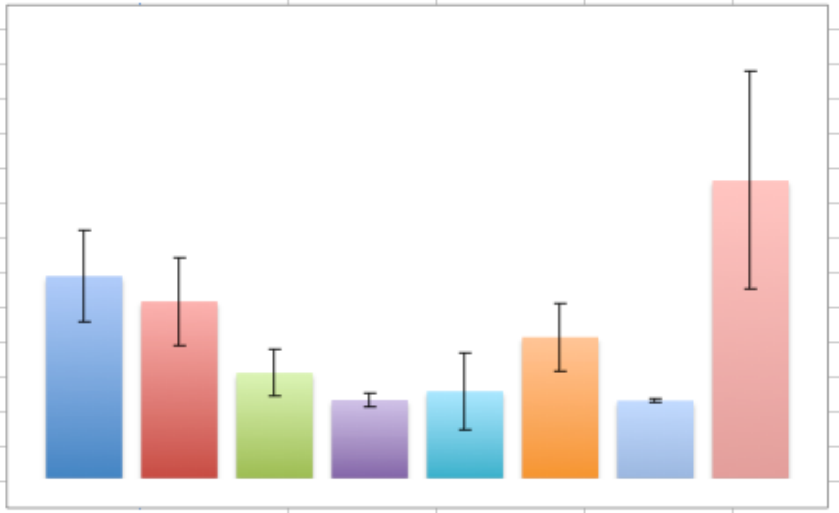
- Text files, presentations, emails, websites, diaries
- Social media, text messages, chat
- image, audio and video files
- **Examples from Science:** satellite imagery, microscope images, space exploration, seismic imagery, atmospheric data, surveillance photos / videos, sensor data



Petroccia, A.. Structural and geomorphological framework of the upper Maira Valley (Western Alps, Italy): the case study of the Gollone Landslide (2020). https://doi.org/10.6084/m9.figshare.12854354.v1

Exercise 3.1: Spot the six problems

	A	B	C	D	E	F	G	H
1	**Mutant**							
2						**average**	**stddev**	
3	Sic18	4059	6415	5938		5471	1246	23%
4	1-273	5004	3486	5870		4787	1207	25%
5	1-225	3212	3218	2129		2853	627	22%
6	210-264	2091	2317	1947		2118	187	9%
7	215-264	1141	3053	2873		2356	1056	45%
8	221-264	3626	3006	4824		3819	924	24%
9	226-264	2038	2090	2176		2101	70	3%
10	Sic18	6947	5823	11405		8058	2952	37%
11								
12						**average**	**stddev**	
13	Sic18	4059	6255	5561		5292	1123	21%
14	1-273	5004	3377	5458		4952	1094	22%
15	1-225	3212	3050	1994		3683	661	18%
16								
17	210-264	2091	6415	1824		3443	2577	75%
18	215-264	1141	3463	2691		2938	1183	40%
19	221-264	3626	3128	4518		3095	704	23%
20	226-264	2038	2038	2038		2898	0	0%
21	Sic18	6947	5678	12622		5227	3698	71%
22								
23						**average**	**stddev**	
24	Sic18	4066	6248	5562		5292	1116	21%
25	1-273	5011	3372	5459		4953	1099	22%
26	1-225	3222	3044	1989		3683	666	18%
27	210-264	2099	6407	1823		3443	2571	75%
28	215-264		3457	2694		3296	540	16%
29	SIC1221-264-3XHA	3630	3123	4513		3483	703	20%
30	226-264	2042	2033	2037		2896	5	0%
31	Sic18	6951	5674			3747	903	24%
32								



Exercise 3.1: Six problems

- **Unlabelled data:** It is completely unclear what kind of data are actually recorded in this sheet. The data do not have a descriptive column header and also no unit is indicated.
- **Column headers should be in a single row.** This spreadsheet contains headers in row 1 and row 2, as well as redundant headers for two columns in rows 14 and 26 that are confusing. There should be a single unique header for each column, preferably in row 1.
- Don’t use units, even percentages together with the value in the field
- **Do not use empty rows or columns.** These may cause problems when data are exported. Empty rows and columns also tend to indicate the presence of multiple tables in one sheet, which appears to be the case with this example. A sheet should contain only one table of data.

- **Empty cells:** unclear. Was this not measured? Did the value seem unreliable so was omitted? Was this value deleted by accident? If a cell must be left empty, make a notation in a column of comments about why the cell is empty.
- Embedded Charts, graphs or images are not included when the data are exported.

Data entry tools

For Spreadsheets


MS Excel


Apple Numbers


Google Sheets


OpenOffice Calc


LibreOffice Calc


Zoho Sheets

Data entry tools

For Surveys

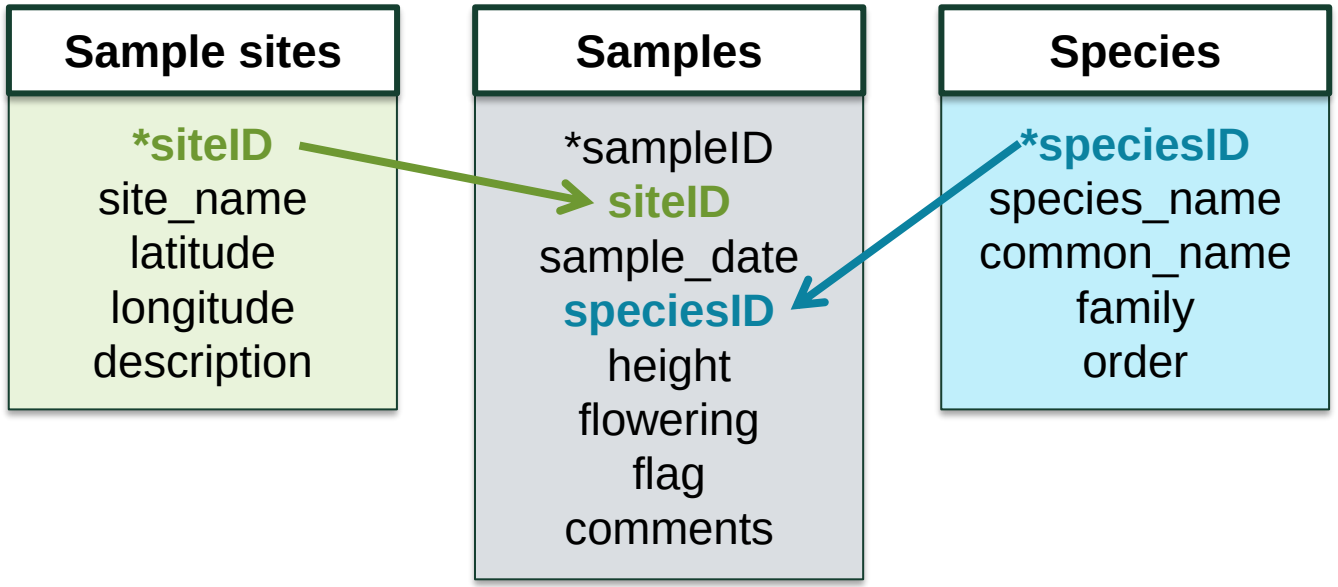
LimeSurvey

Scientific online survey tool
Campus licence available for all UZH members
https://www.uzh.ch/zi/cl/umfragen/index.php/admin/authentication/sa/login

Surveymonkey

Google Forms

What is a relational database?



- Contains more than one table
- Relationships between the tables
- Parent tables and child tables
- Are searched with a declarative programming language: **SQL = structured query language**

Spreadsheets vs. databases

Spreadsheets

- Flexible about cell content type—cells in same column can contain numbers or text
- Cells can contain calculations (functions)
- Limited number of rows
- usually not editable by multiple users at the same time
- Allow for extensive analysis

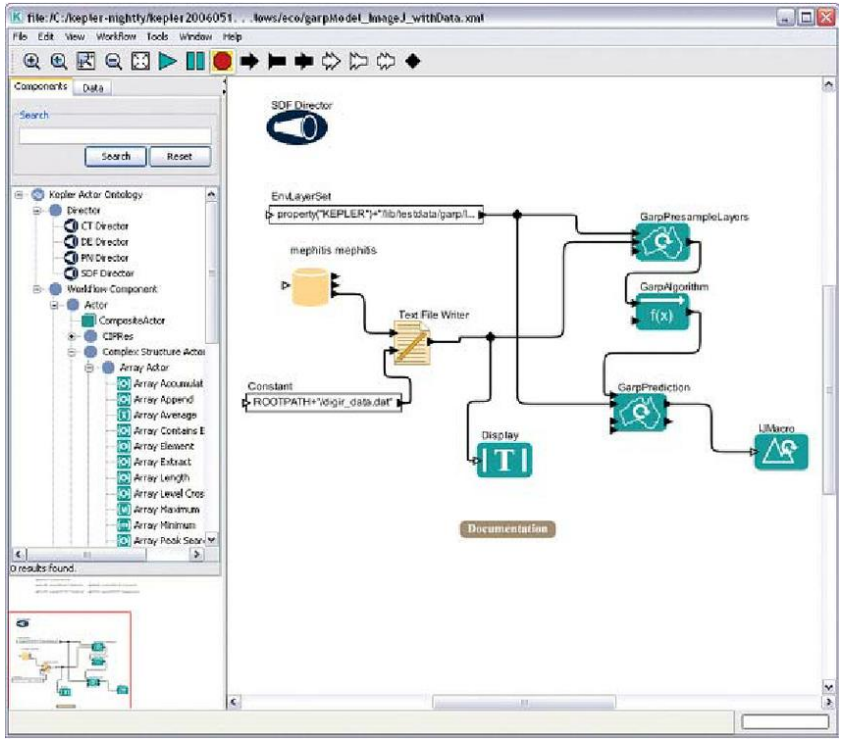
Databases

- Pre-set the type of data contained in a certain field
- Suitable for very large amounts of raw data
- Improved data integrity and consistency
- multiple users can work on it in parallel
- All calculations and operations are done after data retrieval

Tools for documenting scientific workflows

kepler-project.org/

- Open-source, free, cross-platform
- Drag-and-drop interface for workflow construction
- Possible applications
  - Theoretical models or observational analyses
  - Hierarchical modeling
  - Can have nested workflows
  - Can access data from web-based sources (e.g. databases)



Summary of Lesson 3

Create **structured** data whenever possible

Make sure your data is **consistent**, **reproducible**, **accurate** and **complete**.

When using data from different sources: Make sure your data is well **integrated**.

Choose a data entry method that allows for the **validation** of data as it is entered.

Consider investing time in learning how to use a **relational database** if datasets are large or complex.

Remember to **document** your data analysis and manipulation to ensure **reusability** and **reproducibility**.

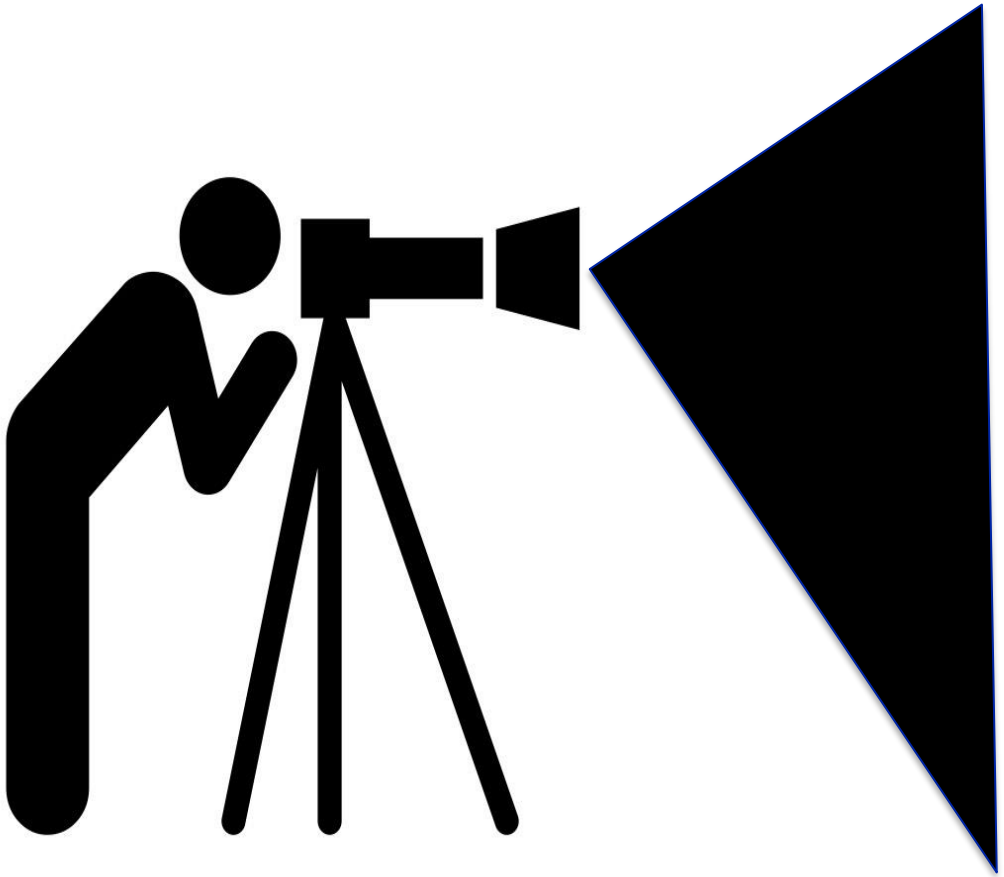
Organizing Your Data

GEO 802, Data Information Literacy

Fall 2021 – Lecture 4

Gary Seitz, MA

Lesson 4 Outline



Luis Prado from The Noun Project

- File-naming conventions
- Data organization
- Documenting your process
- Keeping a [lab] notebook


NAH, I'M NOT WORRIED ABOUT CLOUD SECURITY. MY STORED DATA IS SO DISORGANIZED THEY'D NEVER BE ABLE TO FIND ANYTHING!
© D.Fletcher for CloudTweaks.com

The benefits of consistent data file labelling

- Data files can be **sorted** in logical sequence
- Data files are not accidentally **overwritten** or deleted
- Different **versions** of data files can be identified
- If data files are moved to other storage platform their names will retain useful context

File-naming best practices

Best Practices:

1. **Be Descriptive**: 75092238.txt is not helpful. Instead: 20120814_instrument8_rainyday_raw.txt (up to 255 characters)

2. **Don’t rely on nesting in folders**: 2012/august/instrument8/day14/raw.txt

3. **Use consistent structure** that falls into a useful order (for sorting) and decide on shared terminology

4. **List versions alphanumerically**, eg. v1, v2, v3 rather than last, final, finalfinal, useTHISone

5. **Use numerical dates**, eg. YYYYMMDD rather than Dec09

6. **Use underscores** instead of full-stops or spaces because, like special characters, these are parsed differently on different systems.

7. **File names should provide context for the contents of the file**, making it distinguishable from files with similar subjects or different versions of the same file.

File-naming strategies

Order by date:

- 19550412_notes_MassObs.docx
- 19550412_questionnaire_MassObs.pdf
- 19631215_notes_Gorer.docx
- 19631215_questionnaire_Gorer.pdf

Order by type:

- Notes_Gorer_19631215.docx
- Notes_MassObs_19550412.docx
- Questionnaire_Gorer_19631215.pdf
- Questionnaire_MassObs_19550412.pdf

Order by subject:

- Gorer_notes_19631215.docx
- Gorer_questionnaire_19631215.pdf
- MassObs_notes_19550412.docx
- MassObs_questionnaire_19550412.pdf

Forced order with numbering:

- 01_MassObs_questionnaire_19550412.pdf
- 02_MassObs_notes_19550412.docx
- 03_Gorer_questionnaire_19631215.pdf
- 04_Gorer_notes_19631215.docx

On using number order in file names…

Dates listed in order of collection

US standard	MM/DD/YY	DD/MM/YY	YYYY-MM-DD
January 12, 2011	01/12/11	12/01/11	2011-01-12
February 15, 2011	02/15/11	15/02/11	2011-02-15
March 27, 2011	03/27/11	27/03/11	2011-03-27
April 11, 2011	04/11/11	11/04/11	2011-04-11
May 20, 2011	05/20/11	20/05/11	2011-05-20
June 6, 2011	06/06/11	06/06/11	2011-06-06
July 18, 2011	07/18/11	18/07/11	2011-07-18
August 7, 2011	08/07/11	07/08/11	2011-08-07
September 9, 2011	09/09/11	09/09/11	2011-09-09
October 14, 2011	10/14/11	14/10/11	2011-10-14
November 24, 2011	11/24/11	24/11/11	2011-11-24
December 18, 2011	12/18/11	18/12/11	2011-12-18
January 19, 2012	01/19/12	19/01/12	2012-01-19
February 27, 2012	02/27/12	27/02/12	2012-02-27
March 15, 2012	03/15/12	15/03/12	2012-03-15
April 27, 2012	04/27/12	27/04/12	2012-04-27
May 11, 2012	05/11/12	11/05/12	2012-05-11
June 17, 2012	06/17/12	17/06/12	2012-06-17
July 16, 2012	07/16/12	16/07/12	2012-07-16
August 4, 2012	08/04/12	04/08/12	2012-08-04
September 14, 2012	09/14/12	14/09/12	2012-09-14
October 26, 2012	10/26/12	26/10/12	2012-10-26
November 26, 2012	11/26/12	26/11/12	2012-11-26
December 8, 2012	12/08/12	08/12/12	2012-12-08
January 22, 2013	01/22/13	22/01/13	2013-01-22
February 20, 2013	02/20/13	20/02/13	2013-02-20
March 26, 2013	03/26/13	26/03/13	2013-03-26
April 18, 2013	04/18/13	18/04/13	2013-04-18
May 23, 2013	05/23/13	23/05/13	2013-05-23

On using number order in file names…

If we sort by MM/DD/YY, dates are out of order.

US standard	MM-DD-YYYY	DD-MM-YYYY	YYYY-MM-DD
January 12, 2011	01-12-2011	12-01-2011	2011-01-12
January 19, 2012	01-19-2012	19-01-2012	2012-01-19
January 22, 2013	01-22-2013	22-01-2013	2013-01-22
February 15, 2011	02-15-2011	15-02-2011	2011-02-15
February 20, 2013	02-20-2013	20-02-2013	2013-02-20
February 27, 2012	02-27-2012	27-02-2012	2012-02-27
March 15, 2012	03-15-2012	15-03-2012	2012-03-15
March 26, 2013	03-26-2013	26-03-2013	2013-03-26
March 27, 2011	03-27-2011	27-03-2011	2011-03-27
April 11, 2011	04-11-2011	11-04-2011	2011-04-11
April 18, 2013	04-18-2013	18-04-2013	2013-04-18
April 27, 2012	04-27-2012	27-04-2012	2012-04-27
May 11, 2012	05-11-2012	11-05-2012	2012-05-11
May 20, 2011	05-20-2011	20-05-2011	2011-05-20
May 23, 2013	05-23-2013	23-05-2013	2013-05-23
June 6, 2011	06-06-2011	06-06-2011	2011-06-06
June 17, 2012	06-17-2012	17-06-2012	2012-06-17
June 26, 2013	06-26-2013	26-06-2013	2013-06-26
July 16, 2012	07-16-2012	16-07-2012	2012-07-16
July 18, 2011	07-18-2011	18-07-2011	2011-07-18
July 22, 2013	07-22-2013	22-07-2013	2013-07-22
August 4, 2012	08-04-2012	04-08-2012	2012-08-04
August 7, 2011	08-07-2011	07-08-2011	2011-08-07
August 8, 2013	08-08-2013	08-08-2013	2013-08-08
September 9, 2011	09-09-2011	09-09-2011	2011-09-09
September 14, 2012	09-14-2012	14-09-2012	2012-09-14
September 19, 2013	09-19-2013	19-09-2013	2013-09-19
October 14, 2011	10-14-2011	14-10-2011	2011-10-14
October 26, 2012	10-26-2012	26-10-2012	2012-10-26

On using number order in file names…

If we sort by DD/MM/YY, dates are out of order.

US standard	MM-DD-YYYY	DD-MM-YYYY	YYYY-MM-DD
August 4, 2012	08-04-2012	04-08-2012	2012-08-04
June 6, 2011	06-06-2011	06-06-2011	2011-06-06
August 7, 2011	08-07-2011	07-08-2011	2011-08-07
August 8, 2013	08-08-2013	08-08-2013	2013-08-08
December 8, 2012	12-08-2012	08-12-2012	2012-12-08
September 9, 2011	09-09-2011	09-09-2011	2011-09-09
December 10, 2013	12-10-2013	10-12-2013	2013-12-10
April 11, 2011	04-11-2011	11-04-2011	2011-04-11
May 11, 2012	05-11-2012	11-05-2012	2012-05-11
January 12, 2011	01-12-2011	12-01-2011	2011-01-12
September 14, 2012	09-14-2012	14-09-2012	2012-09-14
October 14, 2011	10-14-2011	14-10-2011	2011-10-14
February 15, 2011	02-15-2011	15-02-2011	2011-02-15
March 15, 2012	03-15-2012	15-03-2012	2012-03-15
July 16, 2012	07-16-2012	16-07-2012	2012-07-16
June 17, 2012	06-17-2012	17-06-2012	2012-06-17
April 18, 2013	04-18-2013	18-04-2013	2013-04-18
July 18, 2011	07-18-2011	18-07-2011	2011-07-18
December 18, 2011	12-18-2011	18-12-2011	2011-12-18
January 19, 2012	01-19-2012	19-01-2012	2012-01-19
September 19, 2013	09-19-2013	19-09-2013	2013-09-19
February 20, 2013	02-20-2013	20-02-2013	2013-02-20
May 20, 2011	05-20-2011	20-05-2011	2011-05-20
January 22, 2013	01-22-2013	22-01-2013	2013-01-22
July 22, 2013	07-22-2013	22-07-2013	2013-07-22
May 23, 2013	05-23-2013	23-05-2013	2013-05-23
November 24, 2011	11-24-2011	24-11-2011	2011-11-24
March 26, 2013	03-26-2013	26-03-2013	2013-03-26
June 26, 2013	06-26-2013	26-06-2013	2013-06-26

On using number order in file names…

If we sort by YY/MM/DD, dates are in order.

US standard	MM-DD-YYYY	DD-MM-YYYY	YYYY-MM-DD
January 12, 2011	01-12-2011	12-01-2011	2011-01-12
February 15, 2011	02-15-2011	15-02-2011	2011-02-15
March 27, 2011	03-27-2011	27-03-2011	2011-03-27
April 11, 2011	04-11-2011	11-04-2011	2011-04-11
May 20, 2011	05-20-2011	20-05-2011	2011-05-20
June 6, 2011	06-06-2011	06-06-2011	2011-06-06
July 18, 2011	07-18-2011	18-07-2011	2011-07-18
August 7, 2011	08-07-2011	07-08-2011	2011-08-07
September 9, 2011	09-09-2011	09-09-2011	2011-09-09
October 14, 2011	10-14-2011	14-10-2011	2011-10-14
November 24, 2011	11-24-2011	24-11-2011	2011-11-24
December 18, 2011	12-18-2011	18-12-2011	2011-12-18
January 19, 2012	01-19-2012	19-01-2012	2012-01-19
February 27, 2012	02-27-2012	27-02-2012	2012-02-27
March 15, 2012	03-15-2012	15-03-2012	2012-03-15
April 27, 2012	04-27-2012	27-04-2012	2012-04-27
May 11, 2012	05-11-2012	11-05-2012	2012-05-11
June 17, 2012	06-17-2012	17-06-2012	2012-06-17
July 16, 2012	07-16-2012	16-07-2012	2012-07-16
August 4, 2012	08-04-2012	04-08-2012	2012-08-04
September 14, 2012	09-14-2012	14-09-2012	2012-09-14
October 26, 2012	10-26-2012	26-10-2012	2012-10-26
November 26, 2012	11-26-2012	26-11-2012	2012-11-26
December 8, 2012	12-08-2012	08-12-2012	2012-12-08
January 22, 2013	01-22-2013	22-01-2013	2013-01-22
February 20, 2013	02-20-2013	20-02-2013	2013-02-20
March 26, 2013	03-26-2013	26-03-2013	2013-03-26
April 18, 2013	04-18-2013	18-04-2013	2013-04-18
May 23, 2013	05-23-2013	23-05-2013	2013-05-23

On using leading zeroes in file names…

21	160	021
23	163	023
23	188	023
29	21	029
51	220	051
56	23	056
58	23	058
64	238	064
82	257	082
160	285	160
163	29	163
188	293	188
220	294	220
238	308	238
257	312	257
285	334	285
293	368	293
294	370	294
308	386	308
312	388	312
334	410	334
368	433	368
370	450	370
386	450	386
388	477	388
410	478	410
433	493	433
450	494	450
450	51	450
477	56	477
478	58	478
493	64	493
494	82	494

Version control

It is important to consistently identify and distinguish versions of data files.

This ensures that a clear audit trail exists for tracking the development of a data file and identifying earlier versions especially if data is frequently updated by multiple users.

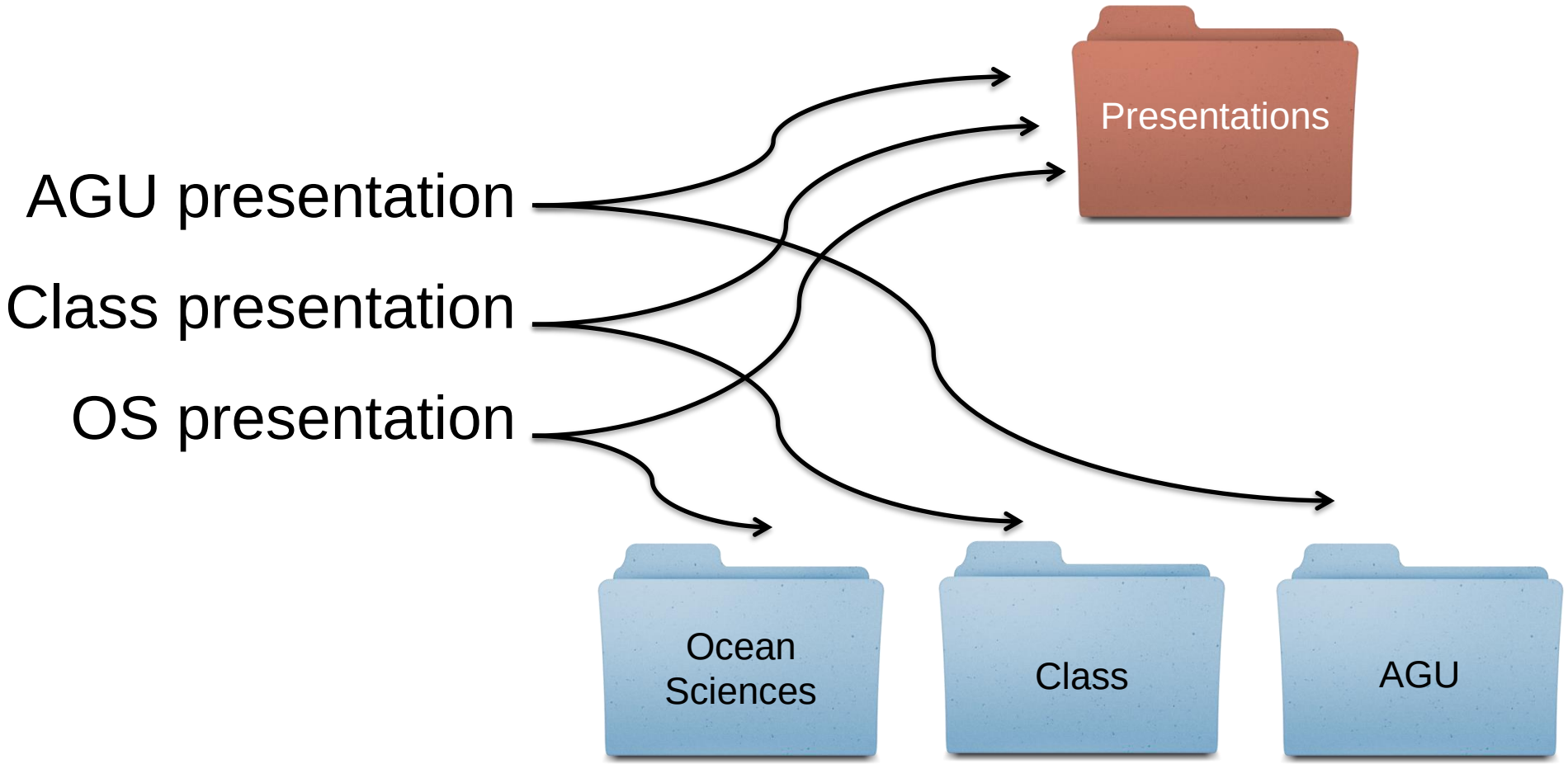
Suggested strategies:

- Use a sequential numbered system: v1, v2, v3, etc.
- Don't use confusing labels: revision, final, final2, etc.
- Record all changes -- no matter how small
- Discard obsolete versions (but never the raw copy)
- Use auto-backup instead of self-archiving, if possible

University of Leichester: Good Practice Document Version Control

List of version-control software

File organization



When **naming** & **organizing**

your files and folders…

be thoughtful

be consistent

document your approach

Exercise 4.4 DMP

4 **Master Research Projects: File Structure and Naming**

Project Title: *Provisional thesis title*

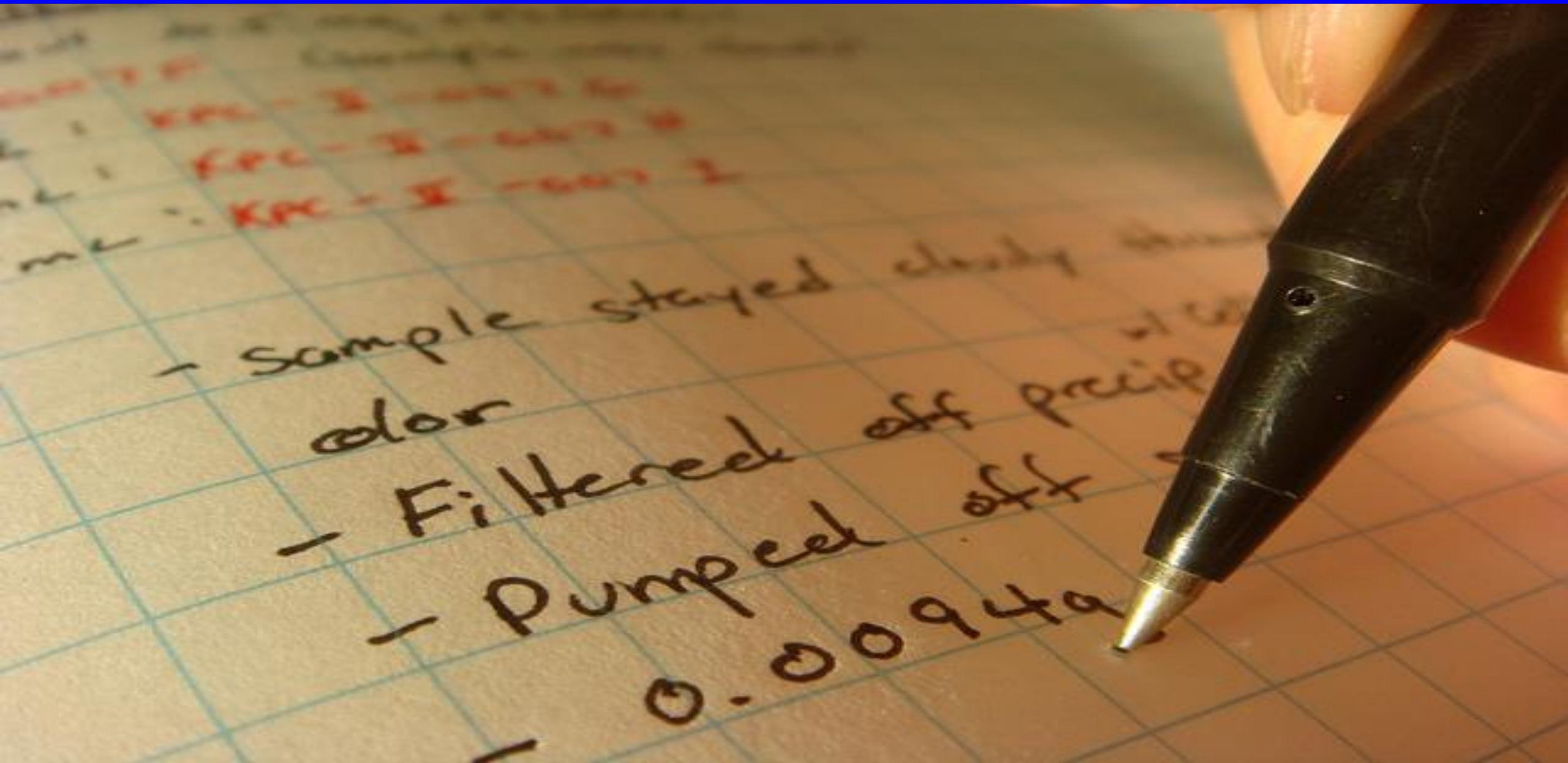
1. File Structure

a) *Describe the organisation of computer folders for your research project.*
b) *Does the file structure follow conventions from a host project, laboratory or institution?*
c) *List the primary folders, and then summarise the organisation of their sub-folders.*
d) *How will the computer folders for your master thesis be distinguished from other research projects and work that you might be involved with?*

2. File Naming

a) *Describe the logic behind the file naming system for your project.*
b) *Does the file naming follow conventions from a host project, laboratory or institution?*
c) *Give examples of the file names, from different types of digital data used in your research.*
d) *How will the file names in your master thesis be distinguished from files in other research projects and work that you might be involved with?*
e) *If a coding or numbering system is used to name files, where will the explanation of this system be saved?*

Documenting your process



Electronic notebooks ELN

SCINOTE

PROS:

- Very user-friendly and quick to set up
- Unique experimental workflow
- Open source license (MPL)
- Free account with unlimited project users

CONS:

- Drawing molecules still in development

LABFOLDER

PROS:

- Sketching
- Free account for smaller teams and free mobile app
- Integration with Mendeley

CONS:

- Not very intuitive
- Unflattering structured design
- Free version is limited up to 3 team members

BENCHLING

PROS:

- Very user friendly and quick to set up
- Useful DNA tools (CRISPR guide and primer design)
- Templates for sequence mapping and sharing
- Free account with 10 GB of storage space

CONS:

- Free account is tied to a single user
- Report structure is not flexible

HIVEBENCH

PROS:

- Plate designer
- Free account with 10 GB of storage space

CONS:

- Creating protocols is very rigid
- No possibility to create tables
- Free account is tied to ten users
- Works only on iOS

LabCollector

http://labcollector.com/

Findings

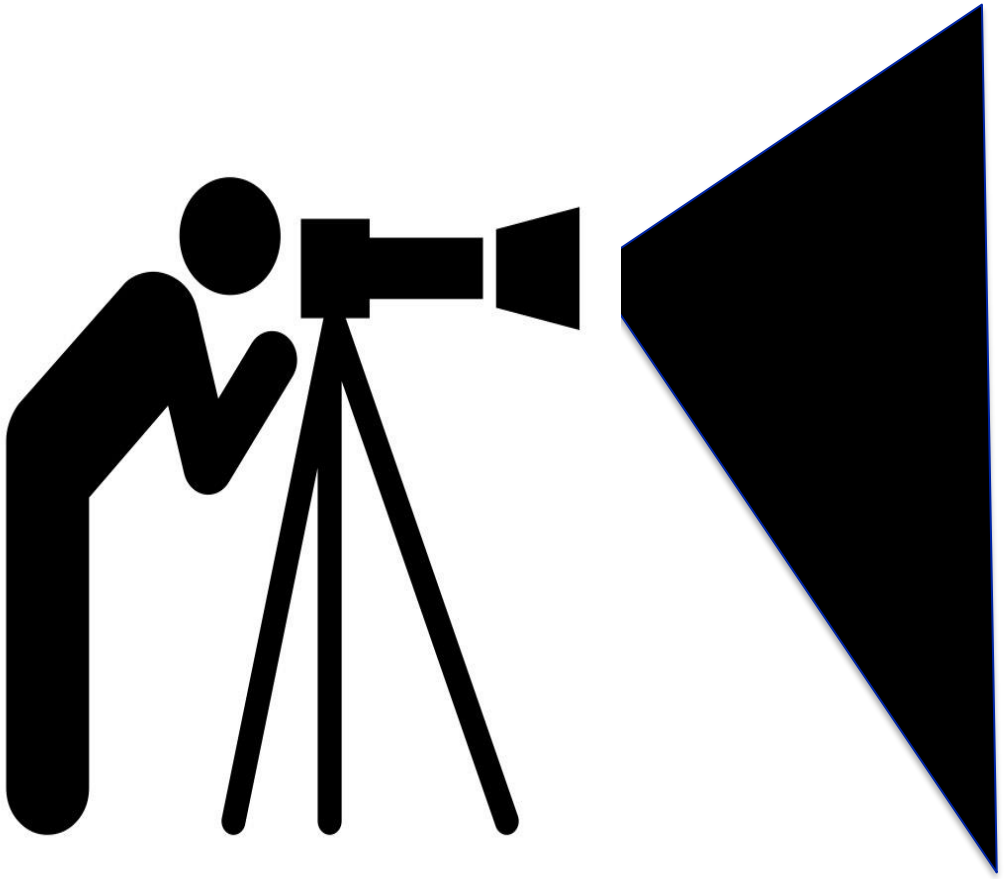
Types, Formats & Stages of Data

GEO 802, Data Information Literacy

Fall 2021 - Lecture 5

Gary Seitz, MA

Lesson 5 Outline



Luis Prado from The Noun Project

Data types & formats

Research lifecycle & stages of data

Observational data

- Observational data are historical records that were collected at a unique place and time, which renders them:
  - Irreplaceable
  - Very critical to archive

- Documenting data collection methods and equipment is critical to enable:
  - Understanding
  - Evaluation
  - Transparency
  - Reproducibility
  - Trust
  - Secondary use

Image: http://www.flickr.com/photos/60379061@N03/5509284578/

Computational models and simulations

- Computational models and simulations produce calculations and predictions of phenomena based on physical theory, algorithms, and observations.
  - We may not need to archive model outputs, but…
  - Archiving the model itself and a robust representation of its metadata (description of the hardware, software, and input data) is essential.

- Ex. The Earth System Documentation project was created to develop a standard way to describe climate models and the data they produce.
  - ES-DOC worked with climate scientists to create a conceptual information model for climate data
  - Tested, developed, and deployed the model for the CMIP5 project
  - ES-DOC: Earth System Documentation: https://es-doc.org/
  - CMIP6: https://pcmdi.llnl.gov/CMIP6/

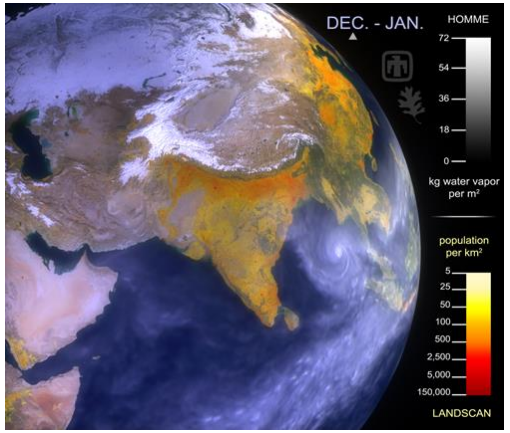


Image: http://www.flickr.com/photos/oakridgelab/4101037894/

Laboratory experimental data

- Experimental data test specific hypotheses in a controlled setting.
  - In principle, we can accurately reproduce experiments without the need to store data indefinitely.
  - However, reproducibility may pose challenges, even with highly controlled lab settings.
    - Can we precisely reproduce all the experimental conditions?
    - Nobody has been able to reliably reproduce cold fusion
    - Who will reproduce the CERN Large Hadron Collider experiments?

- Digital “workflow” systems are available for some types of lab work
  - Precise step-by-step description of a scientific procedure
  - Acts as a script to coordinate research tasks
  - Can enable automated metadata collection and provenance tracking

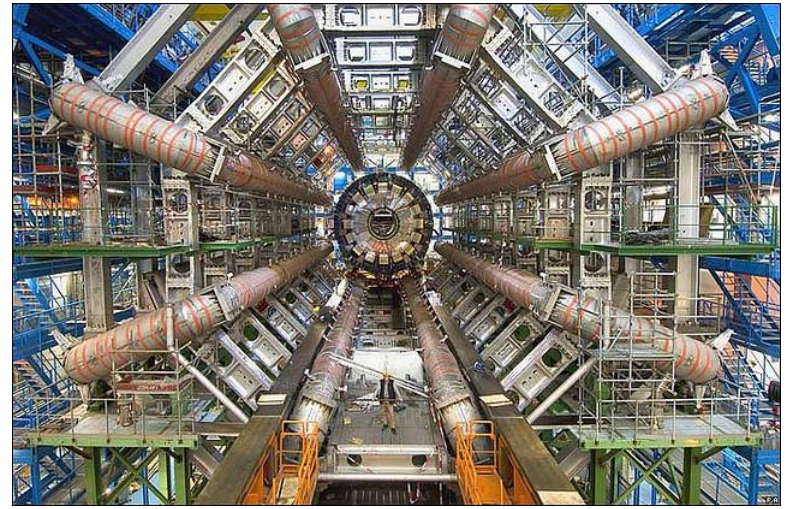


Image: http://www.flickr.com/photos/11304375@N07/2046228644/

Types of research data

- **Observational**
  - data captured in real-time, usually irreplaceable. E.g., sensor data, survey data, sample data, neurological images
- **Experimental**
  - data from lab equipment, often reproducible, but can be expensive. E.g., gene sequences, chromatograms
- **Simulation data**
  - data generated from test models where model and metadata are more important than output data. E.g., climate models, economic models
- **Derived or compiled data**
  - data is reproducible but expensive. E.g., text and data mining, compiled database, 3D models

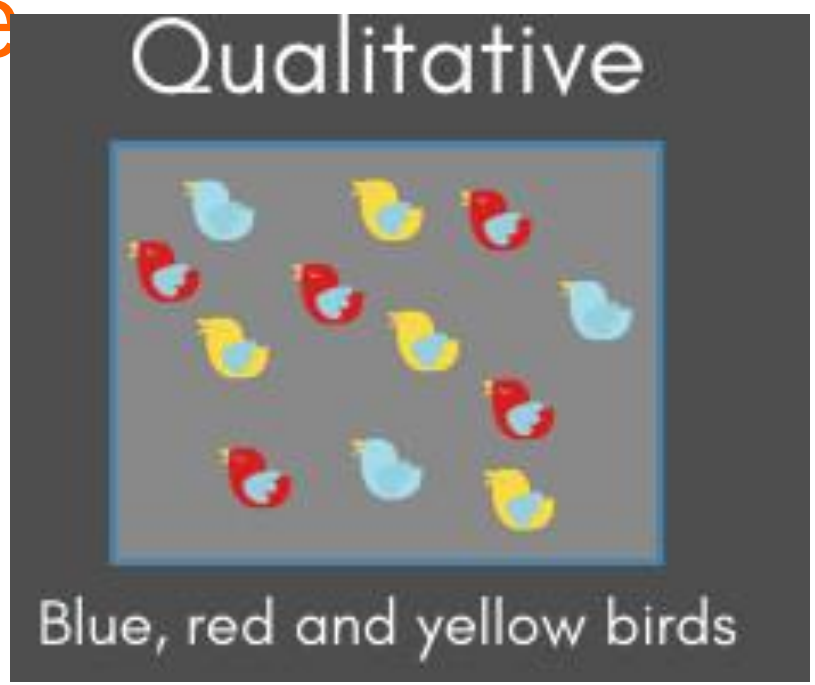
This includes:

- Text or Word documents, spreadsheets
- Laboratory notebooks, field notebooks, diaries
- Questionnaires, transcripts, codebooks
- Audiotapes, videotapes
- Photographs, films
- Test responses
- Slides, artifacts, specimens, samples
- Collection of digital objects acquired and generated during the process of research
- Data files
- Database contents including video, audio, text, images
- Models, algorithms, scripts
- Contents of an application such as input, output, log files for analysis software, simulation software, schemas
- Methodologies and workflows
- Standard operating procedures and protocols

Qualitative data

is everything that refers to the quality of something:

A description of colours, texture and feel of an object. E.g. description of experiences; interview are all qualitative data.



Quantitative data (3)

usually regarded as referring to the collection and analysis of numerical data

….which can be put into categories

or in rank order,

or measured in units of measurement.

Exercise 5.2

Qualitative v. Quantitative Data

1. The baby weighs 20 pounds.
2. My friend is very happy.
3. The sky is greyish-blue.
4. Joe is 6 foot 2.
5. Diana has $100.
6. The number of pairs of shoes you own
7. The type of car you drive
8. The place where you go on vacation
9. The distance it is from your home to the nearest grocery store
10. The number of classes you take per school year.
11. The tuition for your classes
12. The type of calculator you use
13. Movie ratings
14. Political party preferences
15. Weights of sumo wrestlers
16. Amount of money (in dollars) won playing poker
17. Number of correct answers on a quiz
18. Peoples' attitudes toward the government
19. IQ scores

File formats

File formats

Digital data can take countless different form(at)s…

A file format is a specific way of structuring information so that a machine, and therefore a person, can understand it

- should be readable by as many types of system as possible
- without compromising the purpose of the data

Preferable format types for longterm access to data

Data formats that offer the best chance for long-term access are **both**:

- Non-proprietary (also known as *open*), **and**
- Unencrypted & uncompressed

Preferred formats

File type	Recommended	Suitable to only a limited extent	Not suitable for archiving
Text	• PDF/A (*.pdf) • Plain Text (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py, *.r etc.) coded as ASCII, UTF-8, or UTF-16 using byte order mark • XML (inclusive XSD/XSL /XHTML etc.; with included or accessible schema and character encode explicitly specified)	• PDF (*.pdf) with embedded fonts • Plain text (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py, *.r etc.) (ISO 8859-1 coded) • Rich Text Format (*.rtf) • HTML and XML (The ASCII text is readable over long term; try to avoid external links.) Not accepted for publication, OK for supplementary materials: • Word *.docx • PowerPoint *.pptx • LaTeX, TeX (The ASCII text is readable over long term; open source software required for formatting and the resulting PDF should be included.) • OpenDocument formats (*.odm, *.odt, *.odg, *.odc, *.odf)	• Word *.doc • PowerPoint *.ppt
Spreadsheet or table	• Comma- or tab delimited text files (*.csv)	• Excel *.xlsx (container format) • OpenDocument spreadsheets (*.ods)	• Excel *.xls, *.xlsb (binary formats)
Raw data and workspace		• ASCII Text is suitable for long-term use, but the data import may be time-consuming. • S-Plus files (*.sdd) may be saved as text files. • Matlab *.mat files may be saved in HDF Format. Saving nontrivial ASCII Matlab *.mat files should be avoided because they are not readable with the Matlab load command (see table 2). • Network Common Data Format or NetCDF (*.nc, *.cdf) • Hierarchical Data Format (HDF5) (*.h5, *.hdf5, *.he5)	• Binary files such as the standard Matlab files *.mat or the R files *.RData

ETH BIBLIOTHEK

https://documentation.library.ethz.ch/display/RC/File+formats+for+archiving

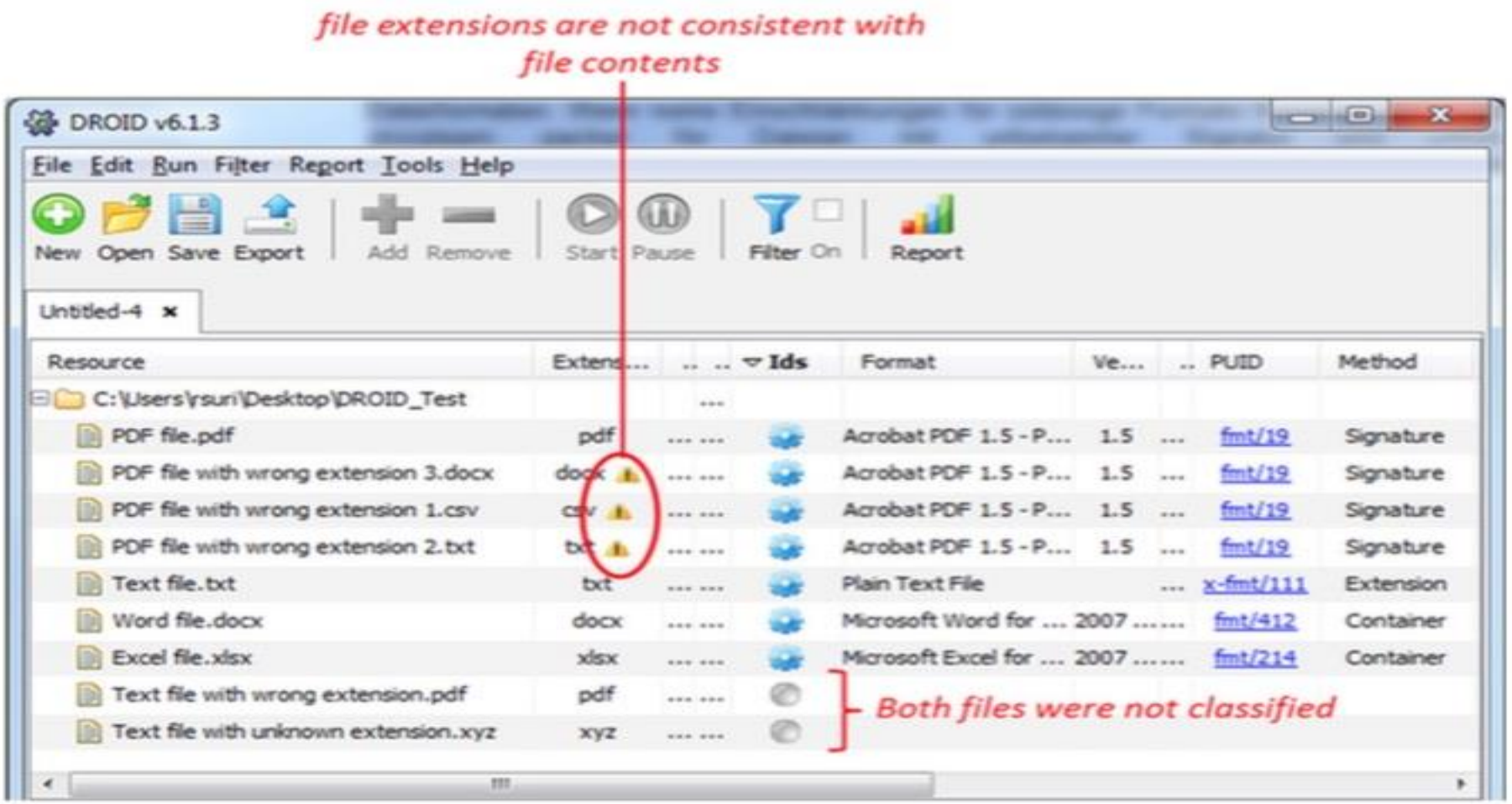
Preferred formats

Raster image (bitmap)	• TIFF (*.tif) (uncompressed, preferentially TIFF 6.0, Part 1: baseline TIFF). TIFF is preferred as compared to PNG or JPEG2000. • Portable Network Graphics (*.png, uncompressed) • JPEG2000 (lossless compression) • Digital-Negative-Format (*.dng) to keep raw data of digital fotos in addition to an second copy in TIFF format	• TIFF (*.tif) (compressed) • GIF (*.gif) • BMP (*.bmp) • JPEG/JFIF (*.jpg) • JPEG2000 (lossy compression) (*.jp2)	
Vector graphics	• SVG without JavaScript binding (*.svg)		• Graphics InDesign (*.indd), Illustrator (*.ait) • Encapsulated Postscript (*.eps) • Photoshop (*.psd)
CAD	• AutoCAD Drawing (*.dwg) • Drawing Interchange Format, AutoCAD (*.dxf) • Extensible 3D, X3D (*.x3d, *.x3dv, *.x3db)		
Audio	• WAV (*.wav) (uncompressed, pulse-code modulated)	• Advanced Audio Coding (*.mp4) • MP3 (*.mp3)	
Video[1)]	• FFV1 codec (version 3 or later) in Matroska container (*.mkv)	• MPEG-2 (*.mpg,*.mpeg) • MP4, which is also called MPEG-4 Part 14 (*.mp4) • QuickTime Movie (*.mov) [2)] • Audio Video Interleave (*.avi) • Motion JPEG 2000 (*.mj2, *.mjp2)	• Windows Media Video (*.wmv)

ETH BIBLIOTHEK

DROID: file format identification tool

For large data collections you can get an overview of your file formats using the free JAVA application DROID. Furthermore, this tool detects unknown file formats as well as inconsistencies between file extensions and file contents



Exercise 5.4

Example of file formats in a DMP

If you were the reviewer of this DMP, how would you asses this part of the plan?

The project will use a variety of open and proprietary formats that will best suit the needs of the project's outcomes. These will be migrated to suitable open standards to facilitate preservation at the end of the project. Text will be transcribed as plain text with HTML markup and will take up roughly 500Mb of space. Images will be in the JPEG format and 5Gb of server space will be set aside for them. Video files will be MOV and 20Gb of space will be available for them.
Visualisations will be SVG files. The map interface will be based around Google Maps. Web pages will follow current HTML and CSS standards

Exercise

Comments by a reviewer

1. Information supplied is **vague**. ‘A variety of open and proprietary formats that will best suit the needs of the project’s outcomes’: What **exactly** are all these formats and how do they best suit the needs of the project’s outcomes? This section doesn’t demonstrate that the project actually knows what it’s talking about.
2. ‘These will be migrated to suitable open standards to facilitate preservation at the end of the project’: again, this sounds impressive but is too vague. **What** open standards? And why couldn’t these open standards just be used from the start of the project?
3. Some file formats and standards are mentioned but the reason for their choice is not made clear, for example:
    a) Why will video files be in the MOV format when this means users will need to install the Quicktime plugin for the files to be playable in their browser?
    b) Why was Google Maps chosen over other maps interfaces?
4. Information about the textual data is far too **vague**. The recipe data will have to be structured in some way and there is no information about how the project intends to do this. Will the data be stored in a relational database? Will recipes be marked up in XML?
5. The section does mention that text will be transcribed with HTML markup but this approach is not considered best practice. HTML should be used to mark up presentation not content, and if markup is to be used to denote ingredients then XML would be better suited. Alternatively the project could record ingredients in a relational database.
6. Insufficient information is provided about the visualisations. What sort of visualisations will be created? How will they be created? Section 1 states that the visualisations will be ‘interactive’ so they presumably won’t just be static SVG files but will change to reflect choices made by the user.

Exercise 5.5 DMP

What data will you collect, observe, generate or re-use?

- What type, format and volume of data will you collect, observe, generate or reuse?
- Which existing data (yours or third-party) will you reuse?
- How will the data be collected, observed or generated?

a) **Data type**:
Briefly describe categories of datasets you plan to generate or use, and their role in the project

b) **Data origin**:
to be mentioned if you are reusing existing data (yours or third-party one). Add the reference of the source if relevant.

c) **Explain how the data will be collected, observed or generated**. De- scribe how you plan to control and document the consistency and quality of the collected data

d) **Format of raw data** (as created by the device used, by simulation or downloaded):
open standard formats should be preferred, as they maximize reproducibility and reuse by others and in the future

e) **Format of curated data** (if applicable):
open standard formats should be preferred

f) **Estimation of volume of raw and curated data**

Data documentation through metadata

GEO 802 Fall 2021, Data Information Literacy

Anna C. Véron, Dr. sc. nat.

Lesson 6: Data documentation through metadata

→ Definition of metadata

→ Why we need metadata

→ FAIR data

→ Metadata standards

→ How to write quality metadata

What is metadata?

– "Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource." *NISO, Understanding Metadata*

– It can be used to describe physical items as well as digital items (documents, audio-visual files, images, datasets, etc.)

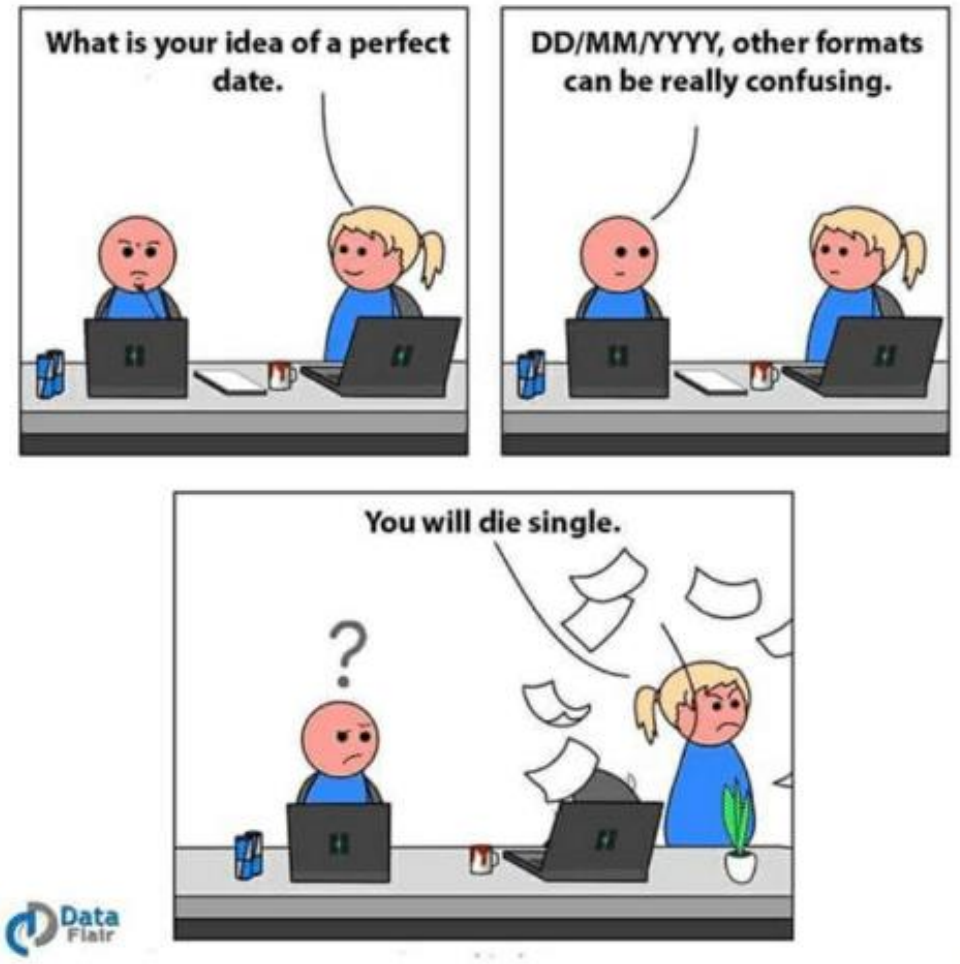
– Metadata can take many different forms, from free text (such as read-me files) to standardized, structured, machine-readable content

– For data to be useful, it will also need subject-specific metadata (reagent names, experimental conditions, population demographic…)

Working with data

When you **receive a dataset** from an external source, what types of details do you want to know about the data?

When you **provide data** to someone else, what types of information should you include with the data?



FAIR principles

- Introduced in 2016 by FORCE 11
  (= representatives from science, funding institutions, publishers, libraries, archives)
- Goal: optimal processing of research data for both human and machine
- 15 Principles
- Explanation by the SNF:
  https://tinyurl.com/SNFfair

Findability

- **Persistent identifier (PID): e.g. Digital Object Identifier (DOI)**

- Descriptive metadata in a machine readable format
  - Title, author / creator of data
  - Context, quality, condition and characterization of the data
  - How was the data generated?
  - Which information is needed to interpretate the data?

Data that really saves lives (and possibly your organisation) by Fredric Landqvist. findwise.com/blog

Accessibility

– Open access to anyone in the world with a computer and internet connection (no charge, no other access restrictions)

  Limitations

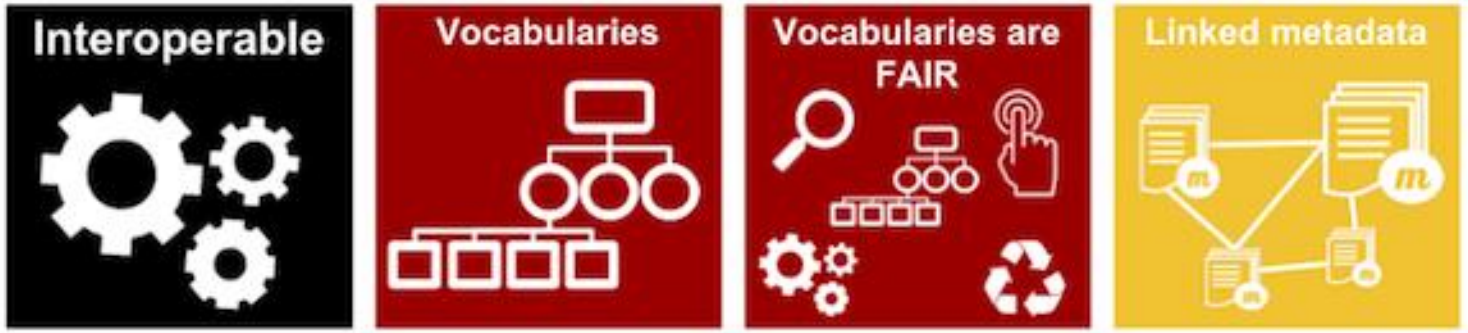
  - Data which are subject to data protection and privacy laws (e.g. involving living individuals)
  - Data from international collaboration with countries that have laws prohibiting the open sharing of data

– At least the **metadata have to be accessible**!

Data that really saves lives (and possibly your organisation) by Fredric Landqvist. findwise.com/blog

Interoperability

- Data and metadata have to be fully compatible between different computer operating systems
- **Open file formats** (files can be used with with freely available software)
- Use of **controlled vocabulary** with an easily findable and accessible documentation
- Citation of relevant / associated data sets



Data that really saves lives (and possibly your organisation) by Fredric Landqvist. findwise.com/blog

Re-usability

- Metadata must contain any **information necessary to properly understand and use the data**. The categories of metadata must be explained or self-explanatory.
- Data needs to be **reliable (reproducible)** and **understandable**!
- Include information about the **license** in the metadata. Whenever possible, the data must be **labelled for reuse**.

Data that really saves lives (and possibly your organisation) by Fredric Landqvist. findwise.com/blog

Exercise 6.1: FAIR Data

- Compare the metadata of two different datasets
- Are the FAIR principles implemented? If yes, how? What is missing?

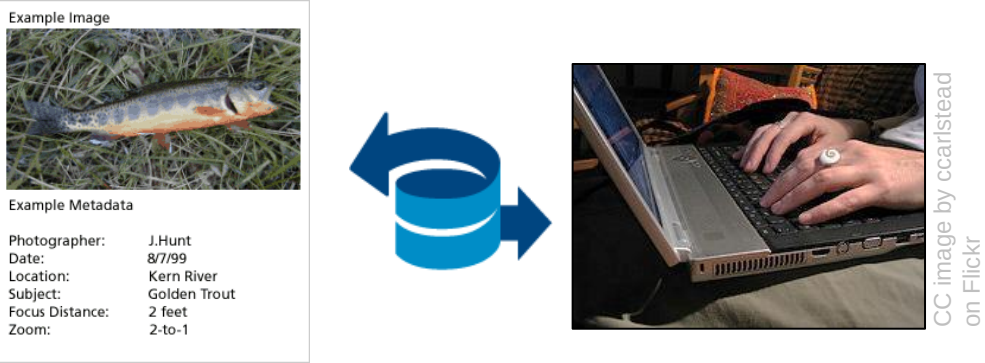
https://tinyurl.com/Zenodo

PANGAEA.
Data Publisher for Earth & Environmental Science

https://tinyurl.com/panga123

What is a metadata standard?

- **A Standard provides a structure to describe data with**
  - Common terms to allow consistency between records
  - Common definitions for easier interpretation
  - Common language for ease of communication
  - Common structure to quickly locate information
- **In search and retrieval, standards provide:**
  - Documentation structure in a reliable and predictable format for computer interpretation
  - A uniform summary description of the dataset



Examples of metadata standards

- **Darwin Core** | biological diversity, taxonomy
- **Dublin Core |** general
- **DDI** (Data Documentation Initiative) | social & behavioral sci.
- **DIF** (Directory Interchange Format) | environmental sci.
- **EML** (Ecological Metadata Language) | ecology, biology
- **ISO** 19115| geographic data

Examples of metadata standards

- **Dublin Core Element Set**
  - Emphasis on web resources, publications
  - http://dublincore.org/documents/dces/

- **FGDC Content Standard for Digital Geospatial Metadata (CSDGM)**
  - Emphasis on geospatial data
  - http://www.fgdc.gov/metadata/geospatial-metadata-standards

- **Biological Data Profile (BDP) of the CSDGM**
  - Profile to the CSDGM emphasis on biological data (and geospatial)
  - https://www.fgdc.gov/standards/projects/metadata/biometadata/index_html

- **ISO 19115/19139 Geographic information: Metadata**
  - Emphasis on geospatial data and services
  - http://www.fgdc.gov/metadata/geospatial-metadata-standards#fgdcendorsedisostandards

Examples of metadata standards

- **Ecological Metadata Language (EML)**
  - Focus on ecological data
  - http://knb.ecoinformatics.org/eml_metadata_guide.html

- **Darwin Core**
  - Emphasis on museum specimens
  - http://rs.tdwg.org/dwc/index.htm

- **Geography Markup Language (GML)**
  - Emphasis on geographic features (roads, highways, bridges)
  - http://www.opengeospatial.org/standards/gml

- **OGC® WaterML**
  - WaterML 2.0 is a standard information model for the representation of water observations data
  - http://www.opengeospatial.org/standards/waterml

Exercise 6.2: Metadata Standards

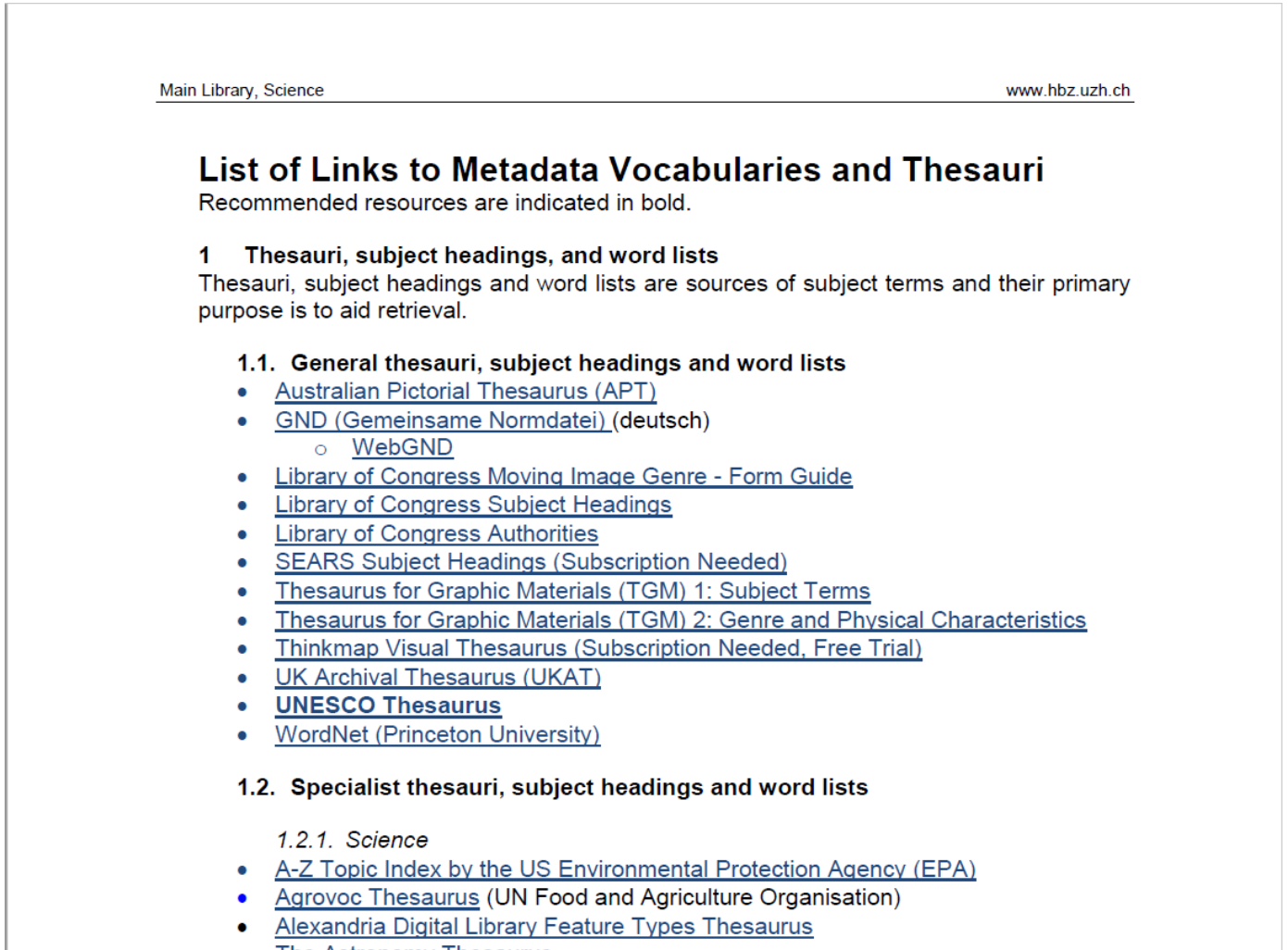
- Browse through metadata standards by discipline.
- Take note of standards that might be relevant for your field.

→ http://www.dcc.ac.uk/resources/metadata-standards

- http://rd-alliance.github.io/metadata-directory/subjects

Exercise 6.3: Controlled Vocabularies

- Check the list of controlled vocabularies in the handout.
- Take note of controlled vocabularies that might be relevant for your field.



List of Links to Metadata Vocabularies and Thesauri

Recommended resources are indicated in bold.

1 Thesauri, subject headings, and word lists

Thesauri, subject headings and word lists are sources of subject terms and their primary purpose is to aid retrieval.

1.1. General thesauri, subject headings and word lists

- Australian Pictorial Thesaurus (APT)
- GND (Gemeinsame Normdatei) (deutsch)
  - WebGND
- Library of Congress Moving Image Genre - Form Guide
- Library of Congress Subject Headings
- Library of Congress Authorities
- SEARS Subject Headings (Subscription Needed)
- Thesaurus for Graphic Materials (TGM) 1: Subject Terms
- Thesaurus for Graphic Materials (TGM) 2: Genre and Physical Characteristics
- Thinkmap Visual Thesaurus (Subscription Needed, Free Trial)
- UK Archival Thesaurus (UKAT)
- **UNESCO Thesaurus**
- WordNet (Princeton University)

1.2. Specialist thesauri, subject headings and word lists

1.2.1. Science

- A-Z Topic Index by the US Environmental Protection Agency (EPA)
- Agrovoc Thesaurus (UN Food and Agriculture Organisation)
- Alexandria Digital Library Feature Types Thesaurus

Summary of Lesson 6

Metadata is documentation of data.

Metadata allows data to be discovered, accessed, and re-used.

A metadata standard provides structure and consistency to data documentation.

Document your process while you are creating and analyzing data.

Metadata completes a dataset.

Creating quality metadata is in your OWN best interest!

Data backup, security, storage & preservation

GEO 802 Fall 2021, Data Information Literacy

Anna C. Véron, Dr. sc. nat.

Lesson 7: Data backup, security, storage & preservation

→ **Why?**

→ **Where to store data**

→ **Data backup**

→ **Data security**

→ **Data preservation**

Why caring about data storage and preservation?

- Media formats age rapidly.
- Storage structures are not documented because they are clear to the person responsible.
- Contact persons are mobile and may only stay at the university for a few years.
- Storage locations are subsequently moved and links broken.
- The variety of subject-specific file and metadata formats makes long-term technical usability more difficult.

Image: Janet McKnight via Flickr, 15422638442 CC BY

Data loss will happen to you

Hard drive failures

Theft or loss of equipment

Dropping your laptop

Research trends (follow the money consequences)

File formats not readable anymore

People move to a new lab

Overwriting data

Media degradation (CDR's, memory sticks, hard drives, etc.)

Obsolescence / upgrades

Non-understandable data (bad metadata)

Where do you store your data?

- PCs & Laptops
- External storage devices
- Network drives
- Cloud servers

→ Go to www.menti.com and use the code **4185 6100**

Storage: PC / Laptop

Pros:

- Convenient
- Accessible

Cons:

- Drive failures are common
- Susceptible to theft and damage
- Not replicated

Bottom Line: Do not use to store master copies of data. Only for «work in progress» files.

Storage: External storage devices

Pros:

- Convenient
- Cheap & portable

Cons:

- Longevity not guaranteed
- Easily damaged, misplaced or lost
  Security risk!
- Might not be big enough to hold all data

Bottom Line: Do not use to store master copies of data. Not recommended for long-term storage.

Storage: Network storage devices

Pros:

- Replicated storage: Less vulnerable to loss due to hardware failure
- Secure storage minimizes risk of loss, theft, unauthorized use

Cons:

- Costs
- Need network connection to access files
  (can be slow, especially when working on files with software)

Bottom Line: Highly recommended for master copies of data. Suitable for long-term storage (5 years or more).

Storage: in the Cloud

Pros:

- Backed up regularly and automatically
- Replicated storage: Less vulnerable to loss due to hardware failure
- Most provide versioning and encryption
- Secure storage minimizes risk of loss, theft, unauthorized use

Cons:

- Stored data *may* not be entirely private
- Service provider may go out of business --- Longevity?
- Possible restrictions by funding agency

Bottom Line: Recommended for master copies of data. Long-term storage?

Exercise 7.1: SWITCH drive

Try SWITCH drive for data entry and collaboration

1. Create an account in SWITCH. https://preview.tinyurl.com/switchdrive
2. Work in groups of three.
3. Create a file and share it with your partners.
4. Try to both work on it simultaneously.

Exercise 7.2: Backup vs. Archive

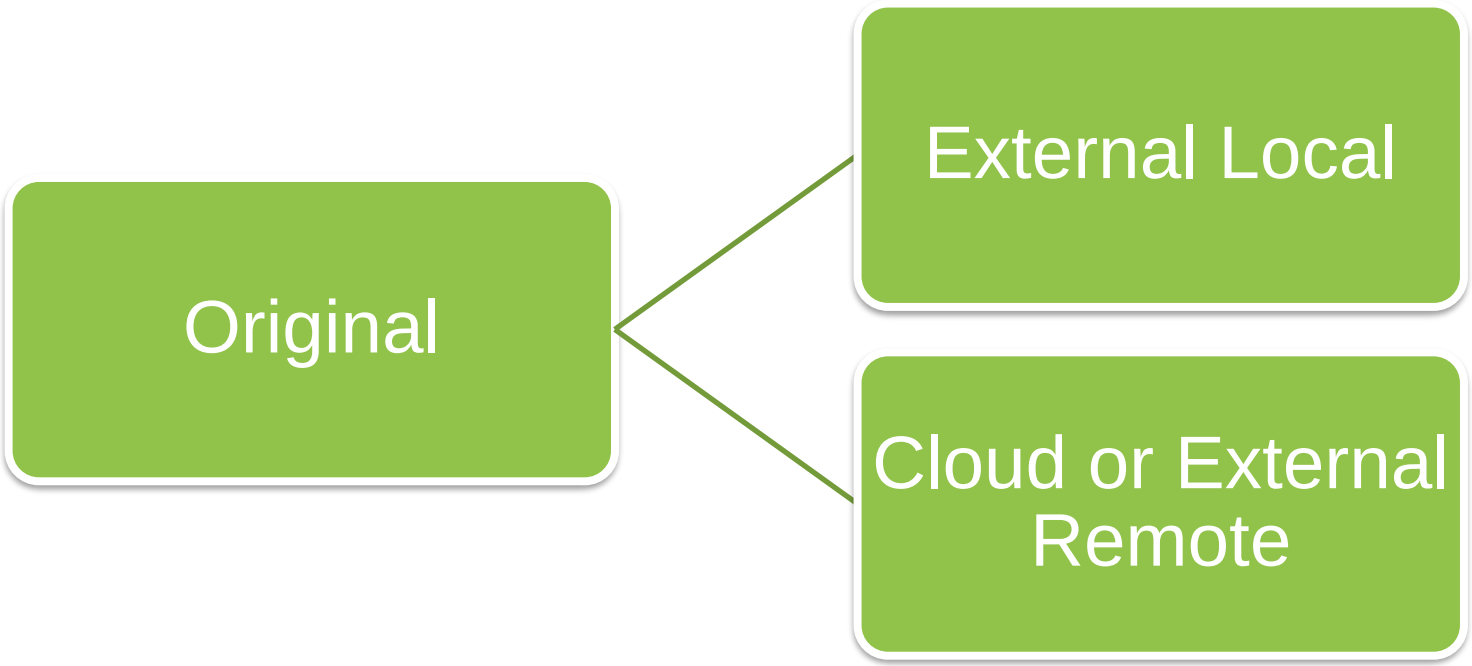
- What is backup and what is archiving?
- Quickly note down your answer:
  Backup =
  Archive =

1 Data may change
2 Finalized data; static record
3 Kept for 5+ years
4 Often stored in official archive
5 Preservation formats
6 Not permanent
7 Usually stored locally (individual, department, institute)
8 “Working” formats

Backup: 3-2-1 Rule

Best Practice:

- 3 Copies of datasets on
- 2 different devices/media of which
- 1 is off site



Information security

Threats & Defences

Undetected, unauthorized access

Backdoors, Exploits

Viruses, Trojans, Worms

Malware, Ransomware, spyware

Phishing

Logic bombs

Keylogging

etc.

Access control

Application security, e.g. Antivirus software

Authentication (e.g. multi-factor auth.)

Intrusion detection system

Encryption

Firewall

Access control

- **Are your data sensitive?**
- Who **has access** to the data?
- Who **is allowed to have access** to the data?
- How can data access be controlled? (Username / Password)

Image: Barry Levine, martechtoday.com @martech_today

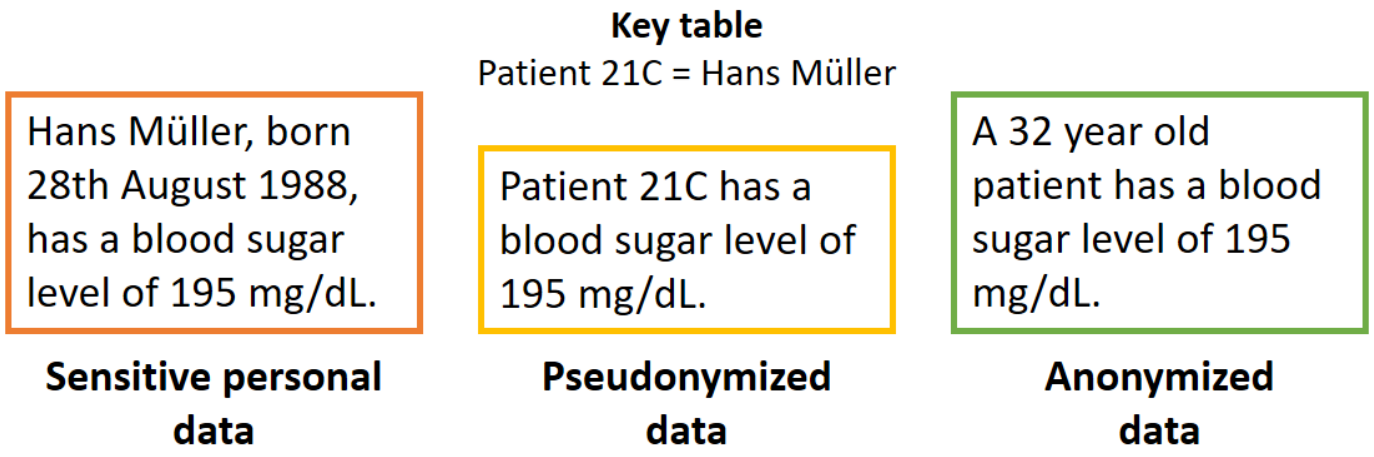
Data security measures for sensitive data

– **Anonymization: irreversibly destroy** any way of identifying the data subject.
Truly anonymized data is not sensitive anymore!

amnesia

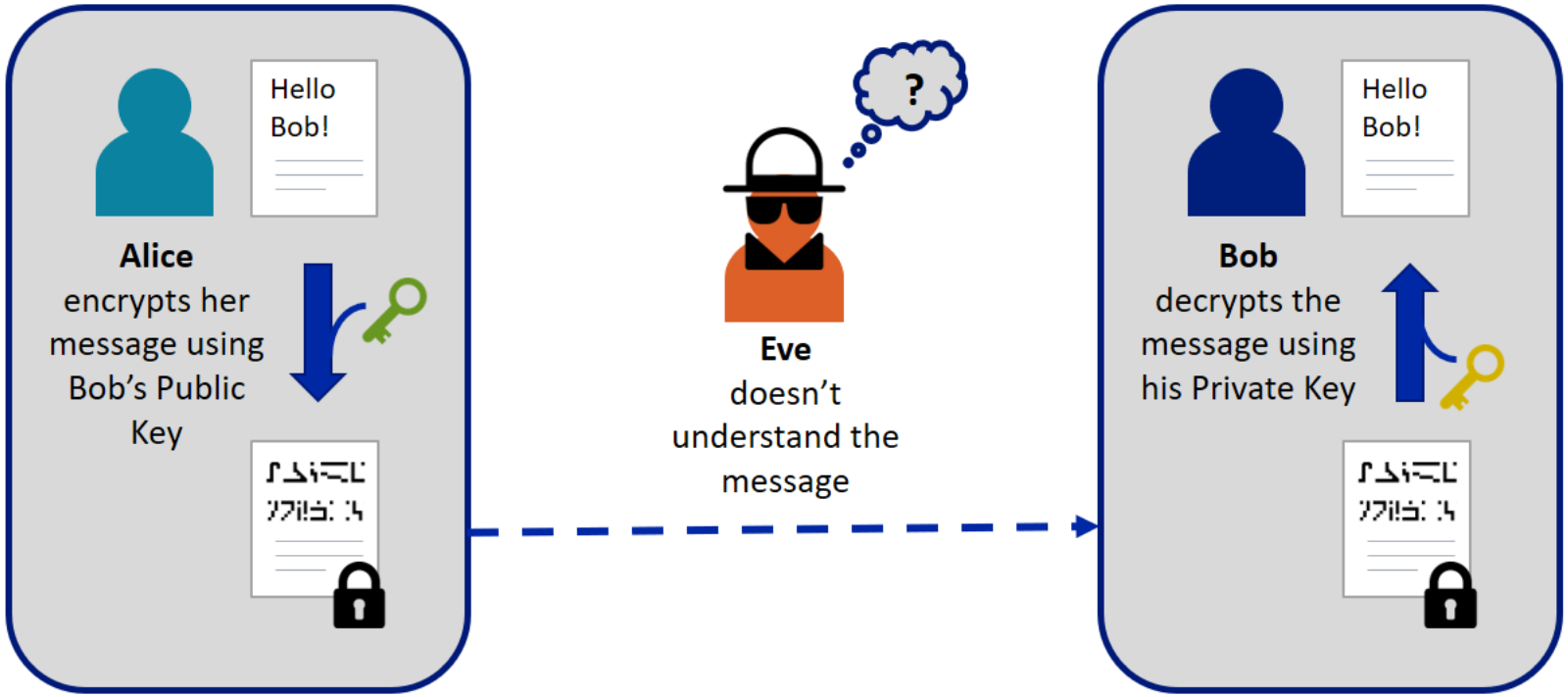
amnesia.openaire.eu/
(data anonymization tool)

– **Pseudonymization:** substitutes the identity of a subject with a «nickname» (code, pseudonym). It is a secure approach if the personal identifiers are stored in a separate location.
Pseudonymized data are still sensitive data, because it can be linked back to the person!



Encryption

– Encryption is the process of encoding digital information in such a way that only authorised parties can view it.



Wikipedia article: Alice and Bob

– It is possible to encrypt individual files, folders, or entire disk volumes or USB storage devices.

– Encryption software: generates encryption / decryption keys (passwords).

Encryption Software

BitLocker
integrated in Windows (since Vista); offers encryption of disk volumes and USB devices

FileVault2
standard for Macs, (OS X Lion or later); full disc encryption

VeraCrypt
free, open source multi-platform encryption software (Windows, Mac and Linux);
full disk, partition and container encryption

Axcrypt
for Windows, Mac and mobile devices; basic version is free; premium version with monthly subscription and extensive features (e.g. secure files in Dropbox, Google Drive etc.).

Exercise 7.3: Group Discussion

What are the pros and cons of portable storage devices?

Can you provide examples where researchers in your community have lost research data and/or when you have lost digital files yourself? How was the data lost and could it have been prevented?

What reasons might there be for researchers not to use networked drives to back up research data?

What are the principal reasons why researchers should back up their data files?

Do you use a cloud-based storage service?
Why or why not?

What methods could you use to manage sensitive data?

Exercise 7.4: Long-term preservation

- Look at the file formats you generally use for your research data and compare them to the list provided by the ETH library.
  https://documentation.library.ethz.ch/display/DD/File+formats+for+archiving
- Are your file formats suitable for long-term preservation?
- If not, which formats could they be converted to?

Summary of Lesson 7

Choose the right storage medium for your data.

Recognize sensitive data and treat them accordingly.

Don’t be lazy with backups! Remember the 3-2-1 rule.

Regularly assess the security of your data. Consider using an encryption software.

Choose strong passwords and change them regularly to keep your data secure.

Think of the long-term preservation of your data. Curate your data, choose appropriate file formats for archiving and create quality metadata.

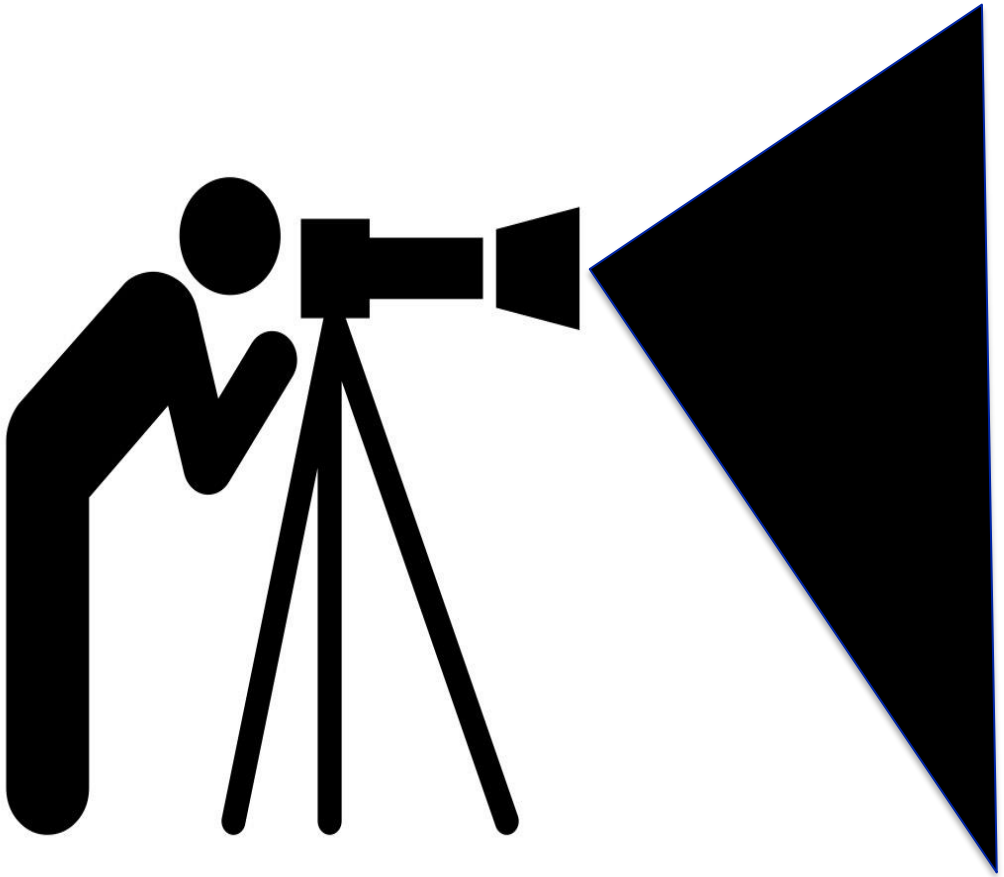
Data sharing & reuse

GEO 802, Data Information Literacy

Fall 2021 – Lecture 8

Gary Seitz, MA

Lesson 8 Outline



Luis Prado from The Noun Project

Benefits of sharing data

Issues/obstacles related to reuse and sharing of data

Understand open access

Data citation

Value of Data Sharing:
To the Public

A better informed public yields better decision making with regard to:

- Environmental and economic planning
- Federal, state, and local policies
- social choices such as use of tax money and education options
- personal lifestyle and health such as nutrition and recreation
- Better decision making in voting

Value of data sharing:
to researcher sponsor

- Organizations that sponsor research must maximize the value of research money
- Data sharing enhances the value of research investments by enabling:
  - verification of performance metrics and outcomes
  - new research and increased return on investment
  - advancement of the science
  - reduced data duplication expenditures

Value of data sharing:
to scientific community

- Access to related research enables community members to:
  - build upon the work of others and further, rather than repeat, the science
  - perform meta analyses that cannot be performed with individual datasets or laboratories
  - share resources and perspectives so that comprehension is expanded and enhanced
  - increase transparency, reproducibility and comparability of results
  - expand methodology assessment, recommendations and improvement
  - educate new researchers as to the most current and significant findings

Value of data sharing:
to the scientist

Scientists that share data gain the benefit of:

- research sponsor recognition as an authoritative source and wise investment
- improved data quality due to expanded use, field checks, and feedback
- greater opportunity for data exchange
- improved connections to scientific network, peers, and potential collaborators

Barriers to data sharing

– *“Scientists would rather share their toothbrush than their data!”*

[Carole Goble, Keynote address, EGEE (Enabling Grids for EscienceE) ‘06 Conference.]

– **Barriers to sharing can relate to …**

- the Researcher - intellectual property issues (Lecture 9)
- the Institution - unrealised commercial value
- the Subject – confidentiality (Lecture 9)

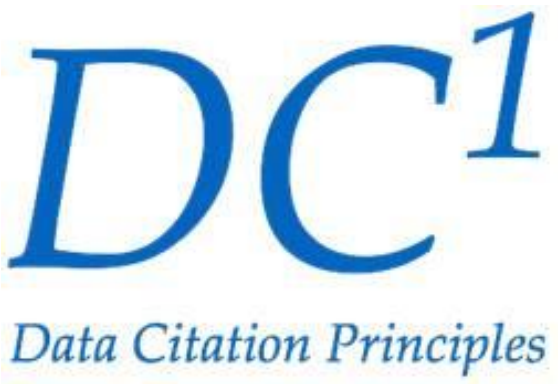
Preservation & sharing platforms

Definitions

- **Data citation**
  - The practice of providing a reference to data in the same way as researchers routinely provide a bibliographic reference to printed resources
  - A key practice underpinning the recognition of data as a **primary research output** rather than as a by-product of research
- **Data author**
  - Individual involved in research, education, or other activities that generate digital data that are subsequently deposited in a data collection
- **Persistent identifier**
  - A unique web-compatible alphanumeric code that points to a resource (e.g., data set) that will be preserved for the long term (i.e., over several hardware and software generations)
  - Should direct to latest available version of resource or to metadata which enables acquisition of desired version or format

8 Principles of Data Citation

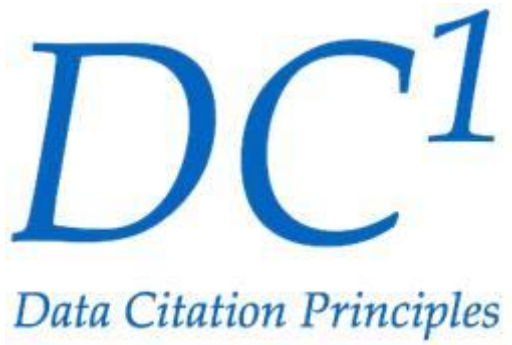
1. **Importance** - Data should be considered legitimate, citable products of research
2. **Credit and Attribution** - Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data
3. **Evidence** - In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
4. **Unique Identification** - A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.



https://www.force11.org/datacitation

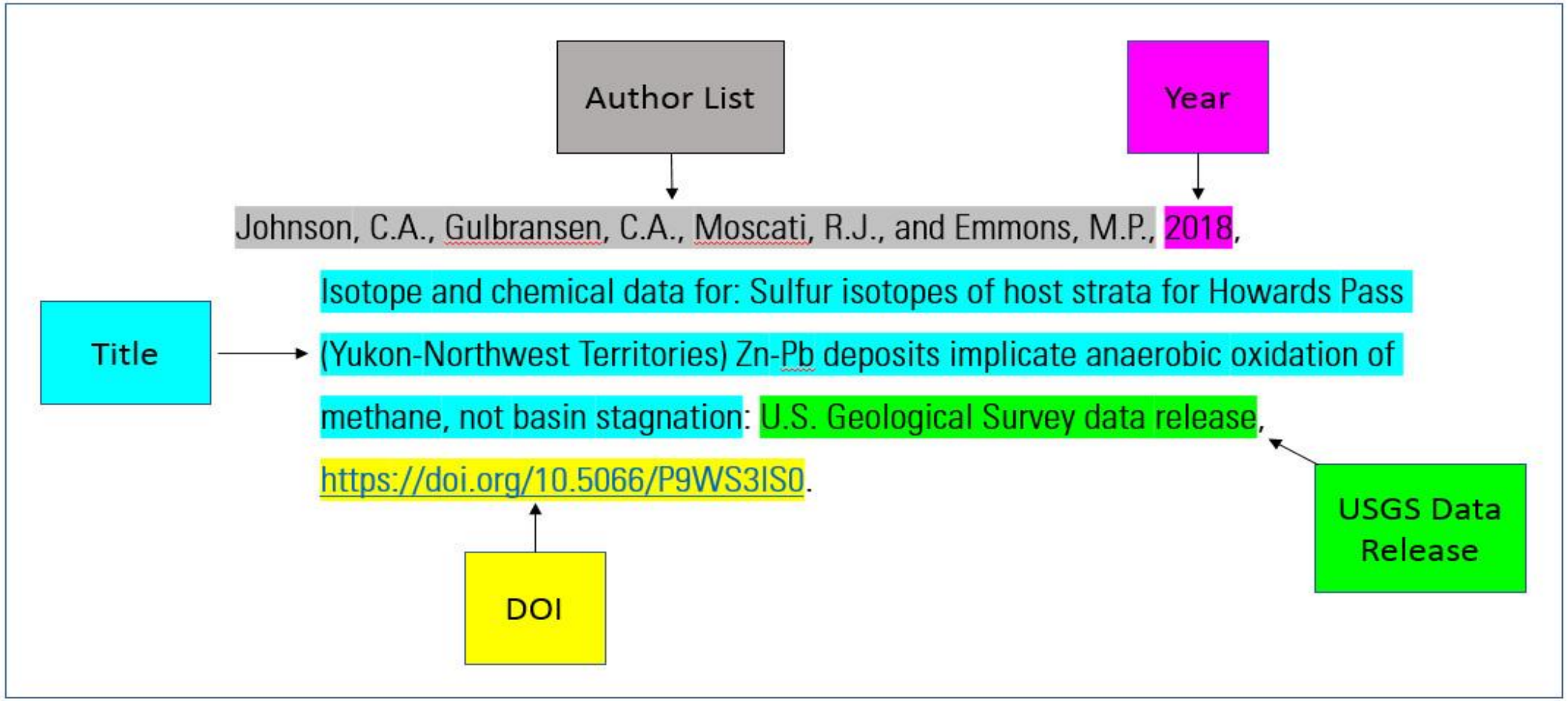
8 Principles of Data Citation

5. **Access** - Data citations should facilitate access to the data, metadata, code, and other materials, as necessary for both humans and machines.
6. **Persistence** - Unique identifiers, data, and metadata should persist beyond the lifespan of the data they describe.
7. **Specificity and Verifiability** - Data citations should facilitate identification of, access to, and verification of the specific data that support a claim.
8. **Interoperability and Flexibility** - Data citation methods should be flexible, but enable interoperability across communities.



https://www.force11.org/datacitation

Examples of data citation formats



Unique identifiers

– **DOI: Digital object identifier**

- *Digital Identifier* of an *Object* (not "Identifier of a Digital Object")
- Object = any entity (thing: physical, digital, or abstract)
  - Resources, parties, licenses, etc.
- Digital Identifier = network actionable identifier ("click on it and do something")

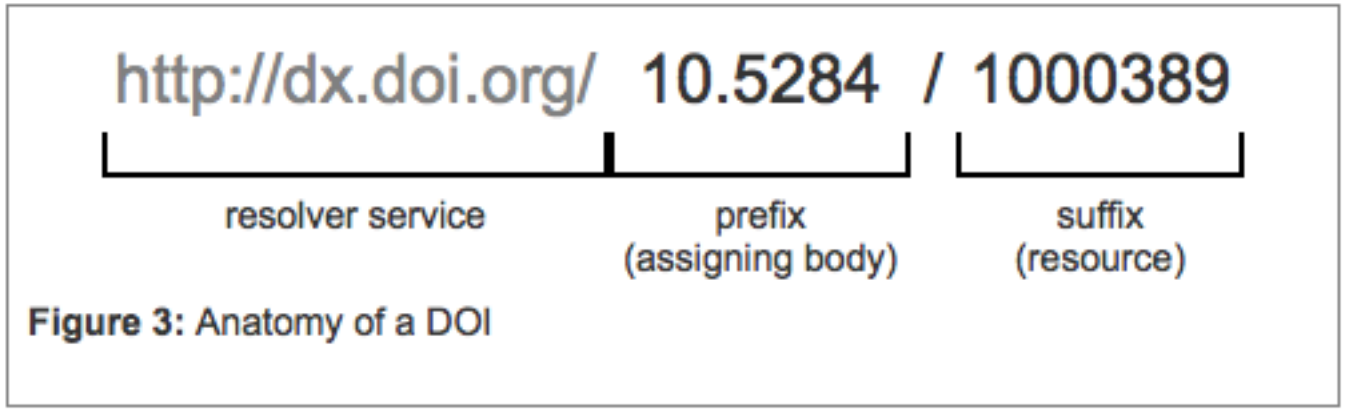


Figure 3: Anatomy of a DOI

http://www.dcc.ac.uk/resources/how-guides/cite-datasets#sthash.assrcuj3.dpuf

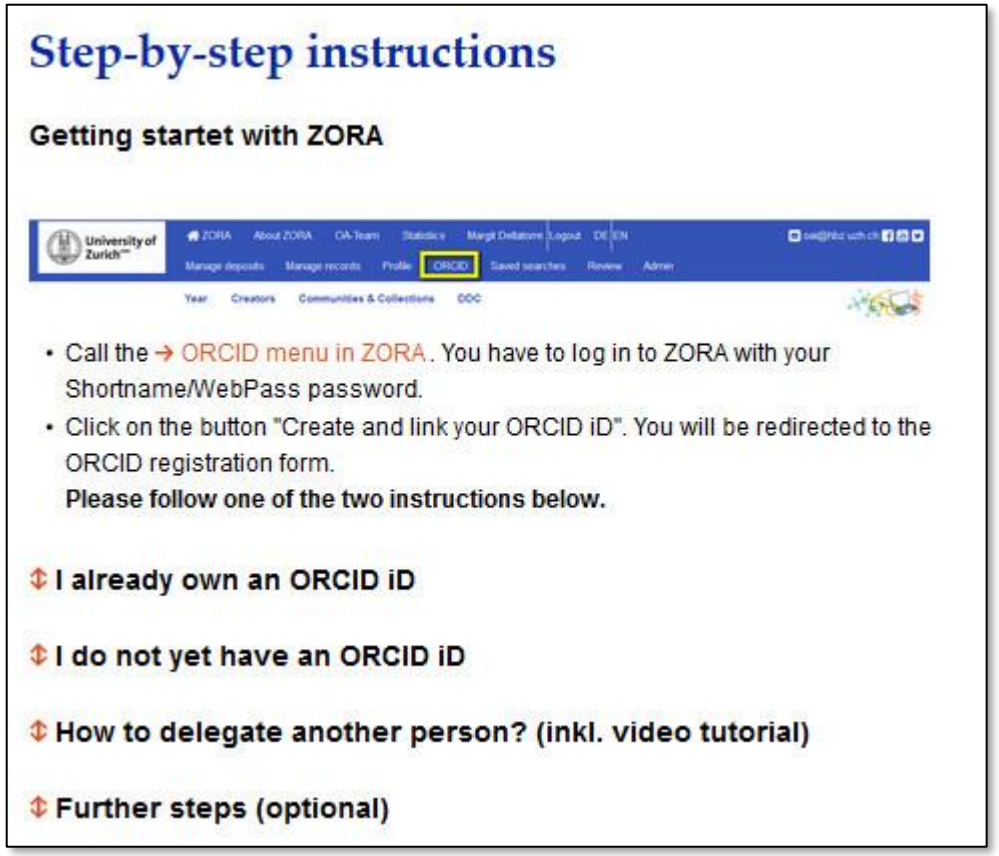
Unique researcher IDs

Disambiguate yourself via:

ORCID
Connecting Research and Researchers

Open Researcher & Contributor ID

ORCID provides a persistent digital identifier that distinguishes you from every other researcher and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between you and your professional activities ensuring that your work is recognized.



Exercise 8.5: Cite the following Datasets

Fluctuations of Glaciers (FoG) Database

- DOI for current scientific data (Identifier): 10.5904/wgms-fog-2018-11
- Creator: World Glacier Monitoring Service (WGMS)
- Title: Fluctuations of Glaciers Database
- Publisher: World Glacier Monitoring Service (WGMS)
- Publication Year: 2018
- Release date: 2018-11-03

Glaciers 2012

- From Repository: Edinburgh DataShare - University of Edinburgh
- By:Gruber, Stephan
- Edinburgh DataShare
- DOI: http://dx.doi.org/10.7488/ds/1878
- Viewed Date: 04 May 2017
- Published: 2017
- Document Type:Data study

CrowdWater game data

- From Repository: Zenodo
- By:Strobl, Barbara; Etter, Simon; Van Meerveld, Ilja; Seibert, Jan
- Zenodo
- DOI: http://dx.doi.org/10.5281/ZENODO.2630586
- Viewed Date: 22 May 2019
- Published: 2019
- Version: 28.02.2019
- Document Type:Data set
- Data Type:Dataset

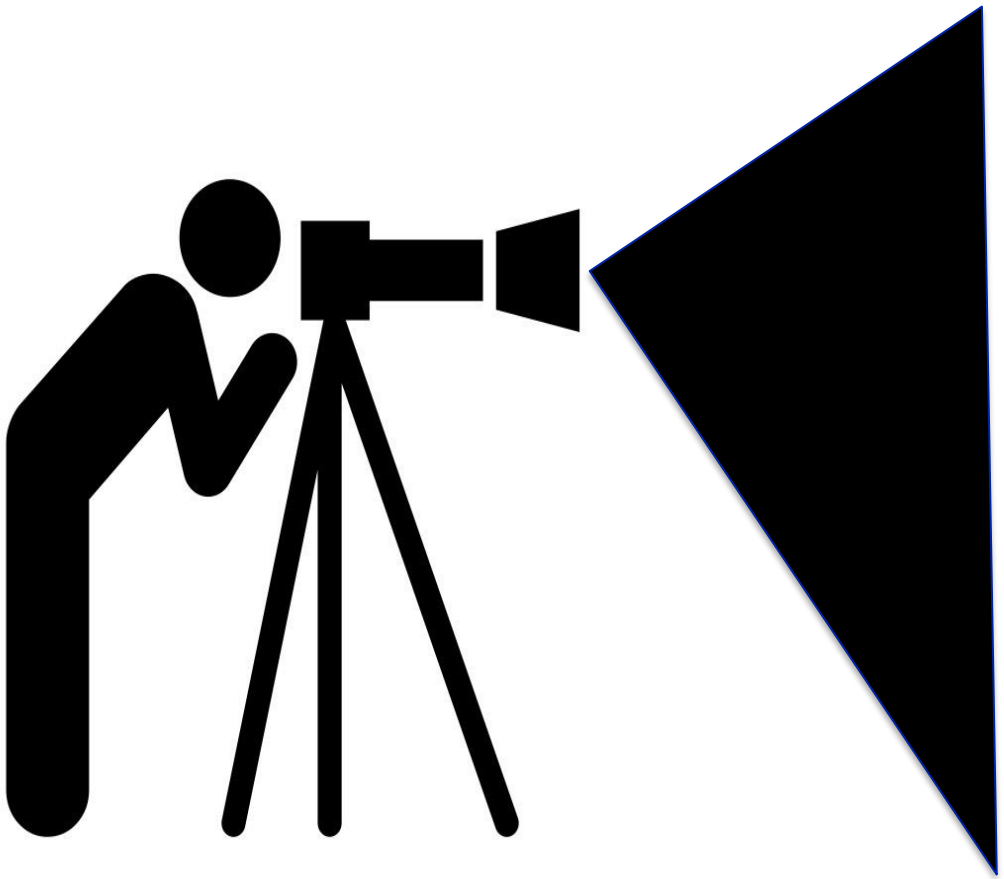
Ethics and copyright

GEO 802, Data Information Literacy

Fall 2021 – Lecture 9

Gary Seitz, MA

Lesson 9 Outline



Identify ethical, legal, and policy issues

Define copyrights, licenses and waivers

Understand reasons behind data restrictions

Discuss ethical considerations

Luis Prado from The Noun Project

Why might data use or sharing be restricted?

- Threatened and endangered species
- National security and classified research
- Export controls
  - Can apply to technologies and data
- Use of Human Subjects
  - Personally identifiable information of any kind

Which research is subject to ethical review?

– Any research involving the use of human participants or live animals will usually be subject to ethical review. Similarly, research which involves referencing individual subjects (people) or storing identifiers for individuals will usually be subject to ethical review, **unless** the data is obtained through pure observational studies of public behaviour.

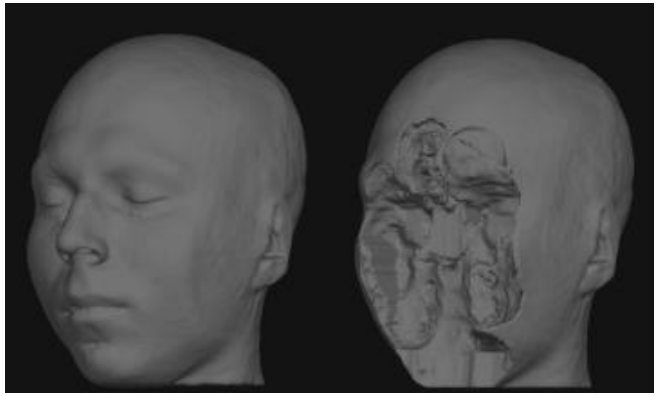
–

'Pure' observational studies

- are of human action that occurs in a forum open to the general public
- are non-invasive and non-interactive (pure observational)
- require no interaction with participants
- do not identify participants

Deidentification of Research Data

The process of anonymizing data to protect the identity of the participants and to remove other private information.



General Guidelines:

- Mark replacements of text clearly, either by using [brackets] or tags:
  <anon> … </anon>
- Keep a secure copy of the non-anonymized data.
- **Create a log of all the replacements, aggregations, or removals made in each data file. Store this log file separately from the de-identified data.**

- Adapted from Alicia Mohr's "Resources for Qualitative Data Management"

In practice: example anonymisation

Ex 1. Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003 (study 5407 in UK Data Archive collection) by M. Mort, Lancaster University, Institute for Health Research.

Date of Interview: 21/02/02

Interview with [Lucas Roberts,] DEFRA field officer
Date of birth: [2 May] 1965
Gender: Male
Occupation: Frontline worker
Location: [Plumpton,] North Cumbria

[Lucas] was living at home with his parents, "but I'm hoping to move out soon" so we met at his parents' small neat house. We sat in a very comfortable sitting room with an open fire and [Lucas] made me coffee and offered shortbread. Although at first [Lucas] seemed a little nervous, quick to speech and very watchful he seemed to relax as we spoke and to forget abut the tape.

I will just start by asking you to tell me a little bit about yourself and your background.

Well it is an agricultural background. I grew up on the farm where my brother is now. After I left school I did work on the farm but went to college and did exams, did land use recreation, sort of countryside/ environmental management course. So I obviously left agriculture, did the course and came back [to the farm] at weekends.

Comment [v1]: Replace: Ken
Comment [v2]: delete
Comment [v3]: delete
Comment [v4]: Replace: Ken
Comment [v5]: Replace: Ken
Comment [v6]: Replace: Ken

Exercise 9.1 DMP

- **How will you address and handle ethical issues in your thesis?**

  - **List any ethical isssues or sensitive data involved in the research project**
    - Human participants
    - Privacy issues (confidential or sensitive data)
    - Animal experiments
    - …

  - **Describe how these ethical issues will be managed.**
    - Consent from ethics committee
    - anonymization of personal data
    - sensitive data not stored on cloud services
    - …

Who Owns Data?

– It depends…

- Who is funding your research?
- What is your institution's data ownership policy?
- Are you working under a grant or contract?
- Who owns my data?

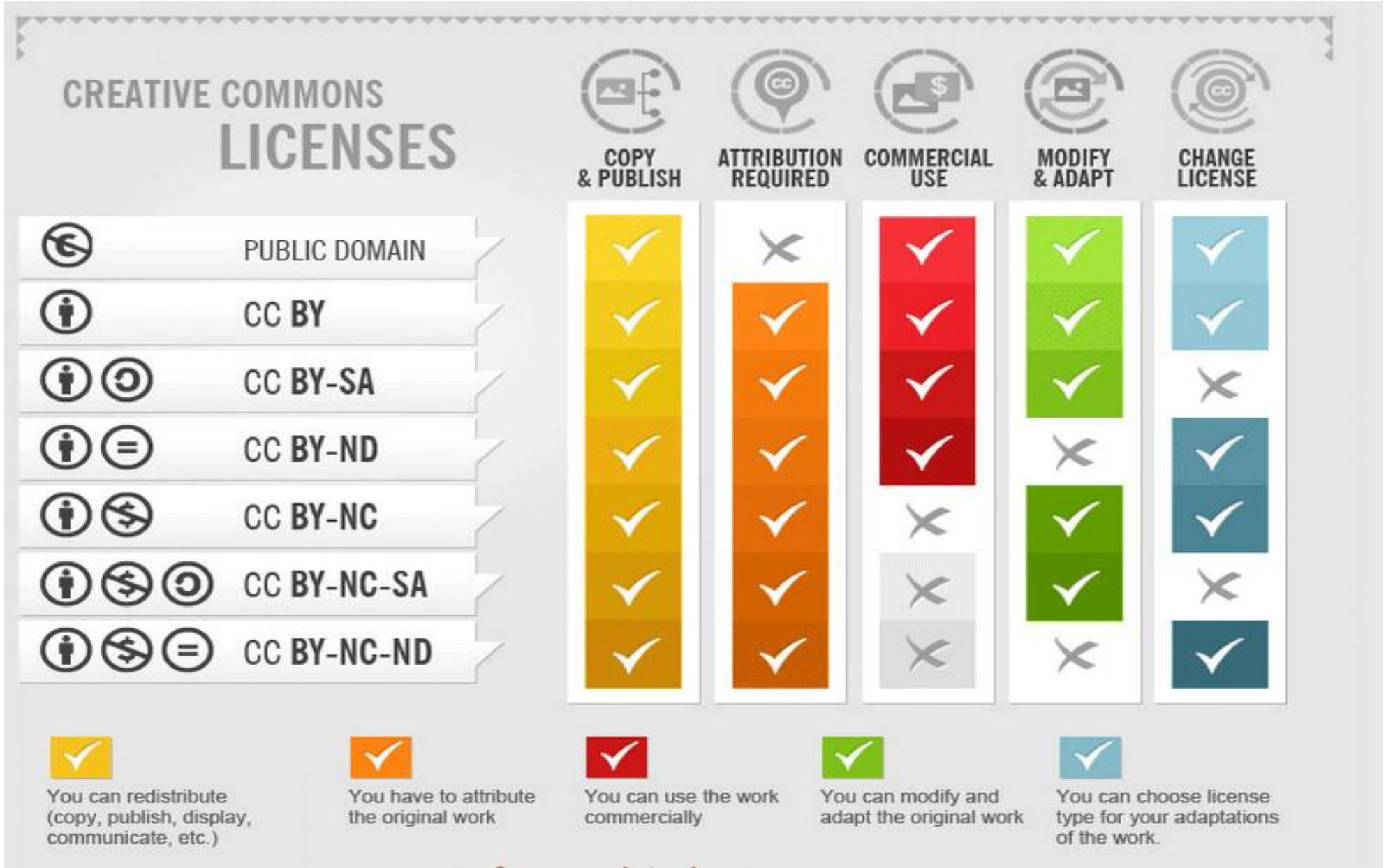
Copyright versus License

- **Copyright**: " [T]he body of exclusive rights granted by law to copyright owners for protection of their work." (U.S. Copyright Office)
  - Facts and *data* cannot be protected by copyright
  - Metadata and data arrangement can be protected (sometimes)
- **License**: States what can be done with the data and how that data can be redistributed (*e.g.*, GPL and CC)
- **Waiver**: (*e.g.*, CC0) relinquishes all rights of ownership and usually commits the "work" to the public domain
- Intellectual property laws will vary depending upon country or region

Choosing an open license

- Why use an open license?
  - Facilitate data sharing and discovery
  - Increase visibility of your data
  - Advance knowledge
- Creative Commons
  - CC0 (not a license, but a waiver)
  - CC-BY (Attribution)
  - CC-BY-ND (Attribution-NoDerivs)
  - CC-BY-NC (Attribution-NonCommercial)
  - CC-BY-SA (Attribution-ShareAlike)

Creative Commons Licensing



Exercise 9.2: Copyright Quiz: True or False?

1. The ownership of copyright is the same for creators of work regardless of their academic status (e.g. students or lecturers), employment status (e.g. employed or self employed)

2. Intellectual Property Rights can be bought, sold, rented, gifted and bequeathed

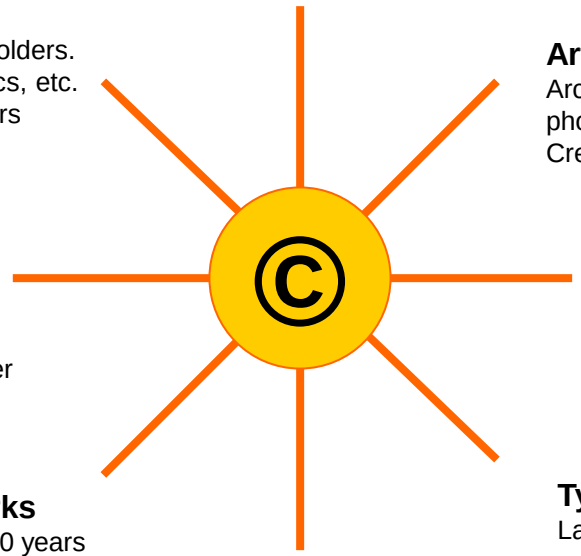
3. Copyright requires registration

4. Copyright protection lasts forever

5. Most web content can be re-used freely

6. The onus of responsibility lies with the user of a work to get permission, even if the rights holder is unknown or cannot be traced

Creative works fixed in material form



Literary works
Published and unpublished works
Creator's life +70 years / +50 years
Unknown creator: 70 / 50 years from creation

Musical works
Multiple © types and holders.
Composition, song lyrics, etc.
Creator's life + 70 years

Artistic works
Archaeological illustrations & photographs, etc.
Creator's life + 70 years

Film
Multiple © types and holders.
70 years after death of last surviving principal director, screenplay authors, composer of film music

Sound recordings
© held by both recorder & recorded
50 years from creation

Dramatic works
Creator's life + 70 years

Typographic arrangements
Layout of text, tables, & arrangement of database etc.
25 years from publication of work

Broadcasts
Multiple © types and holders.
Lasts 50 years from broadcast date

Data Management Plans (DMPs)

GEO 802 Fall 2021, Data Information Literacy

Anna C. Véron, Dr. sc. nat.

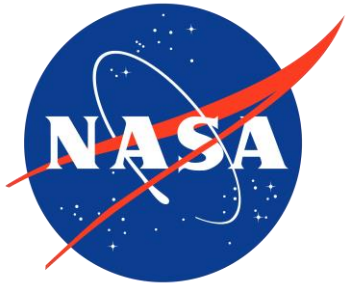
Lesson 10: Data Management Planning

→ Introduction: Why and what for?

→ DMP step by step

→ Your turn!

No DMP, no money!



Features of a DMP

- Formal document
- Outlines what you will do with your data **during** & **after** you complete your research
- Ensures your data is safe for the **present** & the **future**

FNSNF

- The DMP is an integral part of the submitted proposal, but…
  - …it is **not part of the scientific evaluation** of a proposal
  - …a **plausible DMP draft** is sufficient at the time of submission

→ Within the project period the **DMP shall be changed and adapted** at any time.

→ At the end of the project **your DMP will be openly shared** on P3 (SNSF public database)

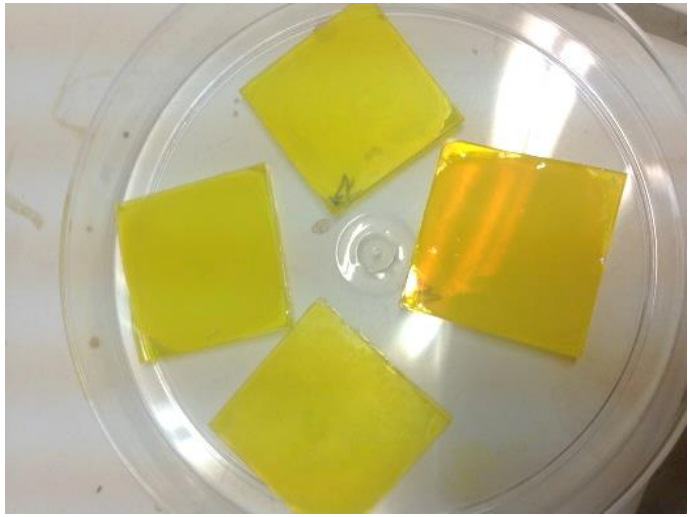
Example Research Project

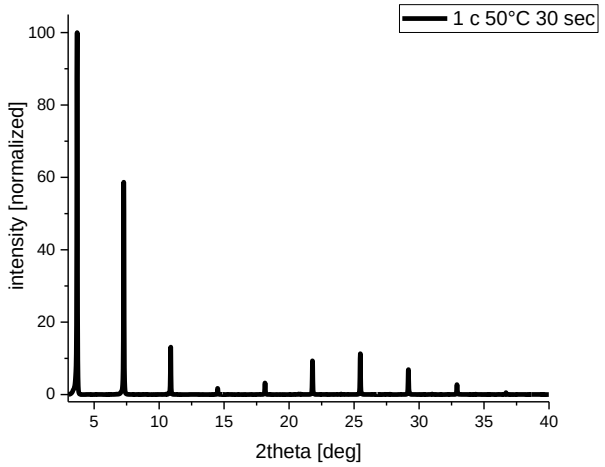
«**Perovskite Thin Films**»

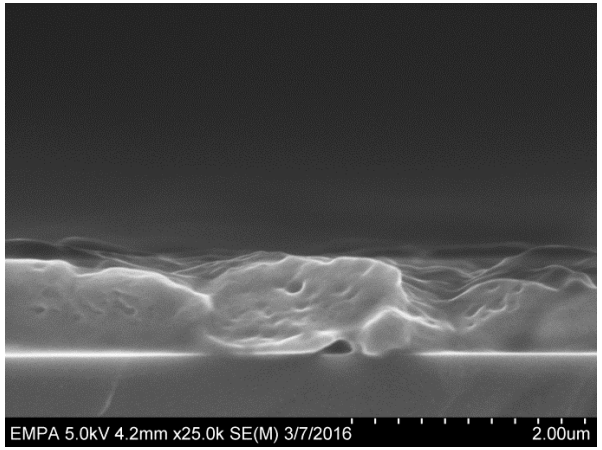
– preparation of perovskite thin films under different conditions
– characterization by X-ray diffraction

Let's create a DMP for this project!

Me in my previous life







Summary of Lesson 10

A DMP is essential for successful data management.

The four main topics of a DMP are:

1. **Data collection and documentation**
2. **Ethics, legal and security issues**
3. **Data storage and preservation**
4. **Data sharing and reuse**

The DMP is a “work in progress”, i.e. it can and *should* be updated during the research project.

Use a tabular format to create your DMP. Make it clear and concise and avoid long “flowery” sentences.

Student's projects: DMP

- Examples of DMPs produced by students of this class
- Pick one and discuss in groups how you would evaluate these DMPs.

tinyurl.com/GEODMP