Data Management Plans

Dr. Tomasz Miksa

TU Wien & SBA Research

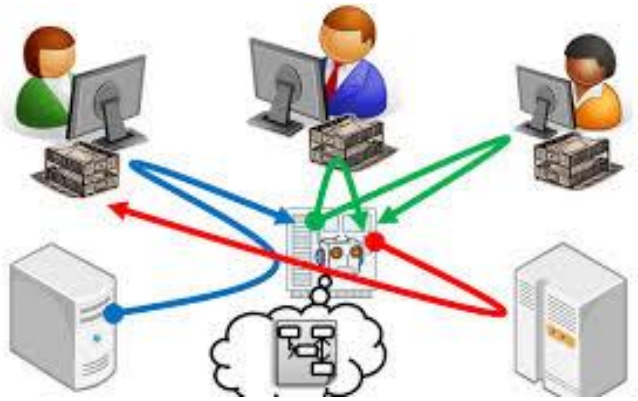
miksa@ifs.tuwien.ac.at

Agenda

- Why do we need to manage data properly?
- What are Data Management Plans (DMPs)?
- How to create a DMP?

e-Science

- Data intensive science
- Research Infrastructures
- Data
  - is fuel for research
  - is the result of processes such as
    - capturing
    - pre-processing
    - transformation
    - integration
    - analysis

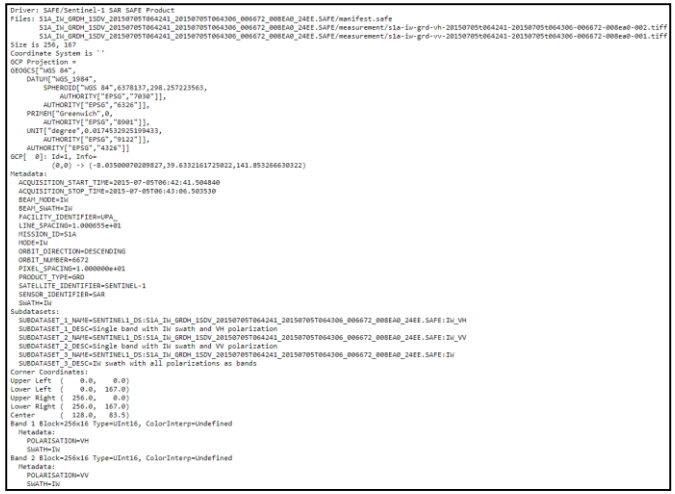


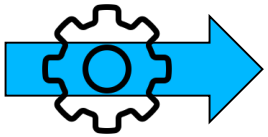
eScience and Research Infrastructures

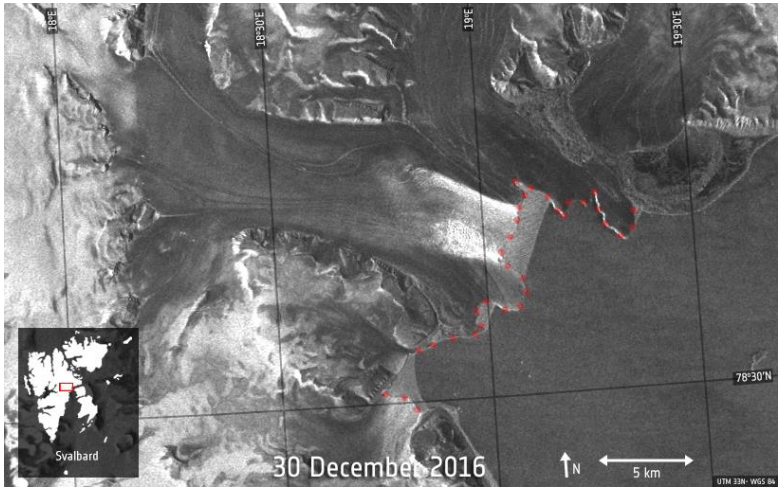
- Examples
  - DNA sequencing
  - Earth observation
- Large Hadron Collider at CERN
  - 300GB per second of raw data from detectors

e-Science example – Earth Observation

- Raw data cannot be interpreted by humans
- Software stack needed to interpret the data

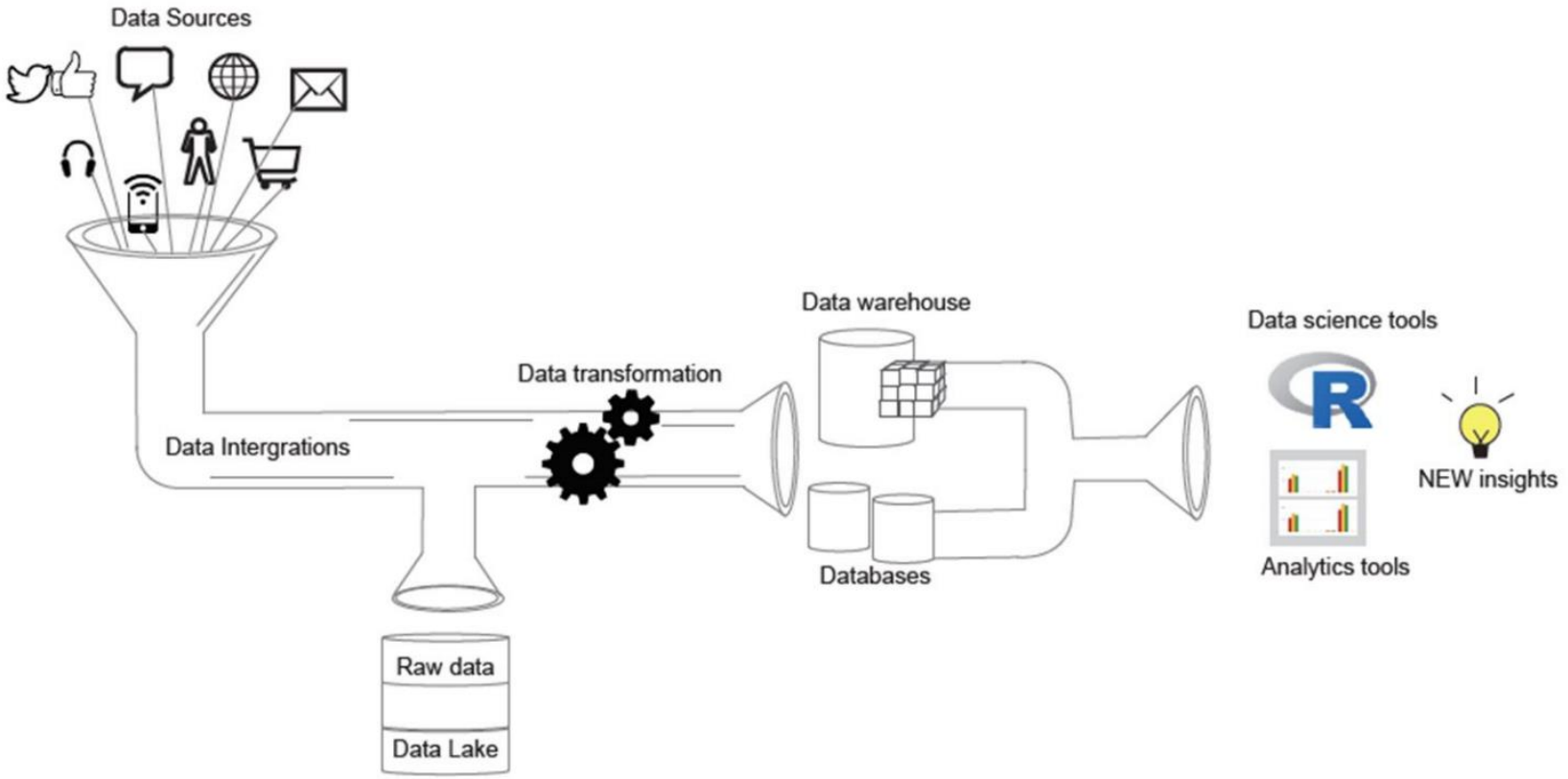






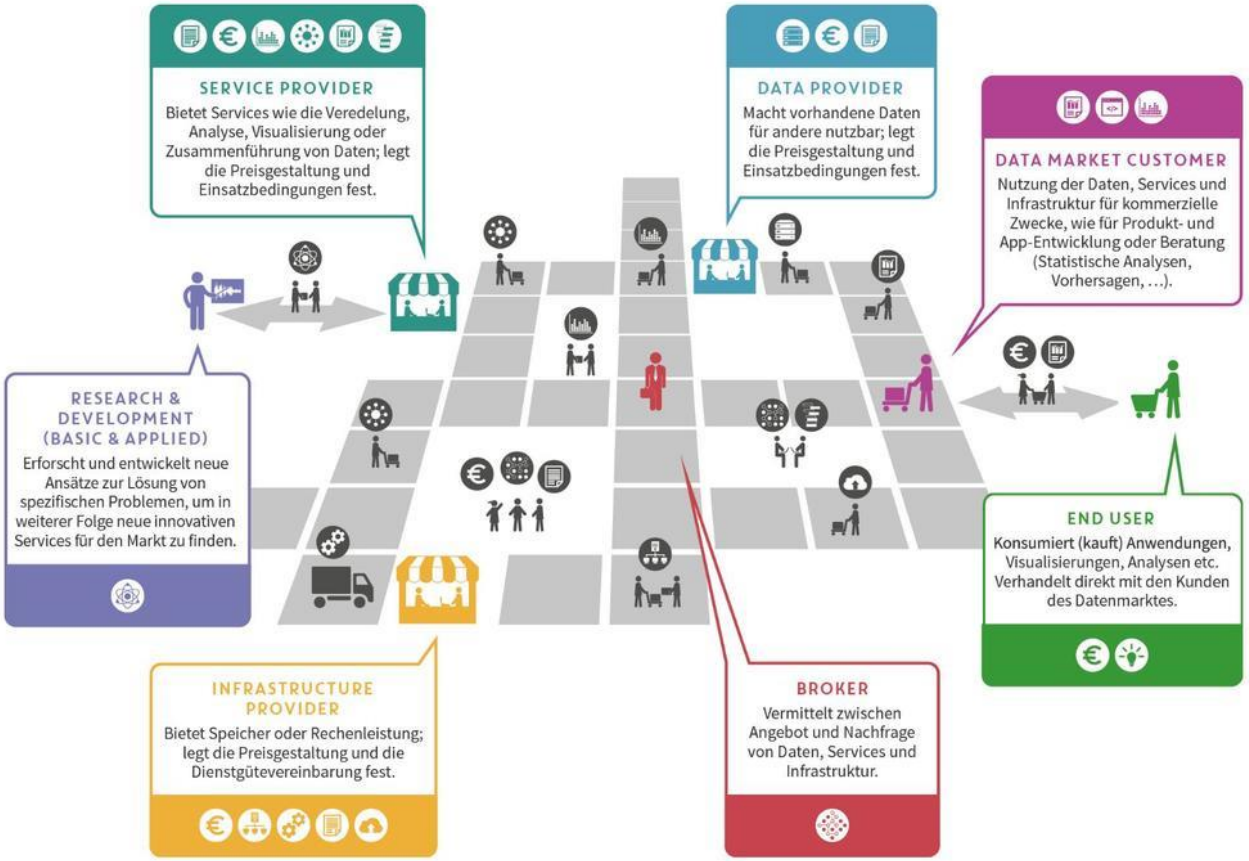
http://www.esa.int/spaceinimages/Images/2017/05/Negribreen_on_the_move

Data Science in Business Domains



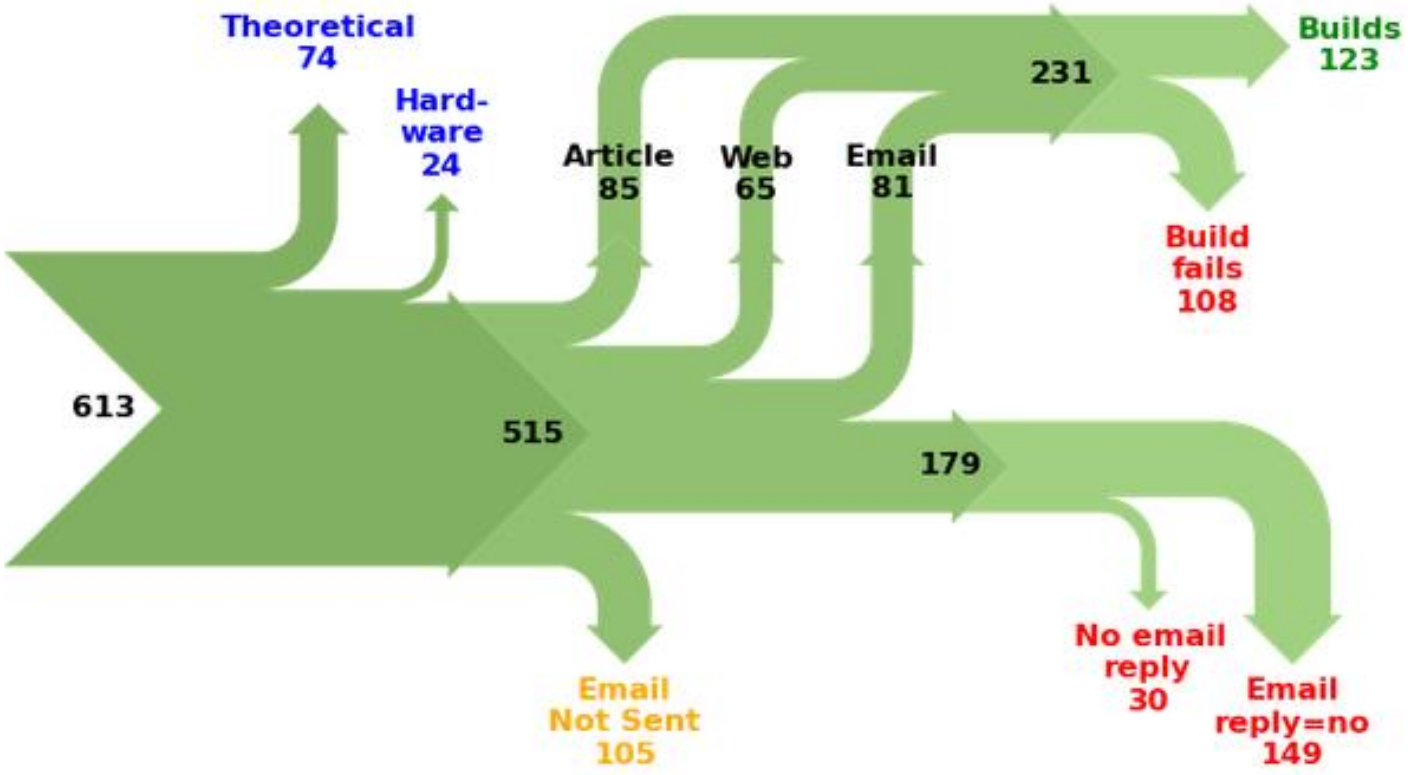
Alice Daish. (2017, February). Data-Driven Museums. Zenodo. http://doi.org/10.5281/zenodo.321812

Data Markets



Reproducibility
Computer Science

- 613 papers in 8 ACM conferences



C. Collberg and T. Proebsting, “Measuring reproducibility in computer systems research,” 2014. [Online]. Available: http://reproducibility.cs.arizona.edu/tr.pdf

Reproducibility
Computer Science

- E-mail responses from authors
  - Wrong version
  - Code will be available soon
  - Programmer left
  - Bad backup practices
  - Commercial code
  - Proprietary academic code
  - Intellectual property
  - No intention to release
  - …

Variety of solutions

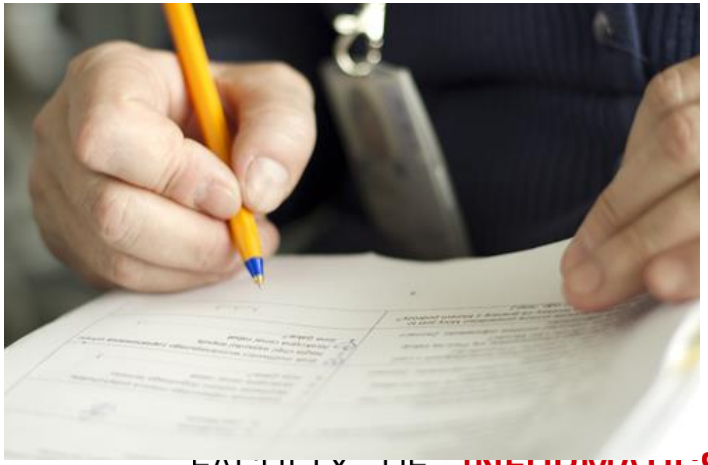
- In response to these needs many solutions were proposed and are being implemented
  - **open access** to scientific publications and data
  - research **data repositories** to host the data
  - **data citation** to reference the datasets
  - **data management plans**

What is a Data Management Plan?

Data Management Plan

- DMP is a formal document
- It outlines what you will do with your data **during** and **after** you complete your research
- It ensures your data is safe for the **present** and the **future**

[from University of Virginia Library]

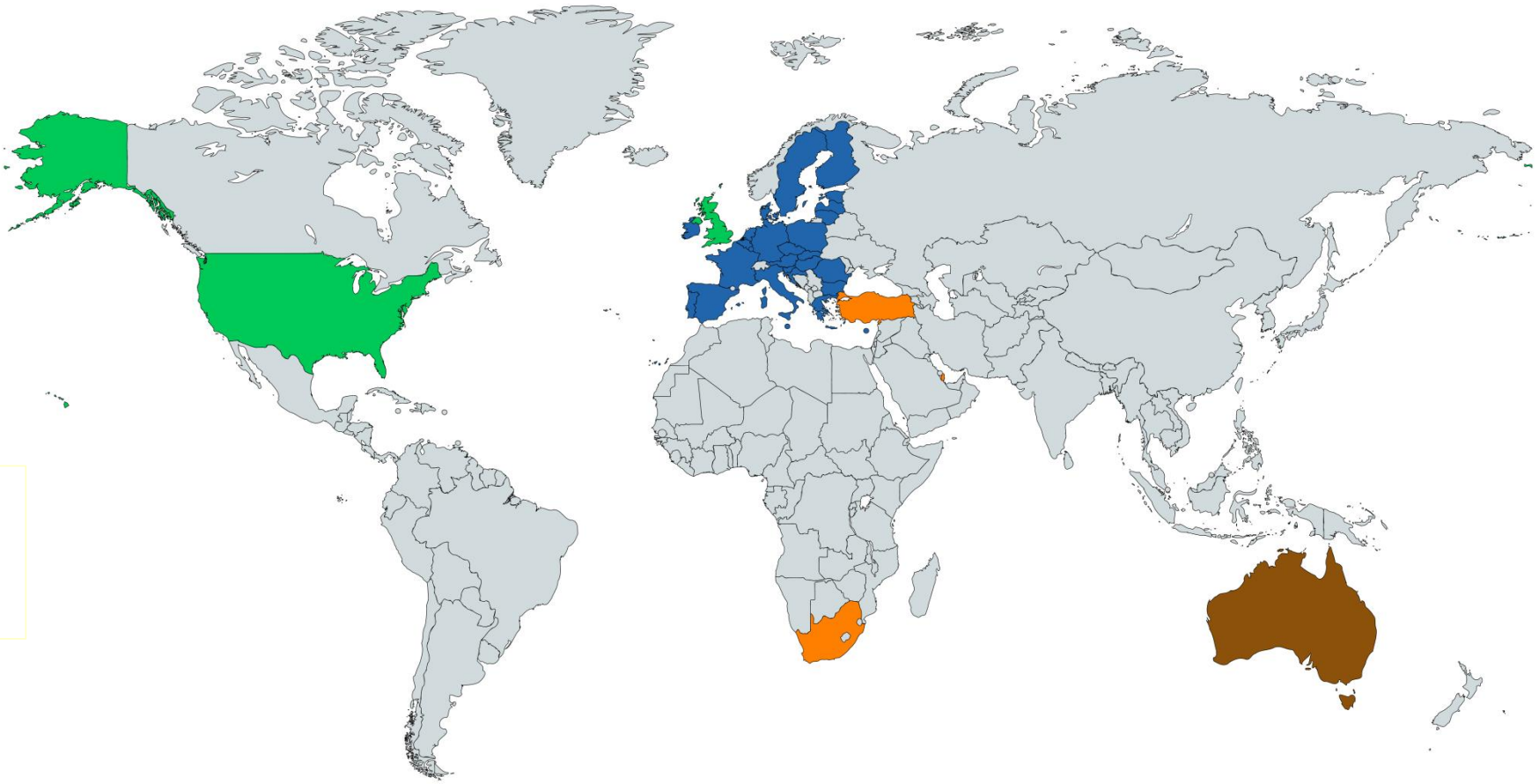


DMP is an awareness tool!

- DMP makes you think
  - what data you will use and where you get it from
  - what infrastructure, software, licenses are needed
  - what will be the output of your research
  - how you will share your research outputs
- DMP helps you organise yourself better

- DMP can reveal how solid your methodology is
  - is it a ‘fishing expedition’?

DMPs worldwide



DMPs in Austria

- Ongoing work on policies to require DMPs
  - FWF
    - https://www.fwf.ac.at/en/research-funding/open-access-policy/research-data-management/
  - Universities
    - http://www.e-infrastructures.at

https://www.tuwien.ac.at/fileadmin/t/ukanzlei/t-ukanzlei-english/Policy_for_Research_Data_Management.pdf

e·infra austria

FWF Der Wissenschaftsfonds.

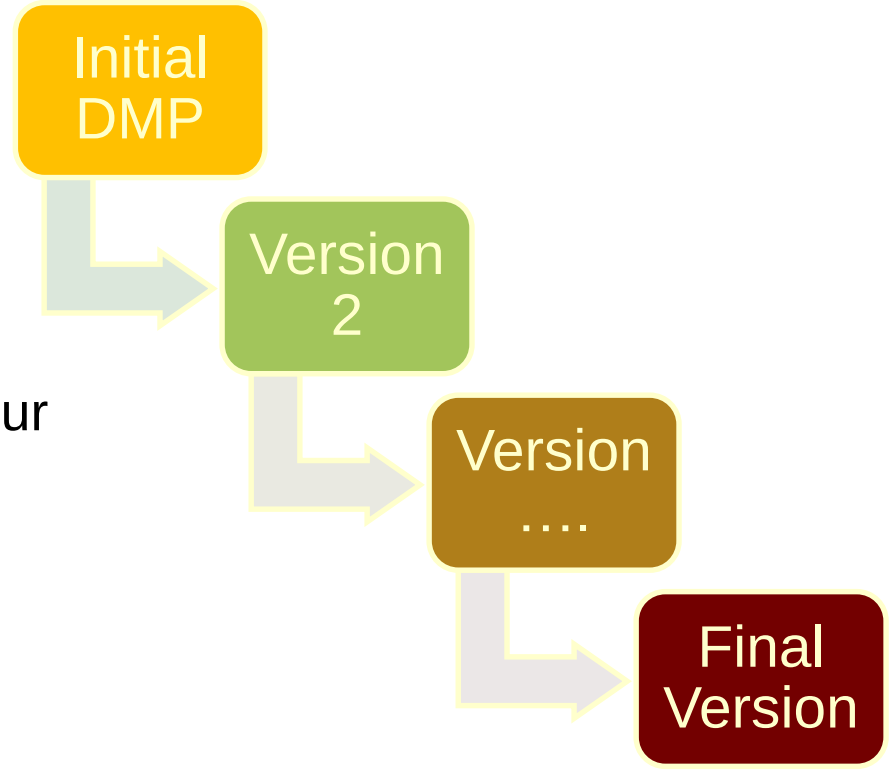
EC Horizon 2020 DMP Template

- Template is recommended but not required
  - 6 sections
  - 31 questions
  - Follows FAIR principles
    - Data must be **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable

DMP component	Issues to be addressed
1. Data summary	• State the purpose of the data collection/generation • Explain the relation to the objectives of the project • Specify the types and formats of data generated/collected • Specify if existing data is being re-used (if any) • Specify the origin of the data • State the expected size of the data (if known) • Outline the data utility: to whom will it be useful
2. FAIR Data	

EC Horizon 2020 DMP versions

- DMP is a living document
- First version
  - within the first 6 months
- Updated versions
  - when significant changes occur
    - new datasets
    - changes in policies
  - periodic reporting
    - project reviews
  - end of project



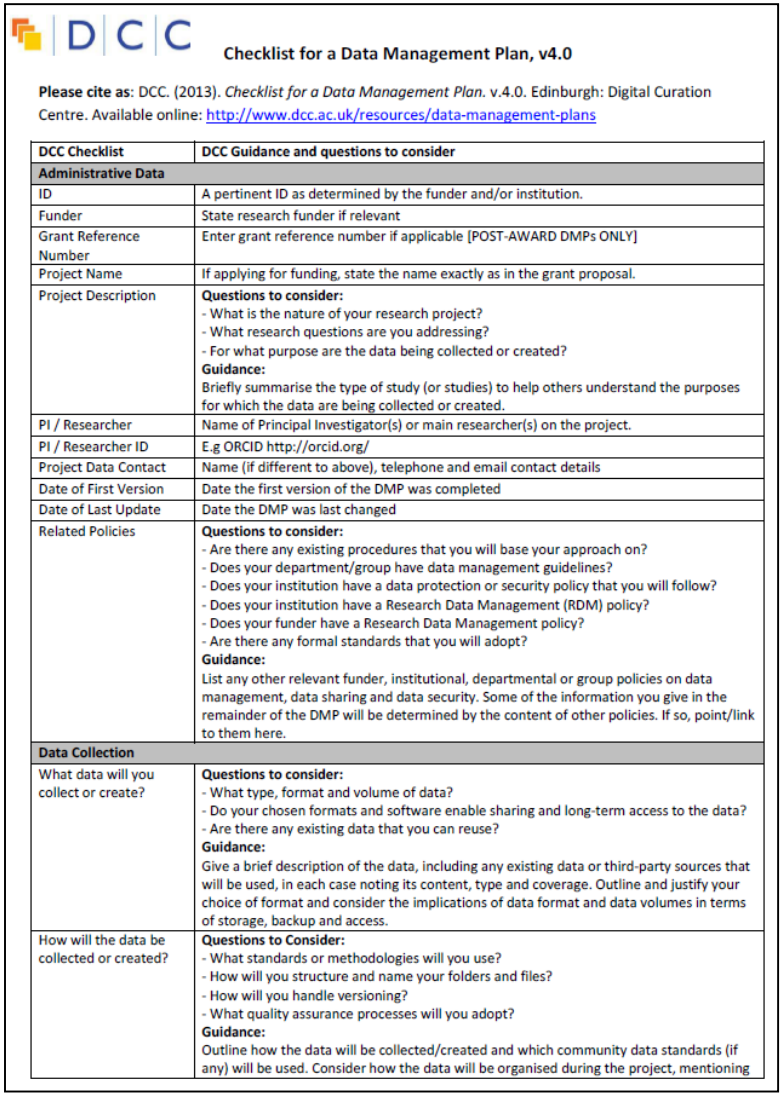
How to create a DMP?

How to create a DMP?

- Most cases by
  - filling out a template
  - answering questions from a checklist
- Using software tools
  - users choose appropriate funders template
  - only relevant questions and guidance is presented

DCC Checklist - example

- Synthesis of
  - funder requirements
  - institutional guidelines
  - good practice
- Contains 8 sections
  - the extent to which they need to be covered depends on a kind of research
- Resulting DMP
  - between a few paragraphs to a few pages long

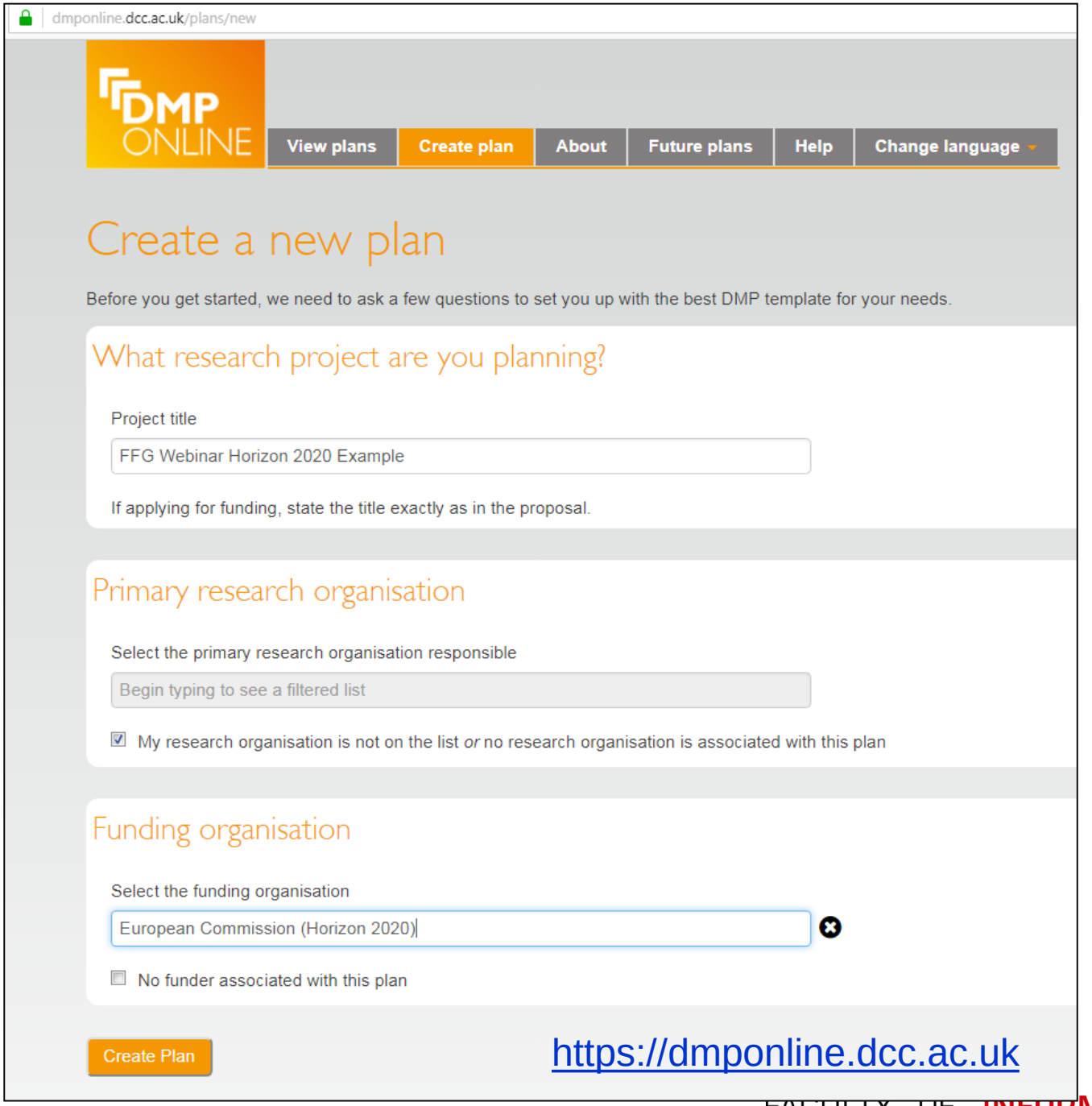


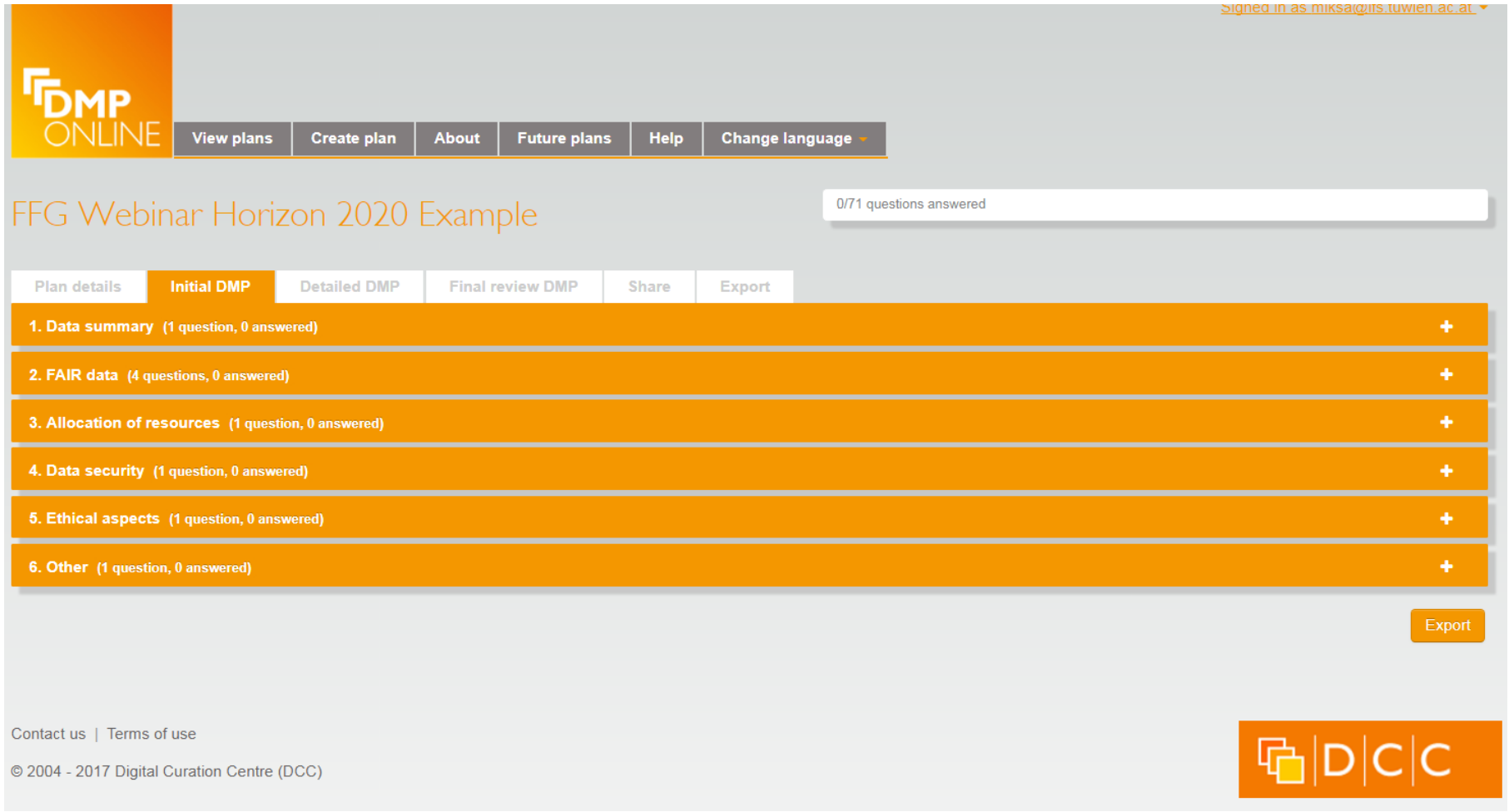
DCC

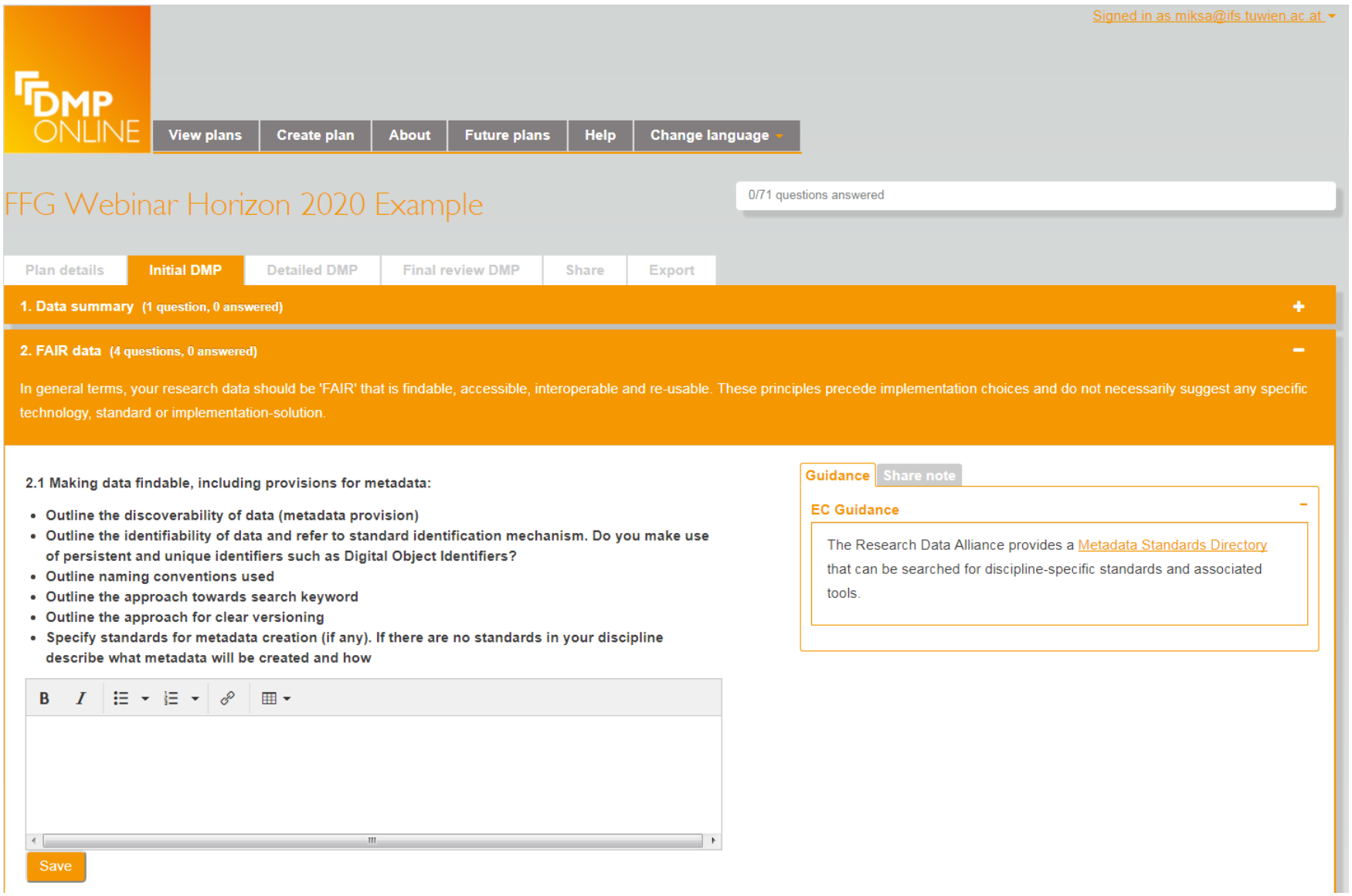
Checklist for a Data Management Plan, v4.0

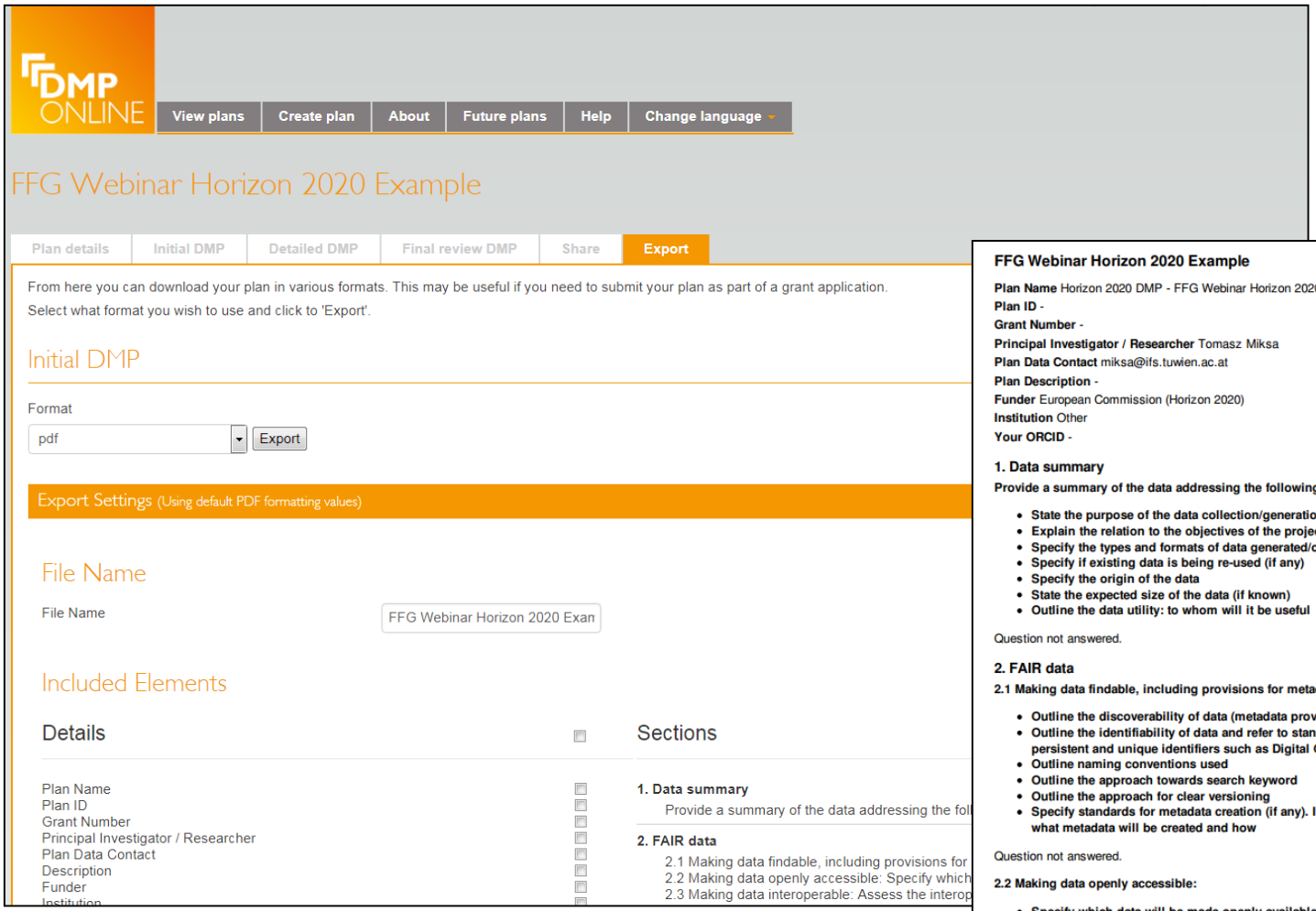
Please cite as: DCC. (2013). *Checklist for a Data Management Plan.* v.4.0. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/data-management-plans

DCC Checklist	DCC Guidance and questions to consider
Administrative Data	
ID	A pertinent ID as determined by the funder and/or institution.
Funder	State research funder if relevant
Grant Reference Number	Enter grant reference number if applicable [POST-AWARD DMPs ONLY]
Project Name	If applying for funding, state the name exactly as in the grant proposal.
Project Description	**Questions to consider:** - What is the nature of your research project? - What research questions are you addressing? - For what purpose are the data being collected or created? **Guidance:** Briefly summarise the type of study (or studies) to help others understand the purposes for which the data are being collected or created.
PI / Researcher	Name of Principal Investigator(s) or main researcher(s) on the project.
PI / Researcher ID	E.g ORCID http://orcid.org/
Project Data Contact	Name (if different to above), telephone and email contact details
Date of First Version	Date the first version of the DMP was completed
Date of Last Update	Date the DMP was last changed
Related Policies	**Questions to consider:** - Are there any existing procedures that you will base your approach on? - Does your department/group have data management guidelines? - Does your institution have a data protection or security policy that you will follow? - Does your institution have a Research Data Management (RDM) policy? - Does your funder have a Research Data Management policy? - Are there any formal standards that you will adopt? **Guidance:** List any other relevant funder, institutional, departmental or group policies on data management, data sharing and data security. Some of the information you give in the remainder of the DMP will be determined by the content of other policies. If so, point/link to them here.
Data Collection	
What data will you collect or create?	**Questions to consider:** - What type, format and volume of data? - Do your chosen formats and software enable sharing and long-term access to the data? - Are there any existing data that you can reuse? **Guidance:** Give a brief description of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage. Outline and justify your choice of format and consider the implications of data format and data volumes in terms of storage, backup and access.
How will the data be collected or created?	**Questions to Consider:** - What standards or methodologies will you use? - How will you structure and name your folders and files? - How will you handle versioning? - What quality assurance processes will you adopt? **Guidance:** Outline how the data will be collected/created and which community data standards (if any) will be used. Consider how the data will be organised during the project, mentioning


dmponline.dcc.ac.uk/plans/new
DMP ONLINE
View plans
Create plan
About
Future plans
Help
Change language
Create a new plan
Before you get started, we need to ask a few questions to set you up with the best DMP template for your needs.
What research project are you planning?
Project title
FFG Webinar Horizon 2020 Example
If applying for funding, state the title exactly as in the proposal.
Primary research organisation
Select the primary research organisation responsible
Begin typing to see a filtered list
My research organisation is not on the list or no research organisation is associated with this plan
Funding organisation
Select the funding organisation
European Commission (Horizon 2020)
No funder associated with this plan
Create Plan
https://dmponline.dcc.ac.uk


Signed in as miksa@ifs.tuwien.ac.at
DMP ONLINE
View plans
Create plan
About
Future plans
Help
Change language
FFG Webinar Horizon 2020 Example
0/71 questions answered
Plan details
Initial DMP
Detailed DMP
Final review DMP
Share
Export
1. Data summary (1 question, 0 answered)
2. FAIR data (4 questions, 0 answered)
3. Allocation of resources (1 question, 0 answered)
4. Data security (1 question, 0 answered)
5. Ethical aspects (1 question, 0 answered)
6. Other (1 question, 0 answered)
Export
Contact us | Terms of use
© 2004 - 2017 Digital Curation Centre (DCC)
DCC


Signed in as miksa@ifs.tuwien.ac.at
DMP ONLINE
View plans
Create plan
About
Future plans
Help
Change language
FFG Webinar Horizon 2020 Example
0/71 questions answered
Plan details
Initial DMP
Detailed DMP
Final review DMP
Share
Export
1. Data summary (1 question, 0 answered)
2. FAIR data (4 questions, 0 answered)
In general terms, your research data should be 'FAIR' that is findable, accessible, interoperable and re-usable. These principles precede implementation choices and do not necessarily suggest any specific technology, standard or implementation-solution.
2.1 Making data findable, including provisions for metadata:
Outline the discoverability of data (metadata provision)
Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?
Outline naming conventions used
Outline the approach towards search keyword
Outline the approach for clear versioning
Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how
Guidance
Share note
EC Guidance
The Research Data Alliance provides a Metadata Standards Directory that can be searched for discipline-specific standards and associated tools.
Save



FFG Webinar Horizon 2020 Example

Plan Name Horizon 2020 DMP - FFG Webinar Horizon 2020 Example
Plan ID -
Grant Number -
Principal Investigator / Researcher Tomasz Miksa
Plan Data Contact miksa@ifs.tuwien.ac.at
Plan Description -
Funder European Commission (Horizon 2020)
Institution Other
Your ORCID -

1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

Question not answered.

2. FAIR data

2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

Question not answered.

2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

Question not answered.

Other DMP checklists examples

- Science Europe (‘federation of research funders’)

https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf

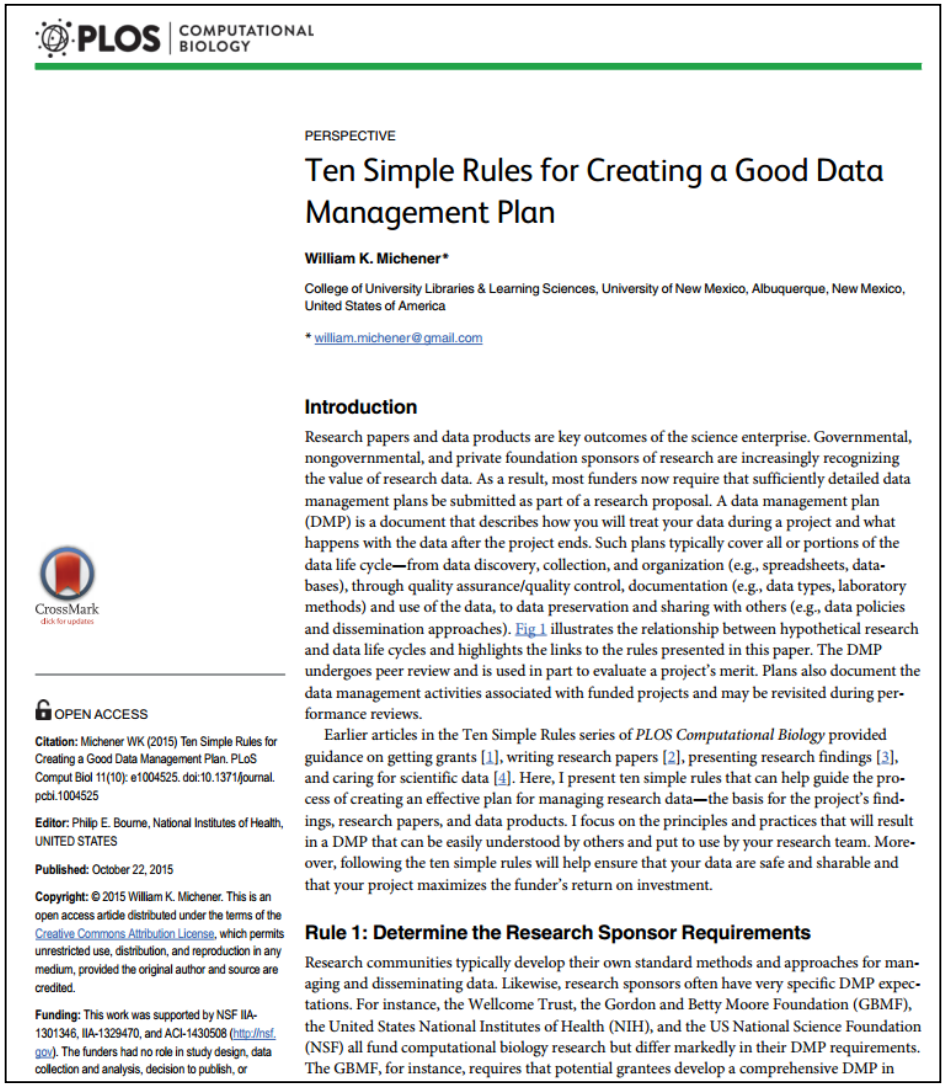
- FWF Austria (based on the one above)

https://www.fwf.ac.at/fileadmin/files/Dokumente/Open_Access/FWF_DMPTemplate_e.pdf

Other DMP tools

- Comprehensive list can be found here
  - https://activedmps.org
- Data Stewardship Wizard
  - Detailed questions, less free form text
  - https://ds-wizard.org
- RDMO
  - https://rdmorganiser.github.io/en/

PLOS – Ten Simple Rules



PLOS | COMPUTATIONAL BIOLOGY

PERSPECTIVE

Ten Simple Rules for Creating a Good Data Management Plan

William K. Michener*

College of University Libraries & Learning Sciences, University of New Mexico, Albuquerque, New Mexico, United States of America

* william.michener@gmail.com

Introduction

Research papers and data products are key outcomes of the science enterprise. Governmental, nongovernmental, and private foundation sponsors of research are increasingly recognizing the value of research data. As a result, most funders now require that sufficiently detailed data management plans be submitted as part of a research proposal. A data management plan (DMP) is a document that describes how you will treat your data during a project and what happens with the data after the project ends. Such plans typically cover all or portions of the data life cycle—from data discovery, collection, and organization (e.g., spreadsheets, databases), through quality assurance/quality control, documentation (e.g., data types, laboratory methods) and use of the data, to data preservation and sharing with others (e.g., data policies and dissemination approaches). Fig 1 illustrates the relationship between hypothetical research and data life cycles and highlights the links to the rules presented in this paper. The DMP undergoes peer review and is used in part to evaluate a project's merit. Plans also document the data management activities associated with funded projects and may be revisited during performance reviews.

Earlier articles in the Ten Simple Rules series of *PLOS Computational Biology* provided guidance on getting grants [1], writing research papers [2], presenting research findings [3], and caring for scientific data [4]. Here, I present ten simple rules that can help guide the process of creating an effective plan for managing research data—the basis for the project's findings, research papers, and data products. I focus on the principles and practices that will result in a DMP that can be easily understood by others and put to use by your research team. Moreover, following the ten simple rules will help ensure that your data are safe and sharable and that your project maximizes the funder's return on investment.

Rule 1: Determine the Research Sponsor Requirements

Research communities typically develop their own standard methods and approaches for managing and disseminating data. Likewise, research sponsors often have very specific DMP expectations. For instance, the Wellcome Trust, the Gordon and Betty Moore Foundation (GBMF), the United States National Institutes of Health (NIH), and the US National Science Foundation (NSF) all fund computational biology research but differ markedly in their DMP requirements. The GBMF, for instance, requires that potential grantees develop a comprehensive DMP in

CrossMark
click for updates

OPEN ACCESS

Citation: Michener WK (2015) Ten Simple Rules for Creating a Good Data Management Plan. PLoS Comput Biol 11(10): e1004525. doi:10.1371/journal.pcbi.1004525

Editor: Philip E. Bourne, National Institutes of Health, UNITED STATES

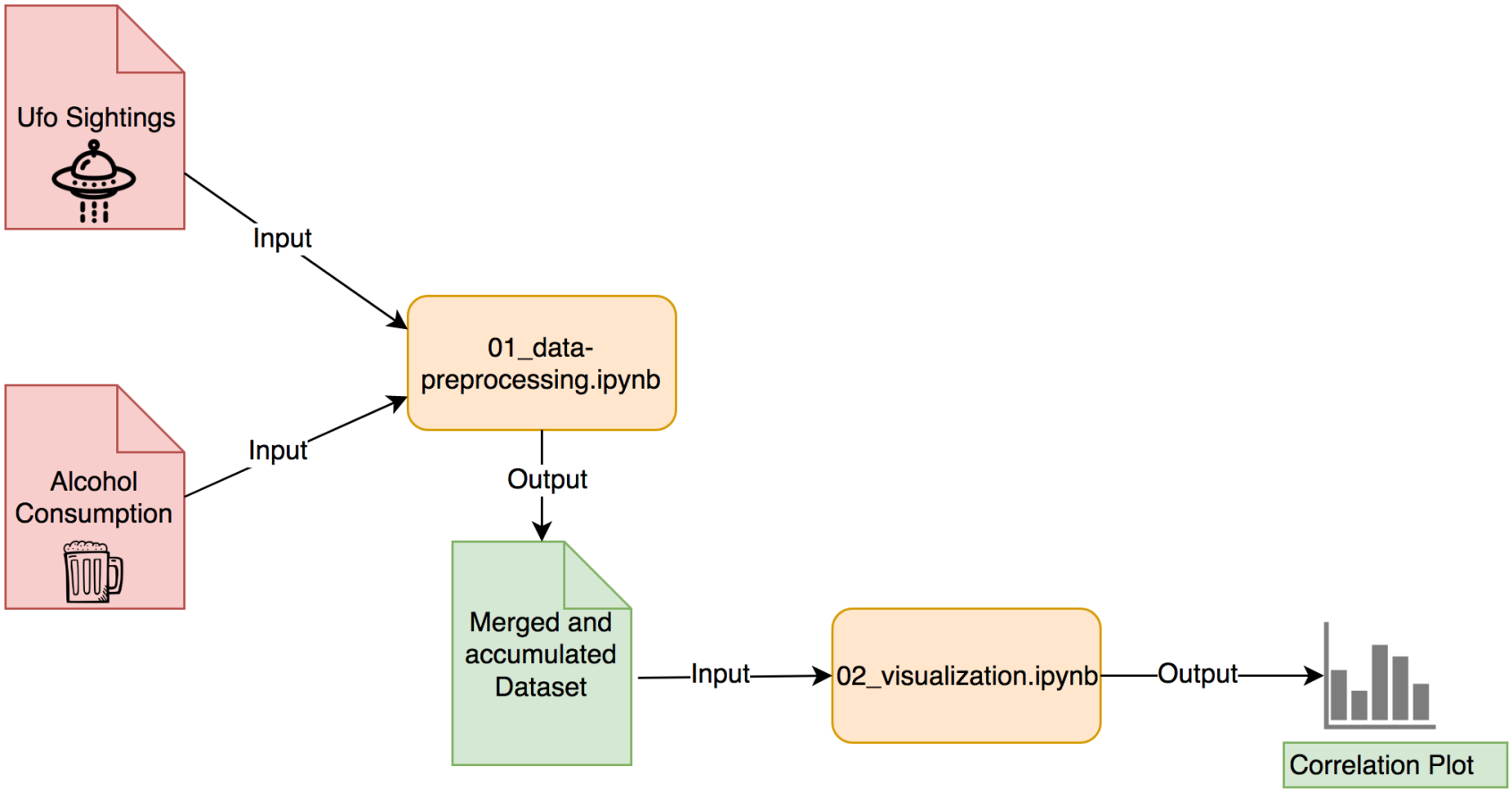
Published: October 22, 2015

Funding: This work was supported by NSF IIA-1301346, IIA-1329470, and ACI-1430508 (http://nsf.gov). The funders had no role in study design, data collection and analysis, decision to publish, or

http://dx.doi.org/10.1371/journal.pcbi.1004525

What should I write in fact?

Correlating Alcohol Consumption and UFO Sightings in the USA (running example)



https://github.com/mdietrichstein/digitalpreservation-dmp

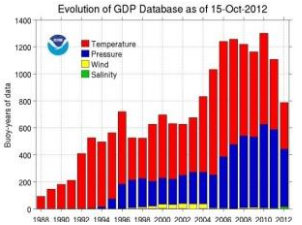
Most templates require

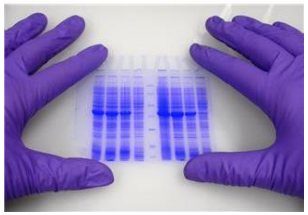
1. Data Summary
2. FAIR data
3. Allocation of resources
4. Data Security
5. Ethical aspects
6. Other issues

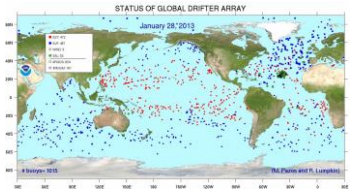
DATA SUMMARY

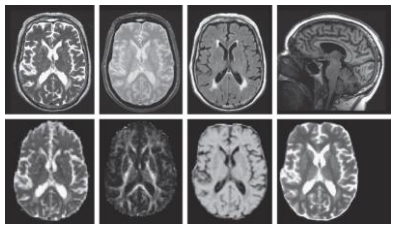
What is data?

- Instrument measurements
- Experimental observations
- Still images, video and audio
- Text documents, spreadsheets, databases
- Quantitative data (e.g. survey data)
- Survey results & interview transcripts
- Simulation data, models & software
- Slides, artefacts, specimens, samples
- Questionnaires
- Sketches, diaries, lab notebooks …









Data Summary

- Type
  - text, spreadsheets, software, models, images, movies, audio, patient records, etc.
- Source
  - human observation, laboratory, field instruments, experiments, simulations, compilations, etc.
- Volume
  - total volume of data, number of files, etc.
- Data and file formats
  - non-proprietary formats
  - used within community

Data Summary - example

Produced Data

This project produces aggregated dataset in CSV format (Filesize ~800K) that contains data points that combine alcohol consumption data with the UFO sighting data and a correlation plot of these in PNG format (Filesize (~100K)).

Input Data

Project accesses two external CSV datasets. Both datasets have been downloaded and saved along with the source code in the folder *data/raw*.

1. Alcohol Consumption: OECD (2018), Alcohol consumption (indicator).

   DOI: 10.1787/e6895909-en (Accessed on 22 March 2018)

   File Location: data/raw/DP_LIVE_22032018202902423.csv

   File Size: 112K

Data Summary - example

- The experiment has been conducted with Jupyter notebooks. The notebooks contain the experiment's code, accompanying documentation, tables and plots.

- We have included instructions (README.md) on how to run the experiment either directly or via Docker.

Running the code

To run the code in this repository you will need to have access to a machine running `python` (at least version `3.5`) and pip.

Run `pip install -r requirements.txt` to install the required dependencies.

Once the dependencies have been installed, start the jupyter notebook server via `jupyter notebook` and open `http://localhost:8888`.

In the `notebooks` folder you'll find the following notebooks:

01_data-preprocessing.ipynb

Running this notebook generates a dataset consisting of the number of ufo sightings and the alcohol consumption in the usa per year by preprocessing and accumulating the data provided by the datasources mentioned above.

FAIR PRINCIPLES

FAIR Principles

- **F**indable
  - contains metadata that facilitates search
- **A**ccessible
  - access conditions are specified
  - software needed to interpret data is known
- **I**nteroperable
  - Follow standards and domain specific conventions
- **R**eusable
  - clear license and documentation
  - ‘sum of the three other rules’
- There is no clear distinction between principles
  - e.g. metadata supports all of them
- FAIR is not equivalent with ‘open data’

https://www.nature.com/articles/sdata201618

Read more about FAIR

- Principles explained (by their authors)
  - https://www.nature.com/articles/sdata201618
  - https://www.go-fair.org/fair-principles/
- Watch (why FAIR matters)
  - https://vimeo.com/143245835
- FAIR underlies European Open Science Cloud
  - https://eosc-launch.eu/declaration/

METADATA

What is in the picture?

Metadata – Atlas Of Living Australia

Atlas Of Living Australia | ALA Apps | ALA Info | Search the Atlas | Search

NatureShare - 2380_Gymnorhina_tibicen

HumanObservation of *Cracticus tibicen* | Australian Magpie recorded on 2011-04-17T12:32:00+1000

Flag an issue | Contact curator

- Dataset
- Event
- Taxonomy
- Geospatial
- Images
- Data quality tests (1 4 21 13 48)
- Additional political boundaries information
- Environmental sampling for this location

Location of record

Images

Photographer: Russell Best

Dataset

Data resource	NatureShare
Catalogue number	2380_Gymnorhina_tibicen
Basis of record	Human observation
Observer	Best, R. Russell *Supplied as "Russell Best"*
Rights	CC BY 2.5 AU
More details	http://natureshare.org.au/observation/2380/
Photographer	Russell Best
Rightsholder	Russell Best via NatureShare
Occurrence remarks	Tags: Female
Occurrence status	present
Abcd identification qualifier	Not provided

Event

Record date	[date not supplied] *Supplied date "2011-04-17T12:32:00+1000"*
Event remarks	Photo date/time used.

Taxonomy

Scientific name	*Cracticus tibicen* *Supplied scientific name "Gymnorhina tibicen"*
Taxon rank	Species
Common name	Australian Magpie
Kingdom	Animalia
Phylum	Chordata
Class	Aves
Order	Passeriformes
Family	Artamidae
Genus	*Cracticus*
Species	*Cracticus tibicen*

http://biocache.ala.org.au/occurrences/544b0271-5f04-47ab-9d8b-0dbe3b5f59d7

Metadata – Atlas Of Living Australia

Dataset

Data resource	NatureShare
Catalogue number	2380_Gymnorhina_tibicen
Basis of record	Human observation
Observer	Best, R. Russell *Supplied as "Russell Best"*
Rights	CC BY 2.5 AU
More details	http://natureshare.org.au/observation/2380/
Photographer	Russell Best
Rightsholder	Russell Best via NatureShare
Occurrence remarks	Tags: Female
Occurrence status	present
Abcd identification qualifier	Not provided

Metadata – Atlas Of Living Australia

Event

Record date	[date not supplied] *Supplied date "2011-04-17T12:32:00+1000"*
Event remarks	Photo date/time used.

Taxonomy

Scientific name	*Cracticus tibicen* *Supplied scientific name "Gymnorhina tibicen"*
Taxon rank	Species
Common name	Australian Magpie
Kingdom	Animalia
Phylum	Chordata
Class	Aves
Order	Passeriformes
Family	Artamidae
Genus	*Cracticus*
Species	*Cracticus tibicen*
Taxonomic issues	No issues
Name match metric	Exact match The supplied name matched the name exactly.

Metadata – Atlas Of Living Australia

Geospatial

Country	Australia
State or territory	Victoria
Local government area	Macedon Ranges (S)
Latitude	-37.421078
Longitude	144.61954
Geodetic datum	EPSG:4326
Biome	Terrestrial
Verbatim longitude	144.619541
Verbatim latitude	-37.421077

Location of record

Standards and metadata

- Metadata
  - helps to understand and interpret data
  - provides details about experiment setup
    - who, when, in which conditions, tools, versions, etc.
  - helps identify and discover new data
- Use community standards to enable interoperability

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

Metadata - example

- The metadata file can be found inside the project folder (/documentation/metadata.xml).
  - experiment title, authors, date, tools, coverage, rights, etc.
- Additionally a descriptive file is added, which explains the axes and units used in the output files, this file can be found inside the project folder as well (/documentation/description.txt).
  - The alcohol consumption is the average consumption rate in liters/capita of USA inhabitants, which are older than 17.

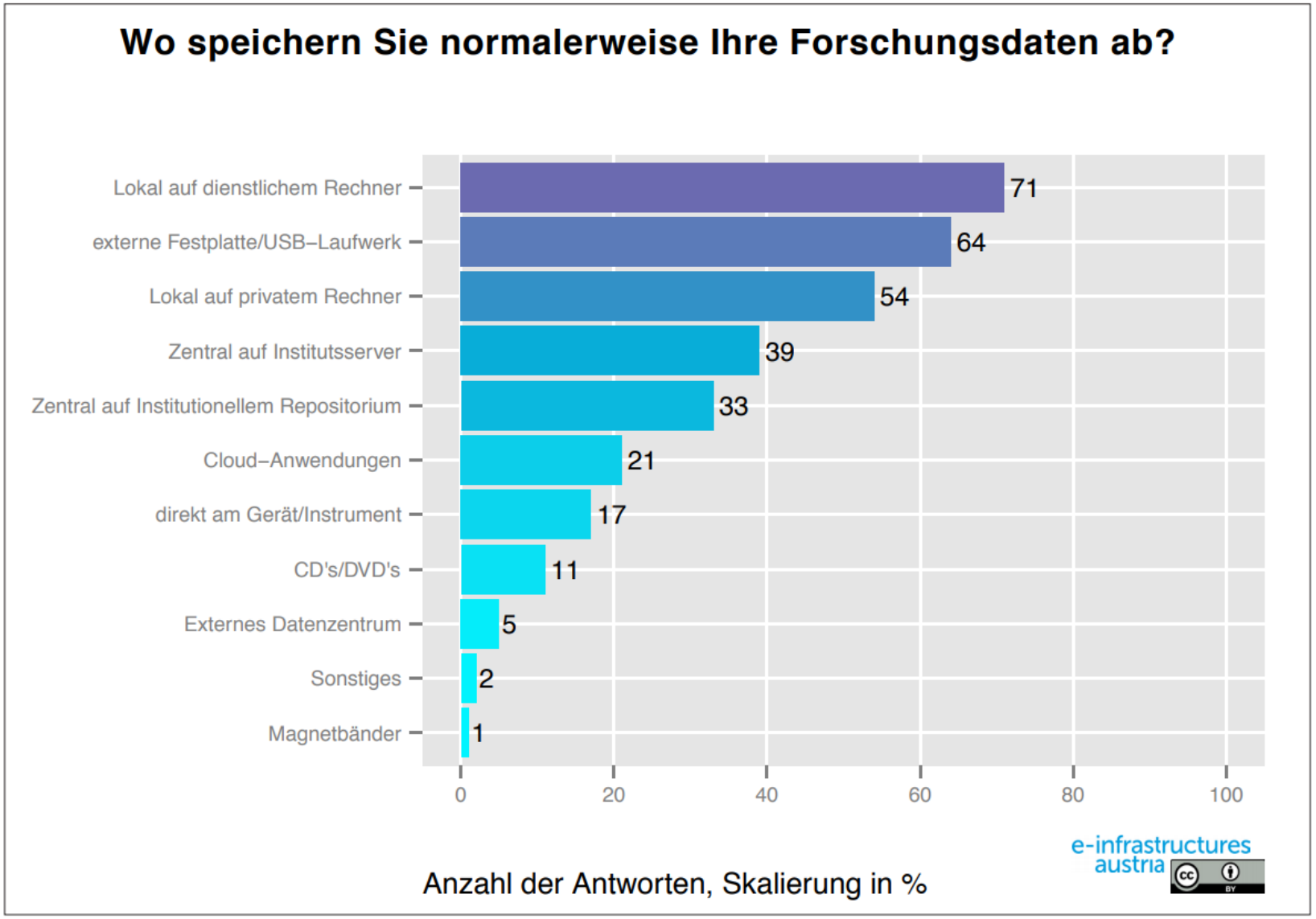
Metadata-example

```
<?xml version="1.0" encoding="UTF-8"?>

<metadata xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://pur
<dc:title>UFOs & Alcohol</dc:title>
<dc:creator>Marc Dietrichstein (https://orcid.org/0000-0003-4890-3498)</dc:creator>
<dc:creator>Markus Neumeyer (https://orcid.org/0000-0002-4081-0716)</dc:creator>
<dc:subject>Correlation of alcohol consumtion and UFO sightings</dc:subject>
<dc:description>Automated tool that invastigates and computes the correlation betwee
<dc:date>23.03.2018</dc:date>
<dc:type>DataGeneration</dc:type>
<dc:format>Jupyternotebook</dc:format>
<dc:source>Ufo Sightings: Sigmond Axel. (2014)</dc:source>
<dc:source>Alcohol Consumption: OECD (2018)</dc:source>
<dc:language>English</dc:language>
<dc:coverage>1960 - 2014 </dc:coverage>
<dc:rights>Free access</dc:rights>
```

MANAGING & SHARING (DURING PROJECT)

Managing data during research



http://phaidra.univie.ac.at/o:407513

Managing data during research

- If you loose your data there will be nothing to share!
- Recreating or recollecting data can be
  - impossible
    - e.g. observational data
  - too expensive
    - e.g. cost of computational power
- How do you manage data during the project?
  - file naming convention
  - versioning
  - backups
  - should the access be restricted?
  - who is responsible?

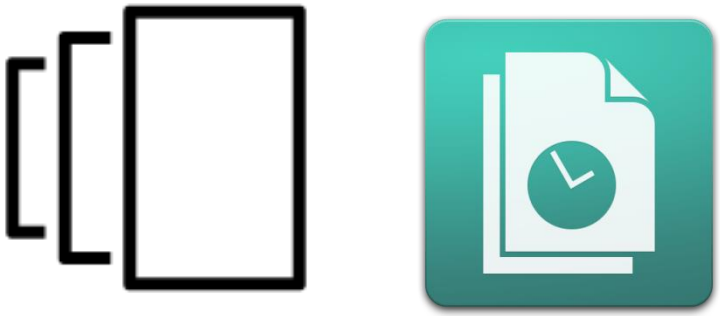
Data sharing - example

- Code and data are hosted in a public git repository on GitHub.
- Read access is open to everyone. Write-access is limited to the researchers working on the project.
- Permissions are managed via Github's account system using SSH keys.

ARCHIVING AND PRESERVATION

Backup vs archiving and preservation

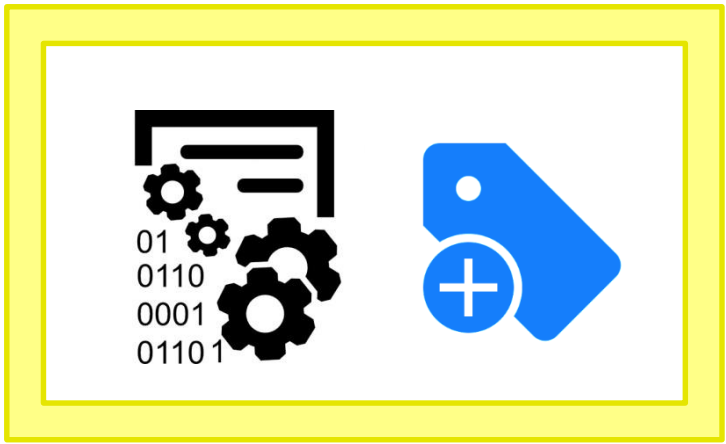
Storing and backing up files while research is active



Likely to be on a networked filestore or hard drive

Easy to change or delete

Archiving or preserving data in the long-term



Likely to be deposited in a digital repository

Safeguarded and preserved

Archiving and preservation

- Which data will be shared?
  - What has to be kept?
  - What can’t be recreated?
  - What is potentially useful to others?
  - What legally must be destroyed?
- Where will the data be deposited?
  - not all of the data must be shared in the same way
- Are there any embargo periods?
- For how long?
- What is the cost and who will pay for it?
- Which license to use?

Archiving and preservation - example

- The following files are relevant to reproduce the experiment and should be preserved
  - *README.md* – Text file containing instructions on how to run the experiment
  - Both *Jupyer notebooks* - The experiment's code and documentation
  - *Dockerfile* - To build a docker container for running the experiment
  - *requirements.txt* - List of python dependencies required by the experiment
  - *documentation/architecture.pn*g - Architectural diagram of the experiment
  - *documentation/description.txt* – Text file describing the correlation plot's content
  - *documentation/metadata.xml* - Metadata relevant to the experiment

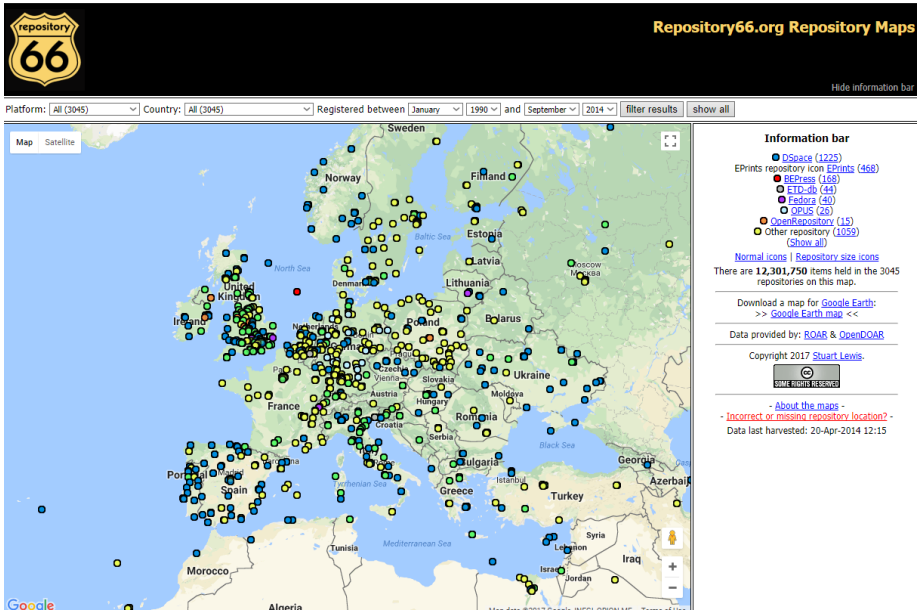
- Note: input data was not selected for preservation
  - it is maintained by existing repository (*easy* to get)

Where to find a repository?

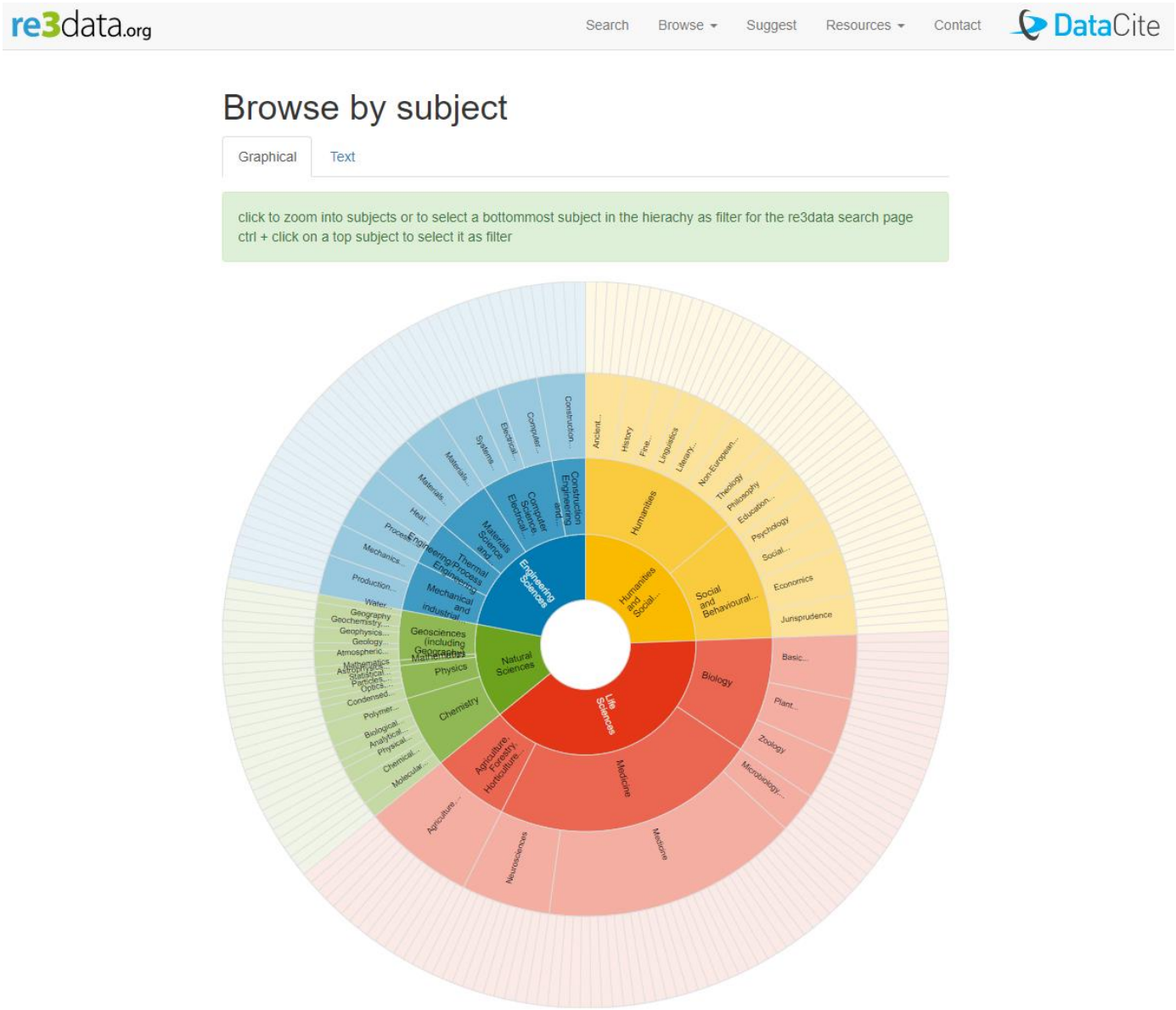
- More information: https://www.openaire.eu/opendatapilot-repository
- Zenodo: http://www.zenodo.org
- Re3data.org: http://www.re3data.org

Repository registries

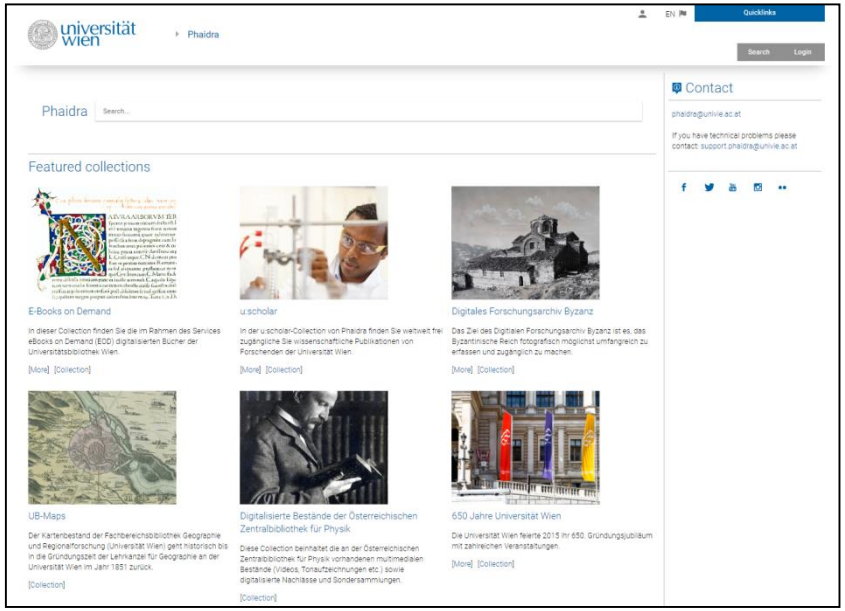
- Directory of Open Access Repositories – DOAR
  - http://www.opendoar.org/
- Registry of Open Access Repositories – ROAR
  - http://roar.eprints.org/
- Projection of DOAR and ROAR onto google maps
  - http://maps.repository66.org



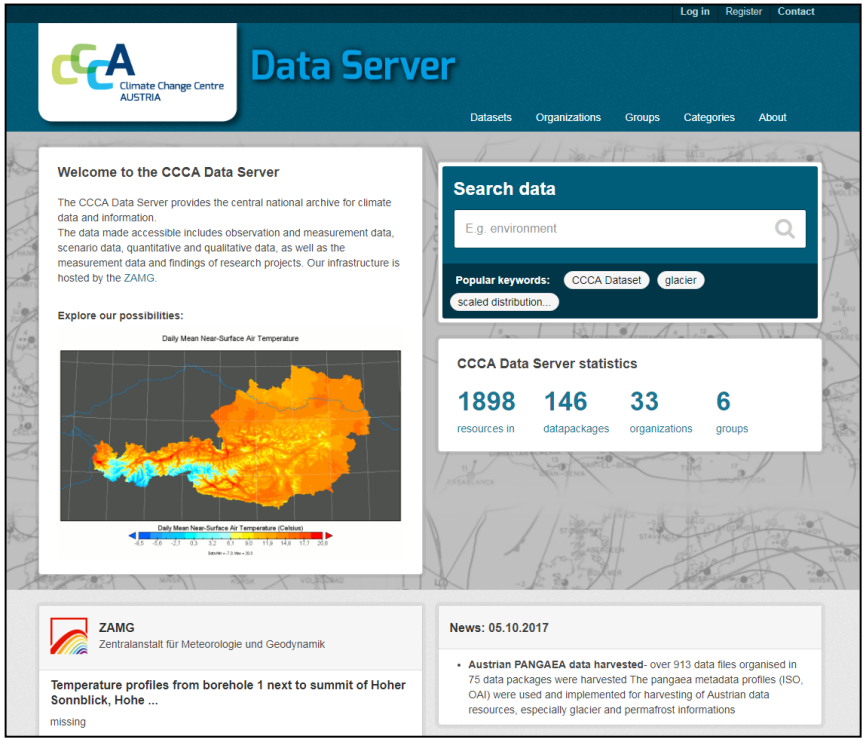
re3data.org


re3data.org
Search
Browse
Suggest
Resources
Contact
DataCite
Browse by subject
Graphical
Text
click to zoom into subjects or to select a bottommost subject in the hierachy as filter for the re3data search page
ctrl + click on a top subject to select it as filter
Humanities and Social...
Engineering Sciences
Natural Sciences
Life Sciences
Humanities
Social and Behavioural...
Construction Engineering and...
Computer Science, Electrical...
Materials Science and...
Thermal Engineering/Process Engineering
Mechanical and Industrial...
Geosciences (including Geography)
Physics
Chemistry
Biology
Medicine
Agriculture, Forestry, Horticulture
Neurosciences
Ancient...
History
Fine...
Linguistics
Literary...
Non-European...
Theology
Philosophy
Education...
Psychology
Social...
Economics
Jurisprudence
Basic...
Plant...
Zoology
Microbiology...
Construction...
Computer...
Electrical
Systems...
Materials...
Heat...
Process...
Mechanics...
Production...
Water
Geography
Geochemistry...
Geophysics...
Geology...
Atmospheric...
Mathematics
Astronomy...
Statistical...
Particles...
Optics...
Condensed...
Polymer...
Biological...
Analytical...
Physical...
Chemical...
Molecular...
Agriculture...

Repositories in Austria - examples



https://phaidra.univie.ac.at

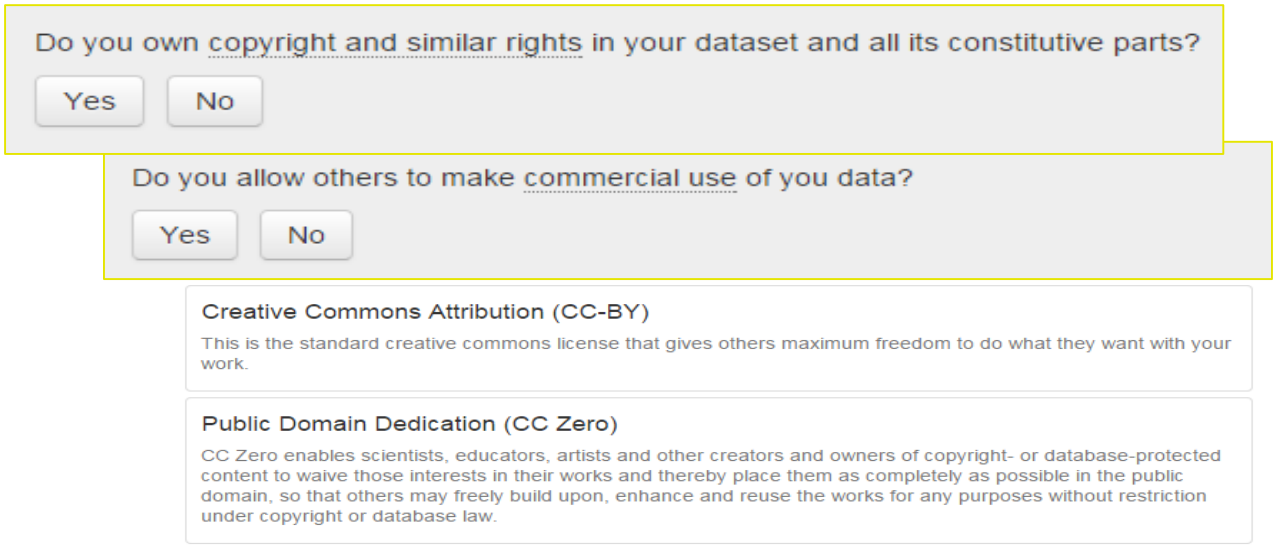


https://data.ccca.ac.at

Licenses

- Horizon 2020 guidelines point to CC-BY or CC-0
- DCC How-to guide helps you to license data

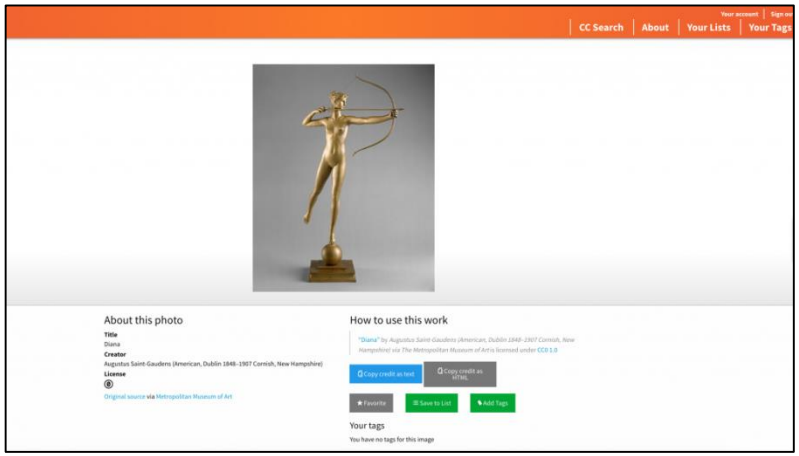
  www.dcc.ac.uk/resources/how-guides/license-research-data
- EUDAT licensing wizard help you pick licence for data & software



http://ufal.github.io/public-license-selector

Licenses

- CC-0
  - waives creator rights -> public domain
  - allows anyone to use, re-use, and remix a work without restriction
  - Example: all images from Metropolitan Museum of Art in New York https://creativecommons.org/2017/02/07/met-announcement/



Creative Commons

- CC-BY (Attribution)
  - allows anyone to use, re-use, and remix a work without restriction, also commercially
  - You must give appropriate credit, provide a link to the license, and indicate if changes were made.

- CC BY-SA (Attribution – ShareAlike)
  - all new works must care the same license

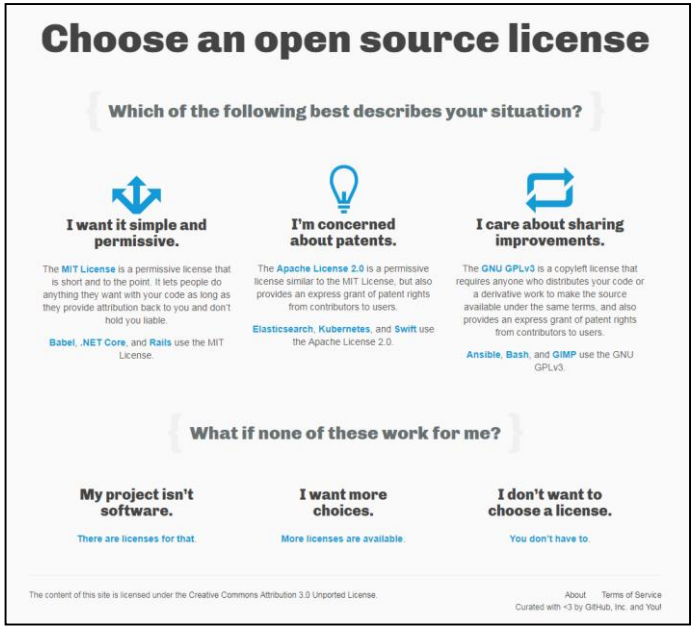
- CC BY-ND (Attribution- NoDerivs)
  - reuse, but no changes

- CC BY-NC
  - no commercial use

- CC BY-NC-SA
- CC BY-NC-ND

Software Licenses

- Choose correct license for your software
  - Apache, MIT, GNU, BSD, …
- Check licenses of libraries you reuse in your software
  - Example: GNU GPL vs GNU **L**GPL
    - GPL enforces the reusing software to be GPL (also public)
    - LGPL code must be clearly marked, rest of the software can have different license (can be private)
- Software licenses can also be used for data



https://choosealicense.com

License - example

- The external datasets are using permissible licenses which allows us the usage and redistribution of the following data:
  - * Ufo Sightings - Creative Commons Attribution 4.0
  - * Alcohol Consumption - Free to use and distribute according to http://www.oecd.org/termsandconditions/ -Section C - Permitted use
- All code, data and documentation is available on Github and is licensed under the MIT license.

Persistent Identifiers

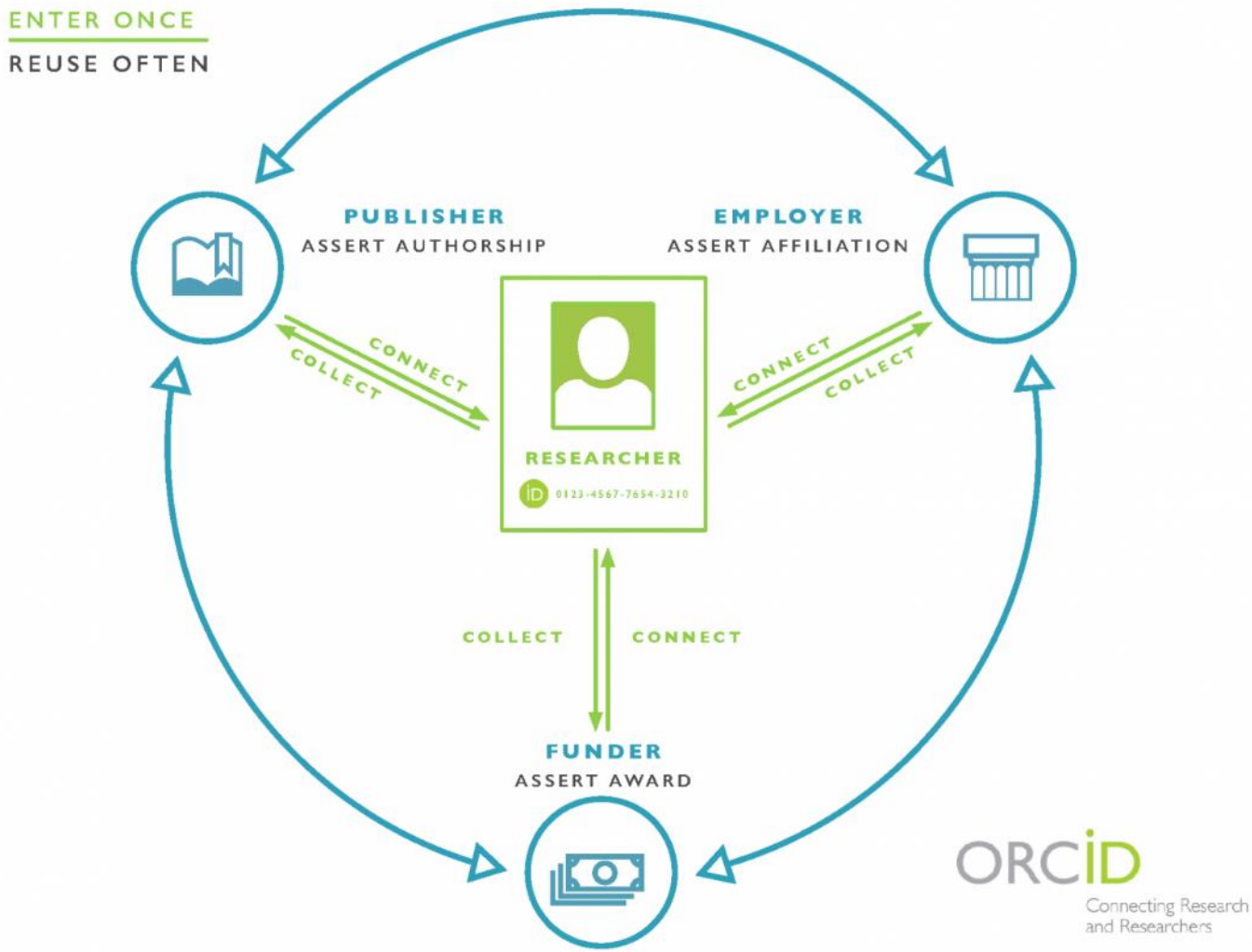
- Digital Object Identifier (DOI)
  - Uniquely identify objects
  - DOI assigned once
  - Physical location of data can change

10.5281/zenodo.1068223

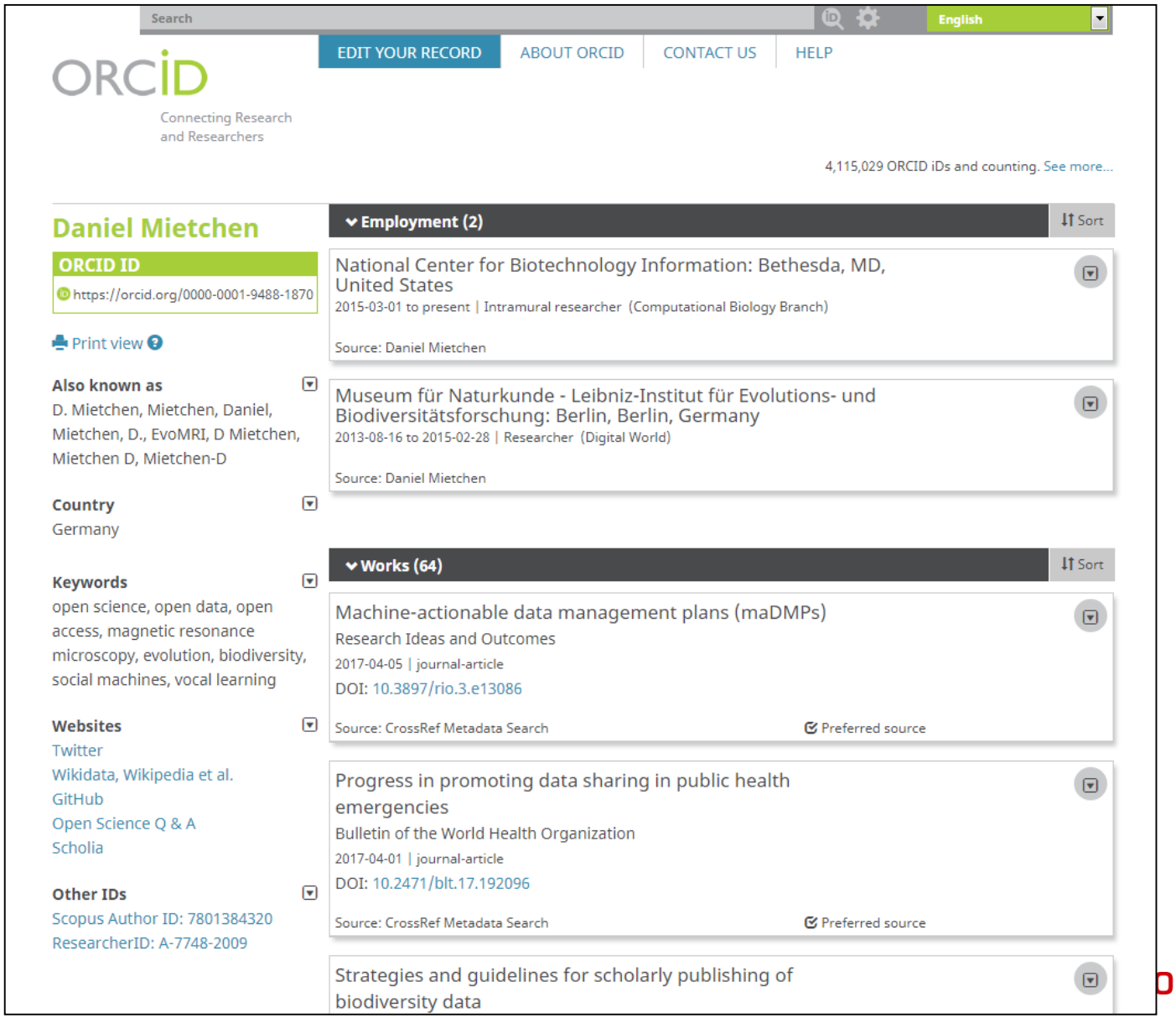
- ORCID ID
  - Unique person ID
  - ORCID assigned once
  - Person can change affiliations (jobs)

0000-0002-4929-7875

Persistent Identifiers - ORCID

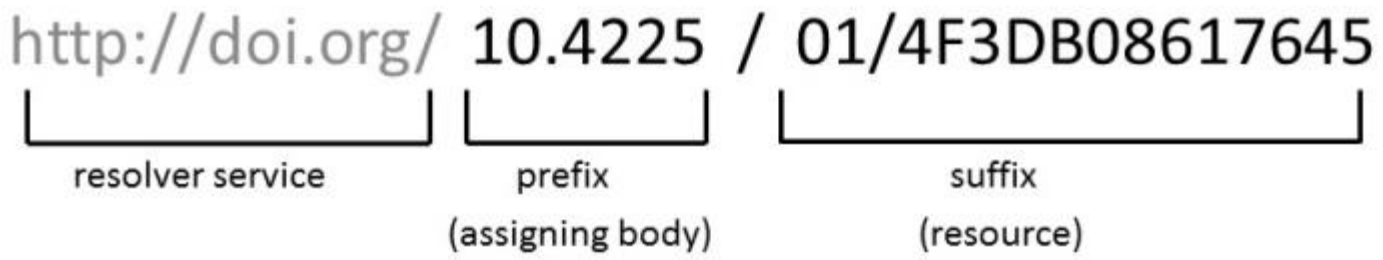


ORCID Example

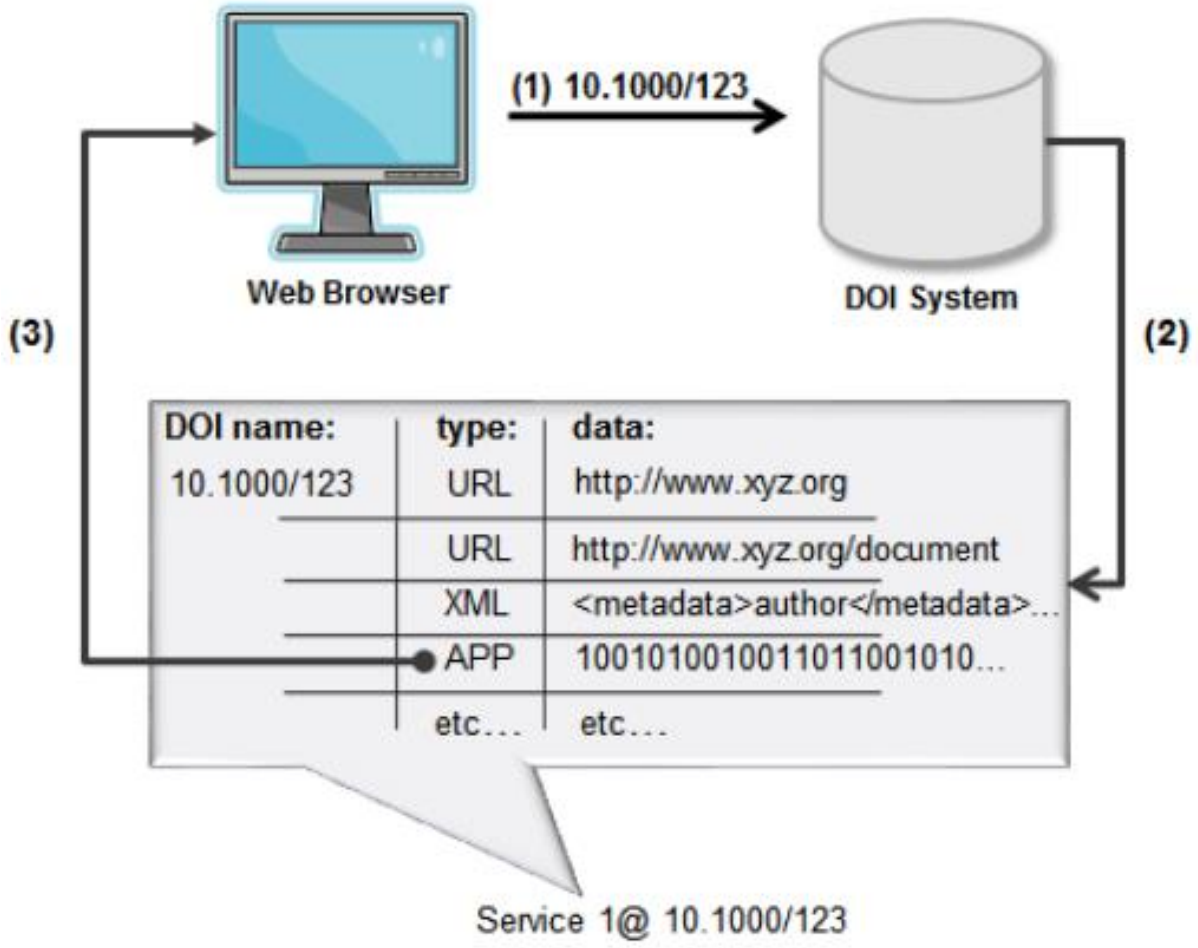


Persistent Identifiers - DOI

- Unlike the URLs, DOIs are associated to documents and not to locations
- DOIs are never deleted
  - if resource does not exist then a message is provided
- Resolver service
- Metadata



DOI – resolver service



https://www.doi.org/doi_handbook/3_Resolution.html

Tips for writing DMPs

Tips for writing DMPs

- DMP can reveal how solid your work is
- Seek advice - consult and collaborate
- When answering questions from checklists write coherent text
- Be specific when referring to tools and standards
- Assign responsibilities and name responsible personnel

Tips for writing DMPs

- Think about things early…
  - Negotiation on licenses and consent agreement may preclude later sharing if not careful
  - Manage your data correctly from the very beginning
    - backups, file naming conventions, access restrictions, metadata collection
  - Plan your budget

Decisions made early on affect what you can do later

DMPs are not that perfect

- Data Management Plans
  - are manually created
  - depend on scientific honesty
  - focus mainly on input and output data
  - provide very general overview of the experiment
  - have scarce information about the process
  - cannot be automatically validated
  - do not support sufficiently the reproducibility of research
- Lecture on 29.4: how to fix some of these problems

Summary

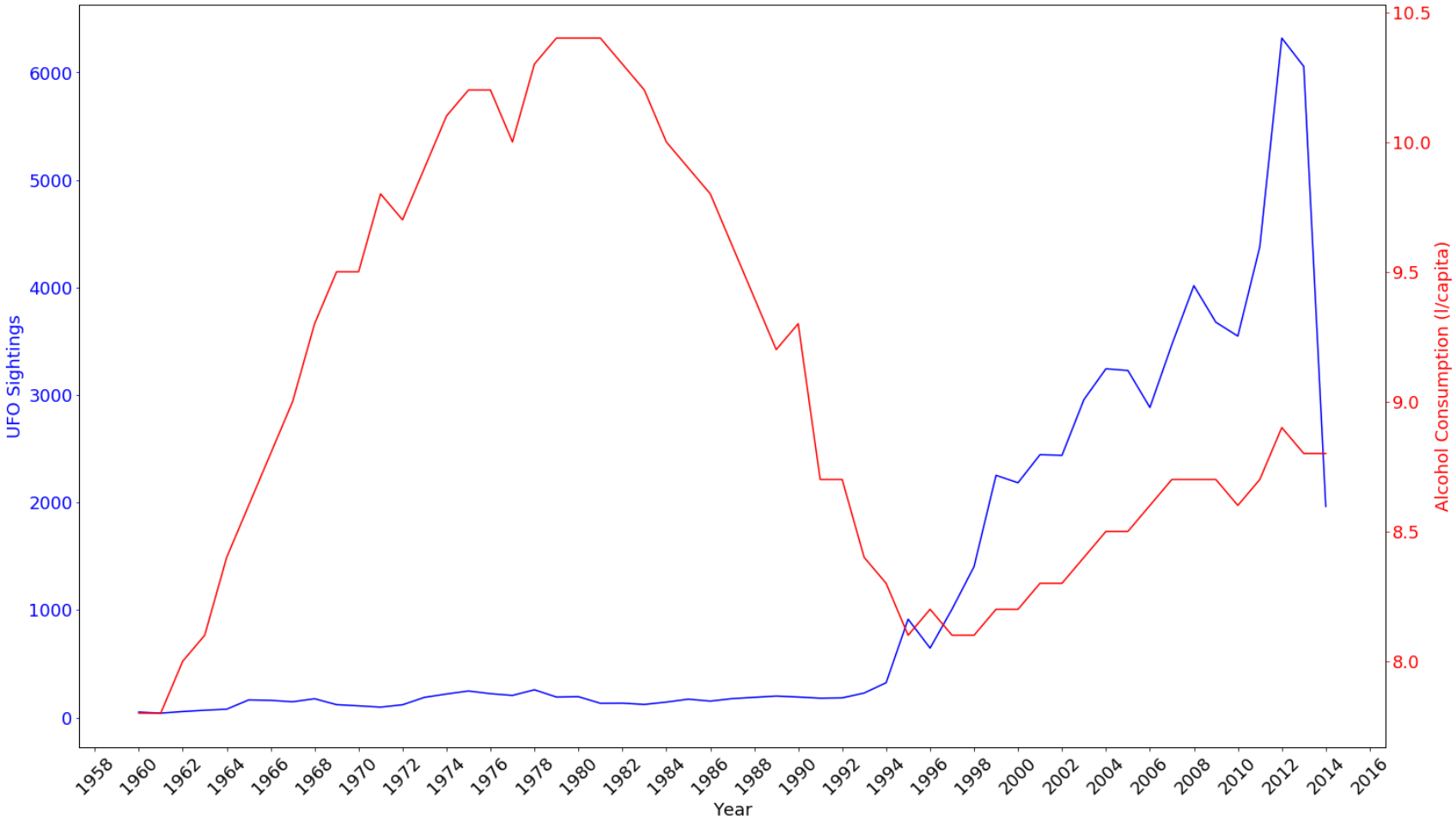
You should know

- how to improve your own data management
- what a DMP is and what kind of information it contains
- what the FAIR principles are
- how to create a DMP
- what to write in a DMP
- how to select a proper license
- what persistent identifiers are

Useful resources

- Managing and sharing data by UK Data Archive
  - http://www.data-archive.ac.uk/media/2894/managingsharing.pdf
- EUDAT webinars on data management
  - https://www.eudat.eu/events/webinar/research-data-management-an-introductory-webinar-from-openaire-and-eudat
- FFG-Akadmie: Der Data Management Plan (DMP) in Horizon 2020 (Webinar)
  - https://www.ffg.at/europa/veranstaltungen/ffg-akademie_2017-10-18
- DMP Online
  - https://dmponline.dcc.ac.uk
- Ten Simple Rules
  - http://dx.doi.org/10.1371/journal.pcbi.1004525
- DMP Checklist
  - http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf

Correlating Alcohol Consumption and UFO Sightings in the USA



https://github.com/mdietrichstein/digitalpreservation-dmp

Thank you! Any questions?

tmiksa@sba-research.org

Acknowledgements:

Thanks to EUDAT, DANS and DCC for reuse of slides, and to the OpenMinTeD and CAPSELLA projects for sharing their Data Management Plans