

# Data Management Plans in Horizon 2020

Dr. Tomasz Miksa

TU Wien & SBA Research

[tmiksa@sba-research.at](mailto:tmiksa@sba-research.at)

18/10/2017

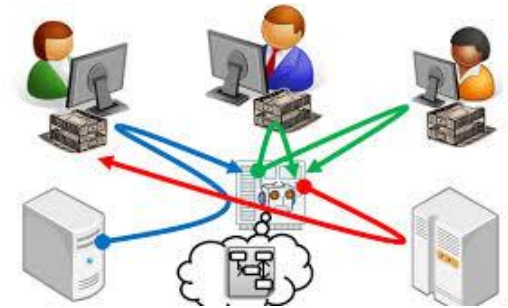
# Agenda

- Why do we need Data Management Plans (DMPs)?
- What is a DMP?
- What are the Horizon 2020 requirements?
- How to create a DMP?
- Tips for writing DMPs

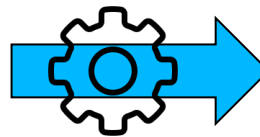
# **WHY DO WE NEED DATA MANAGEMENT PLANS?**

# e-Science and Research Infrastructures

- Scientists exchange
  - data
  - services
  - computational power
- Collaborate to solve challenges
  - DNA sequencing
    - climate change
    - tsunami forecasting
  - Earth Observation
    - climate change
    - tsunami forecasting
  - Large Hadron Collider

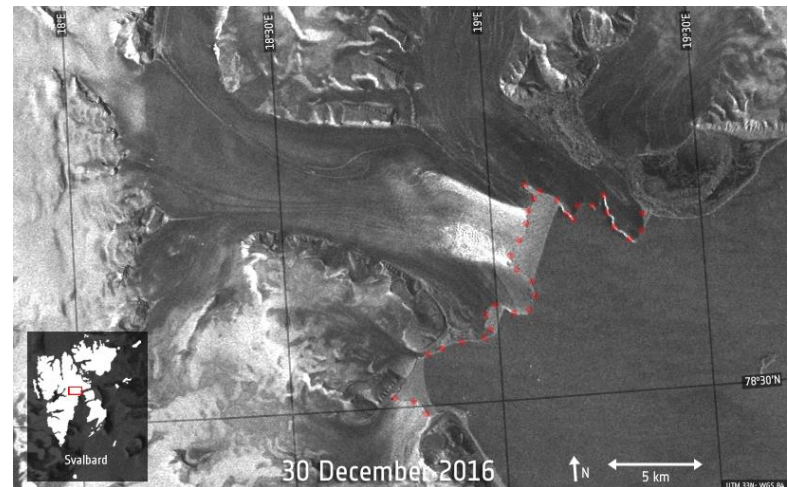


- Research requires **special tooling and software**
  - capture
  - pre-process
  - transform
  - visualize
  - interpret the data



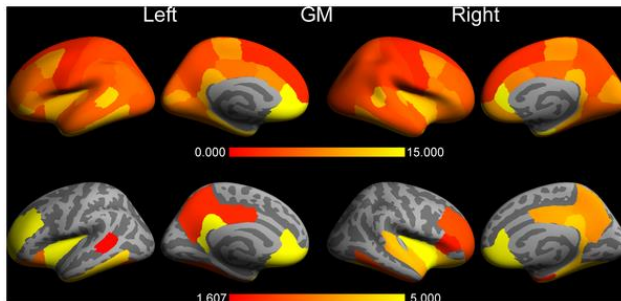
```

Driver: SAFE/Sentinel1/SBA SAFE Product
Files: S1A_IW_GRDM_1SDV_2015070804241_2015070804300_006672_008EAD_24EE.SAFE/manifest.safe
S1A_IW_GRDM_1SDV_2015070804241_2015070804300_006672_008EAD_24EE.SAFE/measurement/s1a-iw-grd-rr-2015070804241-2015070804300-006672-008EAD-001.tif
S1A_IW_GRDM_1SDV_2015070804241_2015070804300_006672_008EAD_24EE.SAFE/measurement/s1a-iw-grd-rr-2015070804241-2015070804300-006672-008EAD-001.tif
Size is 290, 187
Coordinate System is ''
GCP Projection =
PROJCS["WGS_84",
DATUM["WGS_1984",
SPHEROID["WGS 84",6378137,298.257223563,
AUTHORITY["EPSG","7830"]],
AUTHORITY["EPSG","4326"]],
PRIME["Greenwich",0,
AUTHORITY["EPSG","8901"]],
UNIT["degree",0.0174532925199433,
AUTHORITY["EPSG","9122"]],
AUTHORITY["EPSG","4326"]]]
GCP [0]: (x,y, z) = (-8.0350007028827,39.6332161725822,141.853266638322)
Metadata:
ACQUISITION_START_TIME=2015-07-08T06:42:14.1504840
ACQUISITION_STOP_TIME=2015-07-08T06:43:00.503530
BEAM_MODE=IS
BEAM_SWATH=IS
FACILITY_IDENTIFIER=UPA
LINE_SPACING=1.000055e+01
MISSION_ID=S1A
MODE=IW
ORBIT_DIRECTION=DESCENDING
ORBIT_NUMBER=4672
FIXED_SPACING=1.000000e+01
PRODUCT_TYPE=GRD
SATELLITE_IDENTIFIER=SENTINEL-1
SENSOR_IDENTIFIER=ISAR
Swath=IS
Subdatasets:
SUBDATASET_1_NAME=SENTINEL1_DS1A_IW_GRDM_1SDV_2015070804241_2015070804300_006672_008EAD_24EE.SAFE_IW_VH
SUBDATASET_1_DESC=Single band with SW swath and VH polarization
SUBDATASET_2_NAME=SENTINEL1_DS1A_IW_GRDM_1SDV_2015070804241_2015070804300_006672_008EAD_24EE.SAFE_IW_VV
SUBDATASET_2_DESC=Single band with SW swath and VV polarization
SUBDATASET_3_NAME=SENTINEL1_DS1A_IW_GRDM_1SDV_2015070804241_2015070804300_006672_008EAD_24EE.SAFE_IW
SUBDATASET_3_DESC=IS swath with all polarizations as bands
Corner Coordinates:
Upper Left ( 0.0, 0.0)
Lower Left ( 0.0, 167.0)
Upper Right ( 256.0, 0.0)
Lower Right ( 256.0, 167.0)
Center ( 128.0, 83.5)
Band 1 Block=16384x16 Type=Int16, ColorInterp=Undefined
Metadata:
POLARIZATION=VH
Swath=IS
Band 2 Block=16384x16 Type=Int16, ColorInterp=Undefined
Metadata:
POLARIZATION=VV
Swath=IS
  
```



# Reproducibility

- Studies show very low reproducibility in
  - medicine
  - economy
  - computer science
- Reproducibility requires
  - well documented research workflows
  - precise information on the experiment's environment [1] [2]



OPEN ACCESS Freely available online

## The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

Ed H. B. M. Gronenschild<sup>1,2,3</sup>, Petra Habets<sup>1,2</sup>, Heidi I. L. Jacobs<sup>1,2,3</sup>, Ron Mengelers<sup>1,2</sup>, Nico Rozeendaal<sup>1,2</sup>, Jim van Os<sup>1,2,4</sup>, Machteld Marcellis<sup>1,2</sup>

1 Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University Medical Center, Maastricht, Altheimer Center Limburg, The Netherlands, 2 European Graduate School of Neuroscience (EURON), Maastricht University, Maastricht, The Netherlands, 3 Cognitive Neurology Section, Institute of Neuroscience and Medicine-3, Research Centre Jülich, Jülich, Germany, 4 King's College London, King's Health Partners, Department of Psychosis Studies Institute of Psychiatry, London, United Kingdom

**Abstract**

FreeSurfer is a popular software package to measure cortical thickness and volume of neuroanatomical structures. However, little if any is known about measurement reliability across various data processing conditions. Using a set of 30 anatomical T1-weighted 3T MRI scans, we investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). Significant differences were revealed between FreeSurfer version v5.0.0 and the two earlier versions. These differences were on average 8.8±6.6% (range 1.3–64.0%) (volume) and 2.8±1.3% (1–7.7%) (cortical thickness). About a factor two smaller differences were detected between Macintosh and Hewlett-Packard workstations and between OSX 10.5 and OSX 10.6. The observed differences are similar in magnitude as effect sizes reported in accuracy evaluations and neurodegenerative studies. The main conclusion is that in the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system or to switch to a different type of workstation without repeating the analysis; results thus give a quantitative support to successive recommendations stated by FreeSurfer developers over the years. Moreover, in view of the large and significant cross-version differences, it is concluded that formal assessment of the accuracy of FreeSurfer is desirable.

Citation: Gronenschild EHM, Habets P, Jacobs HL, Mengelers R, Rozeendaal N, et al. (2012) The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. PLoS ONE 7(6): e38234. doi:10.1371/journal.pone.0038234

Editor: Satoshi Hayasaka, Wake Forest School of Medicine, United States of America

Received January 12, 2012; Accepted May 1, 2012; Published June 1, 2012

Copyright: © 2012 Gronenschild et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Geeskracht program of the Dutch Health Research Council (ZON-MW, grant number 10-000-1002), and the European Community's Seventh Framework Program under grant agreement No. HEALTH-F2-2009-241909 (Project EU-GES). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

\* E-mail: ed.gronenschild@maastrichtuniversity.nl

**Introduction**

FreeSurfer (Adninsula A. Martinos Center for Biomedical Imaging, Harvard/MIT, Boston) comprises a popular and freely available set of tools for deriving neuroanatomical volume and cortical thickness measurements from automated lesion segmentation (<http://surfer.nmr.mgh.harvard.edu>), recently summarised by Fuchs [1]. A number of reported studies discussed the accuracy of the technique by comparing the volume of specific brain structures, such as the hippocampus or amygdala, with manually derived volumes [2–5]. The measurement of cortical thickness was validated against histological analysis [6] and manual measurements [7,8]. Also the reliability of the measurements was subject of a number of investigations. Some of these studies addressed the effect of scanner-specific parameters, including field strength, pulse sequence, scanner upgrade, and vendor (cortical thickness) [9,10].

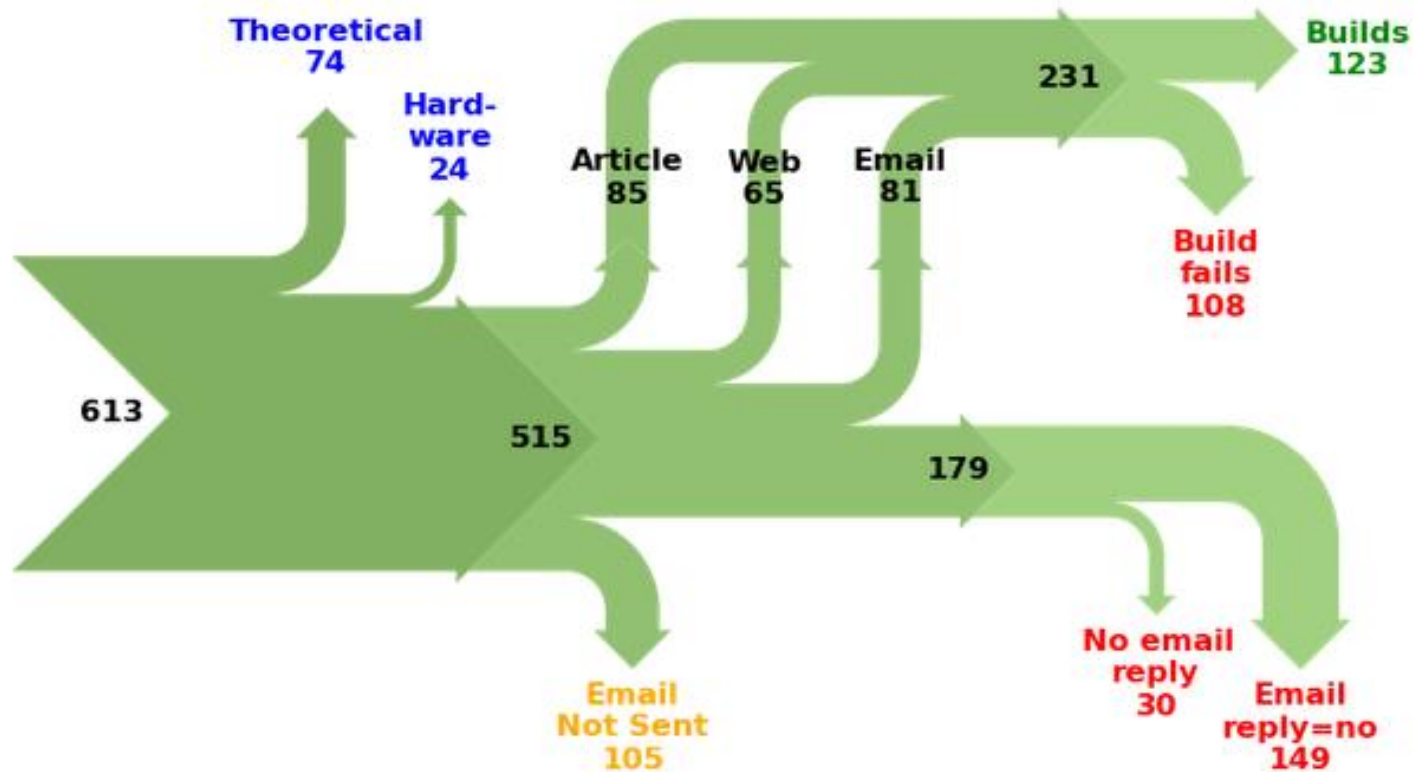
Since FreeSurfer is CPU-intensive (20–30 hours per brain for a full segmentation is not exceptional), it is common practice to distribute the computational load among the available central processor units (CPU) on a single workstation and/or among several workstations. Given this context, a number of questions suggest themselves: (1) does every CPU produce the same results; (2) is there any interaction between the processes running simultaneously on the same workstation; (3) does every workstation produce the same results?

Just like similar neuroimaging packages, new releases of FreeSurfer are issued regularly, fixing known bugs and improving existing tools and/or adding new ones. Each release is accompanied with documentation describing the changes relative to the previous release (<http://surfer.nmr.mgh.harvard.edu/swiki/ReleaseNotes>). However, transition to a new release during the course of a study may affect the results and is therefore

[1] <https://doi.org/10.1016/j.jbi.2016.10.011>  
 [2] <https://doi.org/10.1371/journal.pone.0038234>

# 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 replies 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

- To improve reproducibility and data management many solutions were proposed
  - **open access** to scientific publications and data
  - research **data repositories** to host the data
  - **data citation** to reference the datasets
  - **DATA MANAGEMENT PLANS**

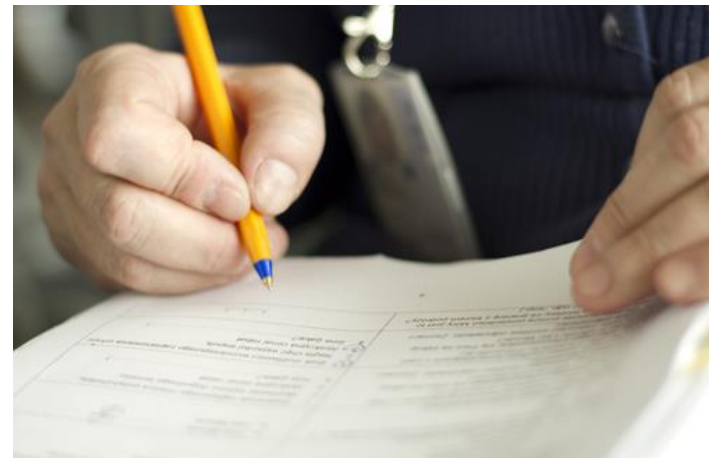


# WHAT IS A DATA MANAGEMENT PLAN?

# 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]

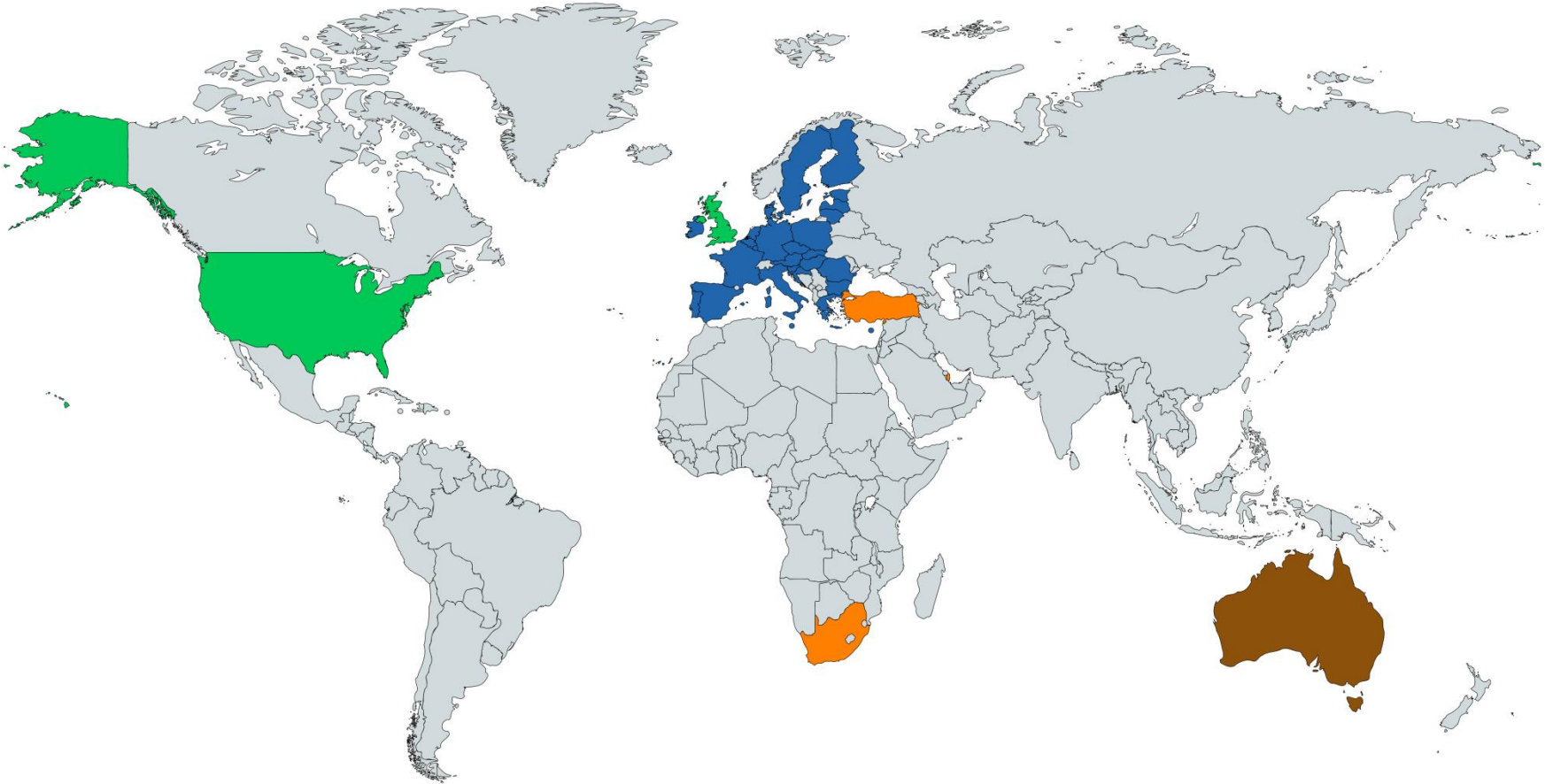


# DMP is an awareness tool!

- 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 be useful for ethics committee
- DMP can reveal how solid your research methodology is
  - is it a 'fishing expedition'?



# DMPs worldwide



Created with mapchart.net ©



# **HORIZON 2020 REQUIREMENTS**

# EC Horizon 2020

## Open Research Data Pilot

- Open access to publications is default
- Open access to research data is default from 2017
  - **NOT** all data must be released
  - data needed to validate scientific publications
  - other data on a voluntary basis
- Opt-out possibility
  - No impact on the proposal evaluation
  - At every stage
- Data management costs can be claimed



# EC Horizon 2020 DMP Template

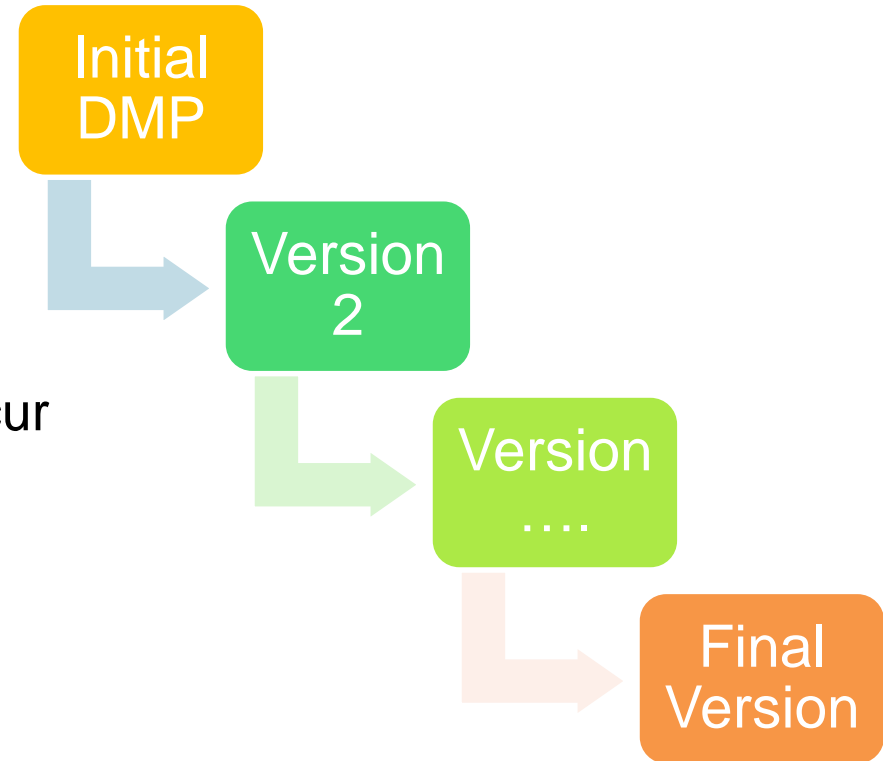
- Template is recommended but not required
  - 6 sections
  - 31 questions
  - Follows FAIR principles
    - Data must be **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable

DMP component	Issues to be addressed
1. Data summary	<ul style="list-style-type: none"> <li>• State the purpose of the data collection/generation</li> <li>• Explain the relation to the objectives of the project</li> <li>• Specify the types and formats of data generated/collected</li> <li>• Specify if existing data is being re-used (if any)</li> <li>• Specify the origin of the data</li> <li>• State the expected size of the data (if known)</li> <li>• Outline the data utility: to whom will it be useful</li> </ul>
2. FAIR Data	

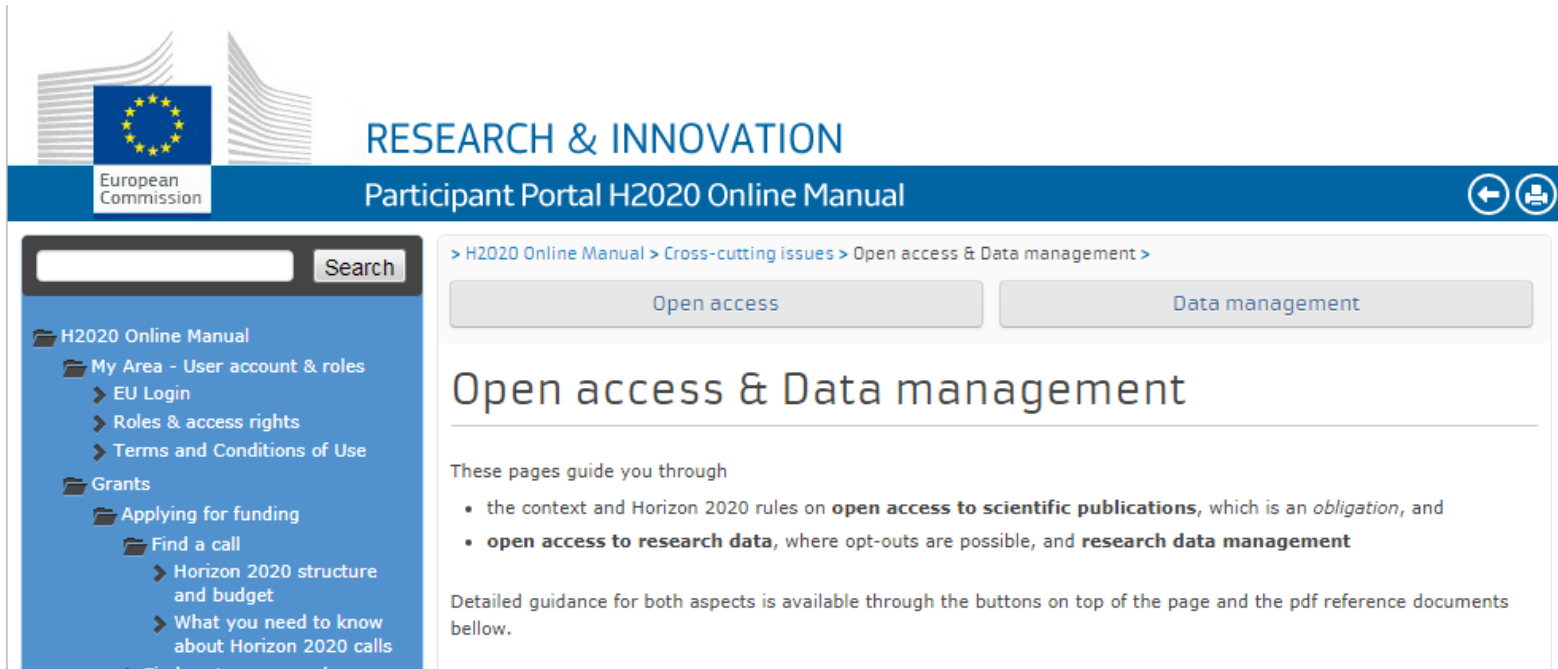


# EC Horizon 2020 DMP versions

- 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



# More Horizon 2020 specific information



The screenshot shows the 'Participant Portal H2020 Online Manual' page. The header includes the European Commission logo and the text 'RESEARCH & INNOVATION'. The main title is 'Open access & Data management'. Below the title, there are two buttons: 'Open access' and 'Data management'. The page content states: 'These pages guide you through' followed by a bulleted list:
 

- the context and Horizon 2020 rules on **open access to scientific publications**, which is an *obligation*, and
- **open access to research data**, where opt-outs are possible, and **research data management**

 Detailed guidance for both aspects is available through the buttons on top of the page and the pdf reference documents below.

- [http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination\\_en.htm](http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm)
- [Guidelines to rules on Open Access to Scientific Publications & Open Access to Research Data in Horizon 2020](#)
- [Guidelines on Data Management in Horizon 2020](#)
- [Template for the Data Management Plan](#)

# HOW TO CREATE A DMP?

# DMP in Horizon 2020

- DMP is a project deliverable
- DMP is a written document
- DMP Template
  - recommended but not required
  - contains auxiliary questions
  - [Template for the Data Management Plan](#)

## 1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

What types and formats of data will the project generate/collect?

Will you re-use any existing data and how?

What is the origin of the data?

What is the expected size of the data?

To whom might it be useful ('data utility')?

## 2. FAIR data

### 2.1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

What naming conventions do you follow?

Will search keywords be provided that optimize possibilities for re-use?

Do you provide clear version numbers?

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

### 2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

How will the data be made accessible (e.g. by deposition in a repository)?

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)?

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Have you explored appropriate arrangements with the identified repository?

If there are restrictions on use, how will access be provided?

Is there a need for a data access committee?

Are there well described conditions for access (i.e. a machine readable license)?

How will the identity of the person accessing the data be ascertained?

### 2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

# DMP in Horizon 2020

- Software tools for filling out DMPs
  - users choose appropriate funders template
  - relevant questions and guidance is presented
  - facilitate co-working
  - results can be exported to PDF
- Usually mixed approach works best
  - check guidance in the tool and create your own document



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

If applying for funding, state the title exactly as in the proposal.

### Primary research organisation

Select the primary research organisation responsible

My research organisation is not on the list or no research organisation is associated with this plan

### Funding organisation

Select the funding organisation

No funder associated with this plan

Create Plan <https://dmponline.dcc.ac.uk>



- 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





- 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

**B** *I*

Save

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.



## FFG Webinar Horizon 2020 Example

Plan details Initial DMP Detailed DMP Final review DMP Share **Export**

From here you can download your plan in various formats. This may be useful if you need to submit your plan as part of a grant application. Select what format you wish to use and click to 'Export'.

### Initial DMP

Format

pdf

Export Settings (Using default PDF formatting values)

### File Name

File Name

### Included Elements

#### Details

Plan Name   
 Plan ID   
 Grant Number   
 Principal Investigator / Researcher   
 Plan Data Contact   
 Description   
 Funder   
 Institution

#### Sections

**1. Data summary**  
 Provide a summary of the data addressing the fol  
**2. FAIR data**  
 2.1 Making data findable, including provisions for  
 2.2 Making data openly accessible: Specify which  
 2.3 Making data interoperable: Assess the interop

### 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.

# WHAT SHOULD I WRITE?

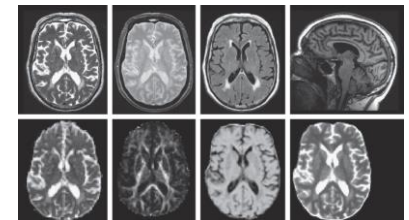
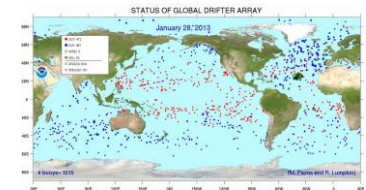
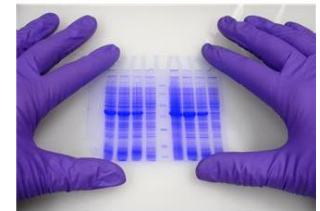
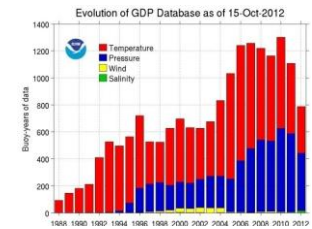
# DMP Horizon 2020

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



# 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

Every two days, we will subsample *E. affinis* populations growing under our treatment conditions. We will use a microscope to identify the life stage and sex of the subsampled individuals. We will **document the information first in a laboratory notebook and then copy the data into an Excel spreadsheet**. The Excel spreadsheet will be saved as a comma separated value **(.csv) file**.

From DataOne – *E. affinis* DMP example

# FAIR Principles

- **Findable**
  - the data is available online, likely in a repository
  - contains metadata that facilitates search
- **Accessible**
  - access conditions are specified
  - software needed to interpret data is known
- **Interoperable**
  - Follow standards and domain specific conventions
- **Reusable**
  - clear license and documentation
  - 'sum of the three other rules'
- There is no clear distinction between principles
  - e.g. metadata supports all of them

# Standards and Metadata





# Metadata – Atlas Of Living Australia

## 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 0)
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	<a href="http://natureshare.org.au/observation/2380/">http://natureshare.org.au/observation/2380/</a>
Photographer	Russell Best
Rights holder	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	<i>Cracticus tibicen</i> Supplied scientific name 'Gymnorhina tibicen'
Taxon rank	Species
Common name	Australian Magpie
Kingdom	Animalia
Phylum	Chordata
Class	Aves
Order	Passeriformes
Family	Artamidae
Genus	<i>Cracticus</i>
Species	<i>Cracticus tibicen</i>

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

# Metadata – Atlas Of Living Australia

## Dataset

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

# 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>
  - <http://rd-alliance.github.io/metadata-directory/standards/>

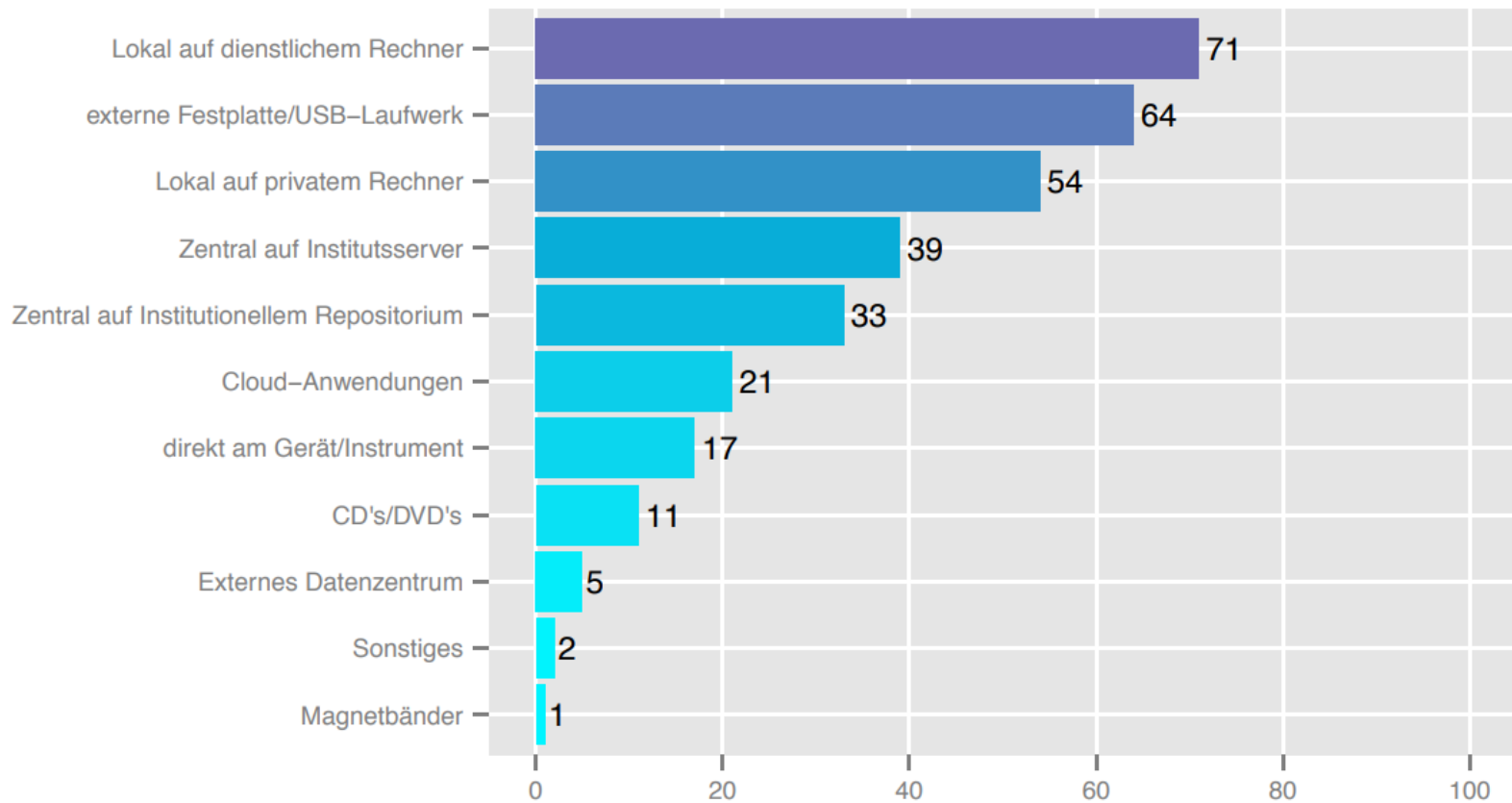
## Standards and metadata

We will first document our metadata by taking careful notes in the laboratory notebook that refer to specific data files and **describe all columns, units, abbreviations, and missing value identifiers**. These notes will be transcribed into a **.txt document that will be stored with the data file**. After all of the data are collected, we will then use EML (Ecological Metadata Language) to digitize our metadata. **EML is one of the accepted formats used in ecology**, and works well for the types of data we will be producing. We will create these metadata using Morpho software, available through KNB. The metadata will fully describe the data files and the context of the measurements.

From DataOne – E. affinis DMP example

# Managing data during research

## Wo speichern Sie normalerweise Ihre Forschungsdaten ab?



Anzahl der Antworten, Skalierung in %

e-infrastructures  
austria



<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

With collaborators while research is active



Data are mutable

(Open) data sharing



Data Repository

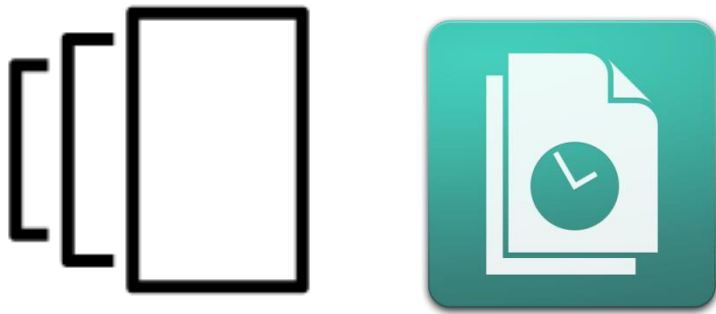


zenodo

Data are stable, searchable, citable, clearly licensed

# Backup vs preservation

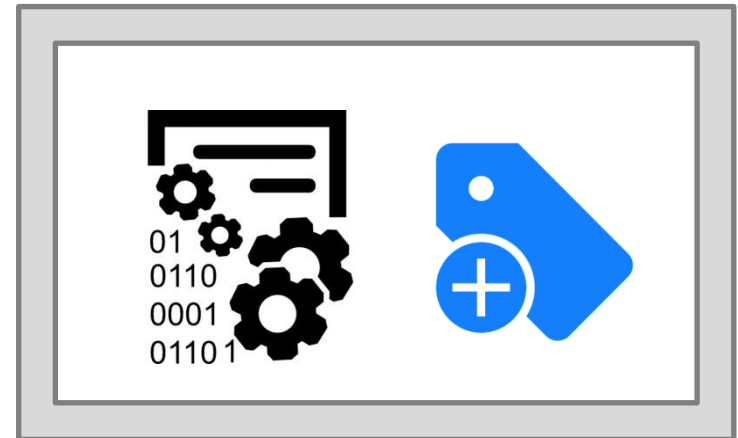
Storing and backing up files while research is active



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 has scientific, cultural or historical value?
  - 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?

# Where to find a repository?

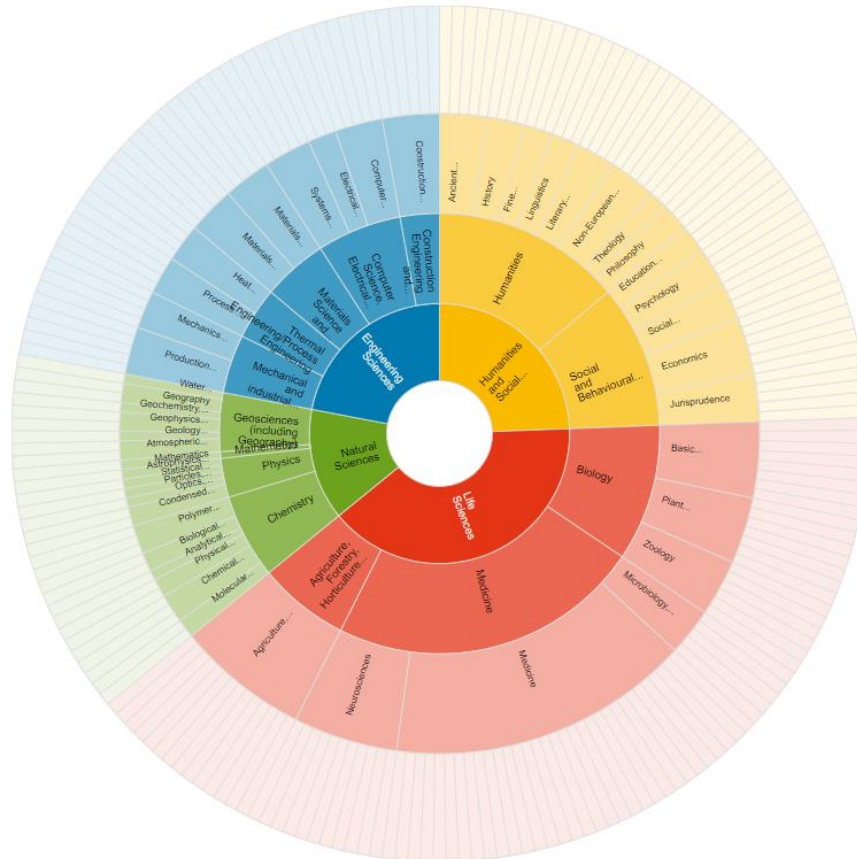


- More information: <https://www.openaire.eu/opendatapilot-repository>
- Zenodo: <http://www.zenodo.org>
- Re3data.org: <http://www.re3data.org>

## Browse by subject

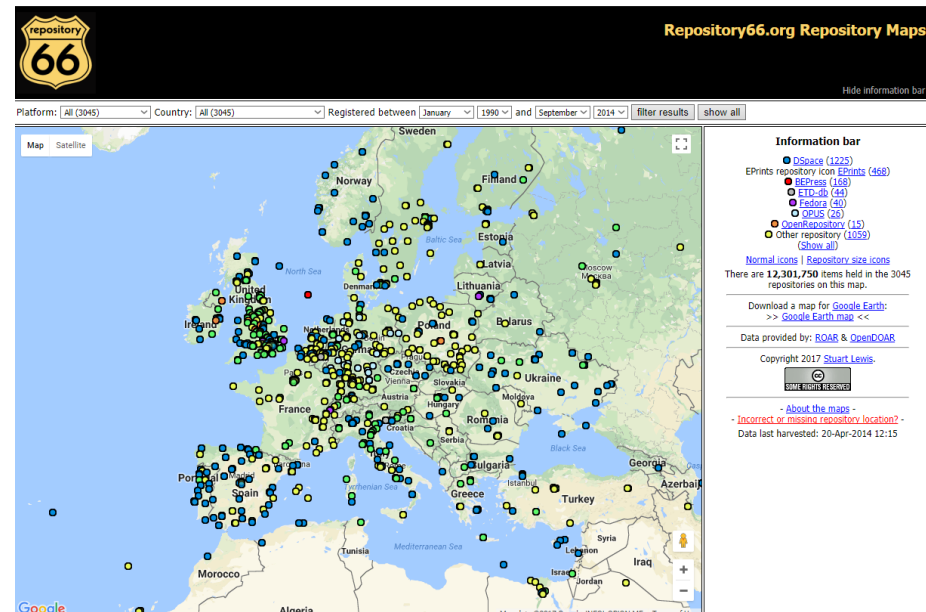
Graphical  Text

click to zoom into subjects or to select a bottommost subject in the hierarchy as filter for the re3data search page  
ctrl + click on a top subject to select it as filter



# 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>



# Repositories in Austria - examples

The screenshot shows the Phaidra website interface. At the top, it features the University of Vienna logo and navigation links for 'Phaidra', 'Search', and 'Login'. Below the header, there is a search bar and a 'Contact' section with the email 'phaidra@univie.ac.at'. The main content area is titled 'Featured collections' and displays six collection cards with images and brief descriptions:

- E-Books on Demand:** In dieser Collection finden Sie die im Rahmen des Services eBooks on Demand (EOD) digitalisierten Bücher der Universitätsbibliothek Wien.
- uschoolar:** In der uschoolar-Collection von Phaidra finden Sie weltweit frei zugängliche wissenschaftliche Publikationen von Forschenden der Universität Wien.
- Digitales Forschungsarchiv Byzanz:** Das Ziel des Digitalen Forschungsarchivs Byzanz ist es, das byzantinische Reich fotografisch möglichst umfangreich zu erfassen und zugänglich zu machen.
- UB-Maps:** Der Kartenbestand der Fachbibliothek Geographie und Regionalforschung (Universität Wien) geht historisch bis in die Gründungszeit der Lehrkanzel für Geographie an der Universität Wien im Jahr 1851 zurück.
- Digitalisierte Bestände der Österreichischen Zentralbibliothek für Physik:** Diese Collection beinhaltet die an der Österreichischen Zentralbibliothek für Physik vorhandenen multimediale Bestände (Videos, Tonaufzeichnungen etc.) sowie digitalisierte Nachlässe und Sonderausstellungen.
- 650 Jahre Universität Wien:** Die Universität Wien feierte 2015 ihr 650. Gründungsjubiläum mit zahlreichen Veranstaltungen.

<https://phaidra.univie.ac.at>

The screenshot shows the CCCA Data Server website interface. At the top, it features the CCCA Climate Change Centre AUSTRIA logo and navigation links for 'Log In', 'Register', and 'Contact'. Below the header, there is a search bar and a 'Search data' section with the input 'E.g. environment'. The main content area is titled 'Welcome to the CCCA Data Server' and includes a description of the server's purpose and a 'Search data' section with a search bar and popular keywords: 'CCCA Dataset', 'glacier', and 'scaled distribution'. Below this, there is a 'Explore our possibilities:' section with a map of Austria showing 'Daily Mean Near-Surface Air Temperature' data. The map is color-coded from blue (colder) to red (warmer). Below the map, there is a legend for 'Daily Mean Near-Surface Air Temperature (Celsius)' ranging from -4.5 to 16.5. To the right of the map, there is a 'CCCA Data Server statistics' section with the following data:

resources in	datapackages	organizations	groups
1898	146	33	6

At the bottom of the page, there is a 'ZAMG' section with the text 'Zentralanstalt für Meteorologie und Geodynamik' and a 'News: 05.10.2017' section with the text 'Austrian PANGAEA data harvested - over 913 data files organised in 75 data packages were harvested. The pangaea metadata profiles (ISO, OAI) were used and implemented for harvesting of Austrian data resources, especially glacier and permafrost informations'.

<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](http://www.dcc.ac.uk/resources/how-guides/license-research-data)

- EUDAT licensing wizard help you pick licence for data & software

Do you own copyright and similar rights in your dataset and all its constitutive parts?

Do you allow others to make commercial use of you data?

**Creative Commons Attribution (CC-BY)**  
This is the standard creative commons license that gives others maximum freedom to do what they want with your work.

**Public Domain Dedication (CC Zero)**  
CC Zero enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

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

# Data sharing and preservation

Data will be provided in **file formats considered appropriate for long-term access**, as recommended by the UK Data Service. For example, SPSS Portal format and tab-delimited text for qualitative tabular data and **RTF** and **PDF/A** for interview transcripts. Appropriate **documentation necessary** to understand the data will also be provided. Anonymised data will be held for **a minimum of 10 years** following project completion, in compliance with LSHTM's Records Retention and Disposal Schedule. Biological samples (output 3) will be **deposited with the UK BioBank** for future use.

From [Writing a Wellcome Trust Data Management and Sharing Plan](#)

# TIPS FOR WRITING DMPS



# Tips for writing DMPs

- DMPs vary across scientific domains
- No good or bad answers – rationale is important
- DMP can reveal how solid your research is
- Seek advice - consult and collaborate
- Discuss any technical issues with the IT personnel
- 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
  - Useful to consider data issues at the consortium negotiation stage to make sure potential issues are identified and sorted asap
  - 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**

# Summary

- DMPs are NOT meant to be yet another paper work imposed on researchers!
- DMPs are an awareness tool
- DMPs help you plan your project
- DMPs help in making data FAIR
- Future work: machine-actionable DMPs
  - automate data management
  - RDA DMP Common Standards WG
    - [https://www.rd-alliance.org/system/files/documents/RDA\\_P10\\_DMPCommonStandardsWG.pdf](https://www.rd-alliance.org/system/files/documents/RDA_P10_DMPCommonStandardsWG.pdf)
    - <https://www.rd-alliance.org/groups/dmp-common-standards-wg>





ABOUT   SPEAKERS   AGENDA   REGISTER   PARTNERS   VENUE

NOVEMBER 23, 2017, VIENNA

# RDA Europe Workshop Data Stewardship Realized: From Planning to Action

Towards the Establishment of an Austrian Research Infrastructure

# Useful resources

- Most links can be found on specific slides
- Managing and sharing data by UK Data Archive
  - <http://www.data-archive.ac.uk/media/2894/managingsharing.pdf>
- 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](http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf)
- EUDAT webinars on data management
  - <https://eudat.eu/training/research-data-management>
- DMP Schulungsunterlagen e-Infrastructures Austria
  - <http://phaidra.univie.ac.at/o:459770>

**Thank you! Any questions?**

[tmiksa@sba-research.org](mailto: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