

# Data Management Plans

Dr. Tomasz Miksa

SBA Research & TU Wien

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

# Agenda

- What is a Data Management Plan (DMP)?
- Where does the idea come from?
- What is inside of a DMP?
- How to create a DMP?

# What is a Data Management Plan?

# What is a Data Management Plan (DMP)?

“Plans typically state **what data will be created and how**,  
and outline the **plans for sharing and preservation**,  
noting what is appropriate given **the nature of the data**  
and any **restrictions** that may need to be applied. “



[Digital Curation Centre, GB]

# What is a Data Management Plan (DMP)?

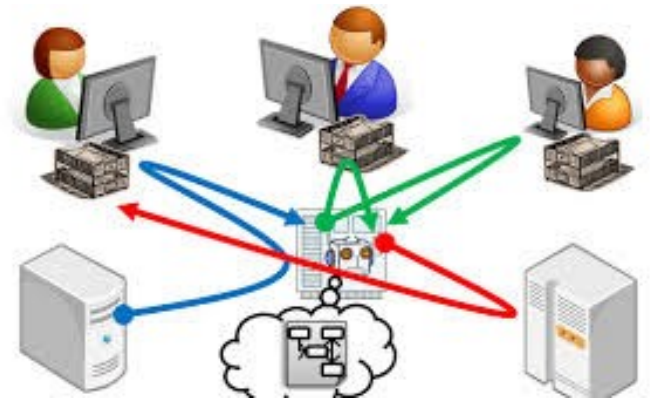
*“... Data Management Plans (DMPs) detailing **what data** the project will generate, whether and how it will be **exploited** or **made accessible** for verification and re-use, and how it will be **curated** and **preserved**.”*



*[European Commission, Guidelines on Data Management in Horizon 2020 ]*

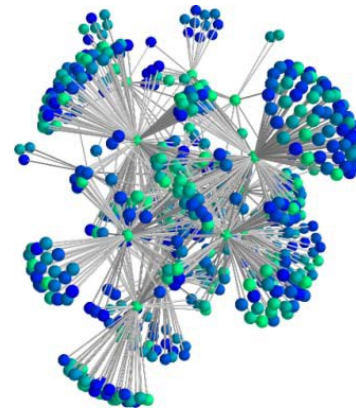
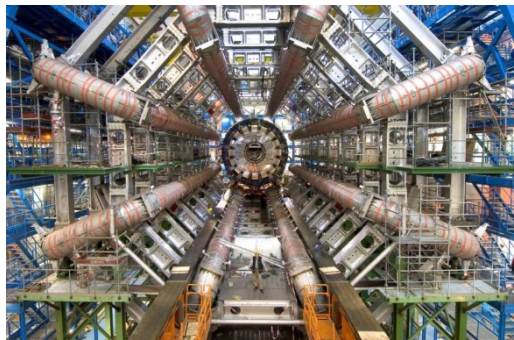
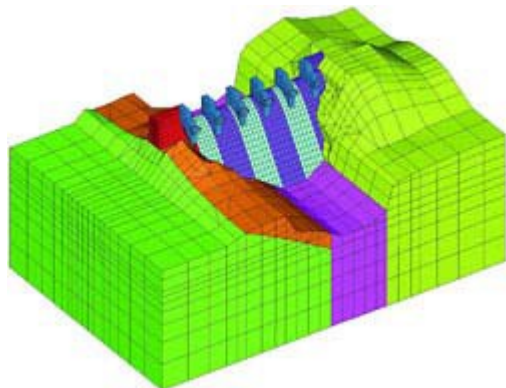
**Where does the idea come from?**

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



# eScience examples

- DNA sequencing
- Earth observation
  - climate change
  - tsunami forecasting
- Large Hadron Collider at CERN
  - 300GB per second of raw data from detectors



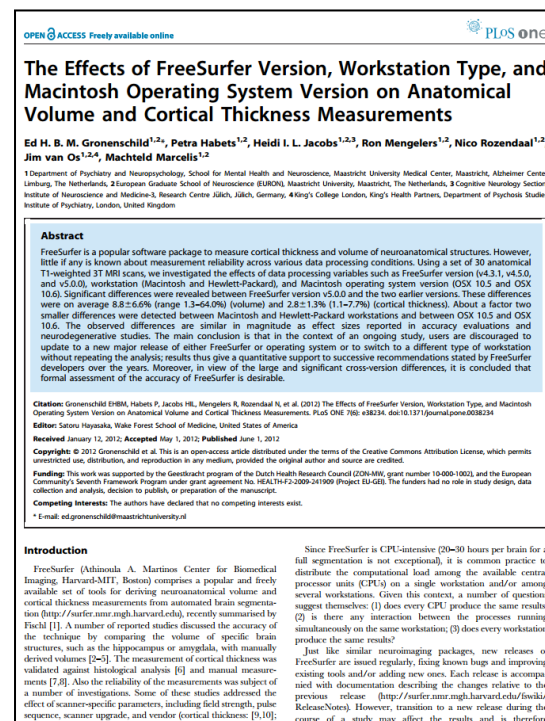
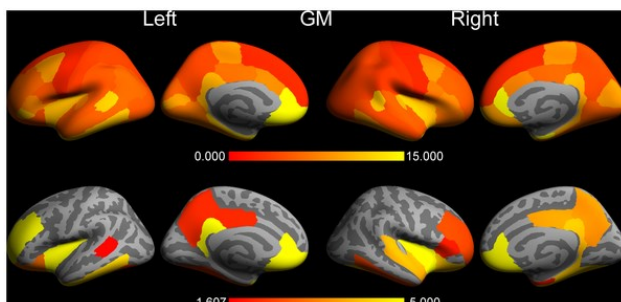


## ■ Current studies show very low reproducibility in

- medicine
- economy
- computer science

## ■ Reproducibility requires

- well documented research workflows
- precise information on the experiment's environment

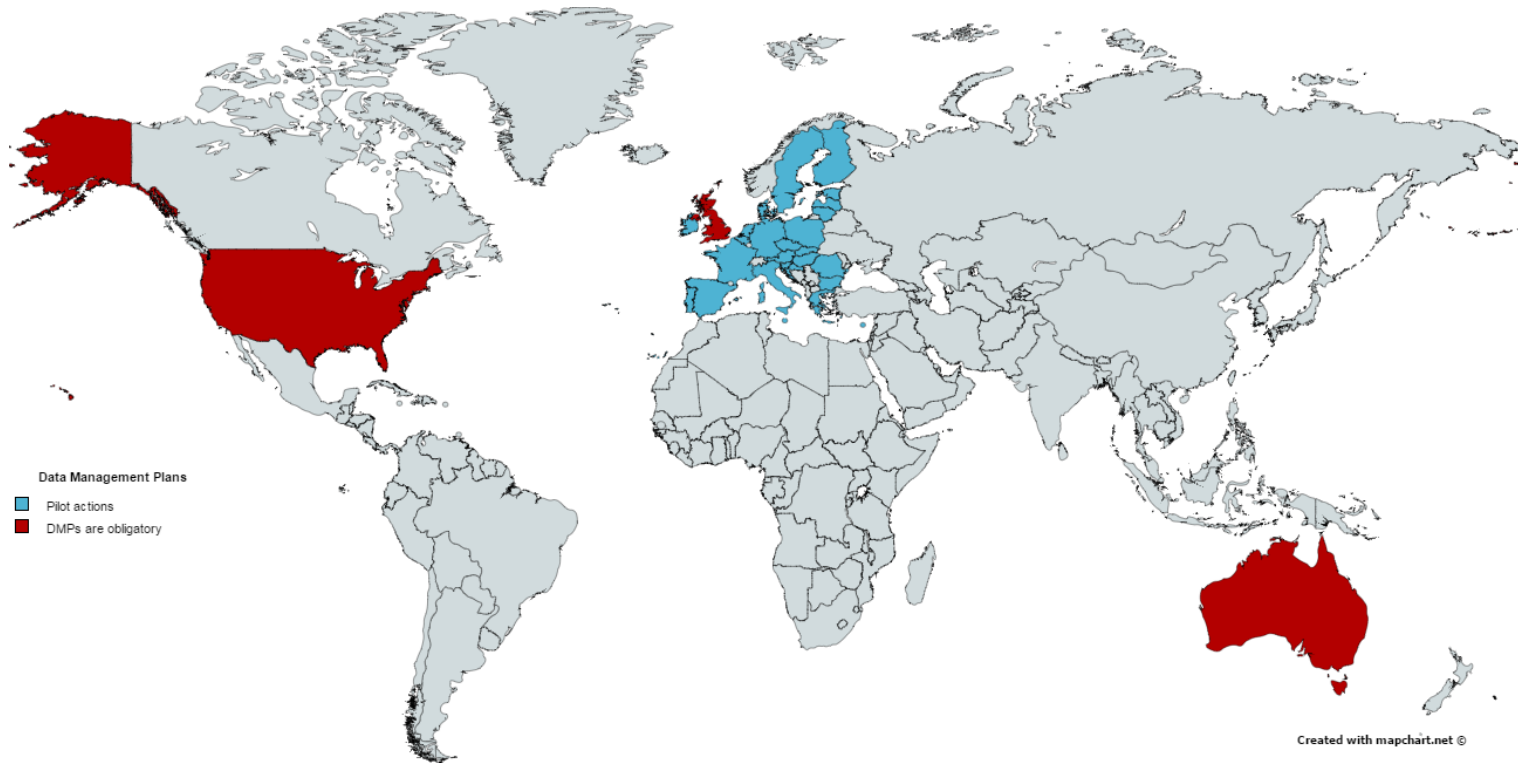


# Variety of solutions

- 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**



# DMPs worldwide



- Obligatory (red)
  - National Science Foundation in US
  - Australian National Data Service
  - Research Councils in GB
- Pilot actions (blue)
  - EU Horizon 2020

# DMPs in Austria

- FFG
  - policy is planned
- FWF
  - pilot action
  - recommended (not required)
- WWTF
  - pilot action
  - required (no impact on proposal evaluation)
- EU Horizon 2020



## What is inside of a DMP?

# What is inside of a DMP?

- It depends...
  - on the institution requiring a DMP
  - field of research
- DMP is
  - usually a written document
  - usually has an enforced structure
- Most templates overlap
- Level of details varies
- DMP creation facilitated by
  - questionnaires, guidance documents, checklists, etc.

# EC Horizon 2020 – Open Data Pilot

“All project proposals (...) include a section on research data management which is evaluated under the criterion 'Impact'. Where relevant, **applicants must provide a short, general outline of their policy for data management (...)**”



[http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-data-mgt\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf)

# EC Horizon 2020 DMP Template

## Annex 1: Data Management Plan (DMP) template

The purpose of the Data Management Plan (DMP) is to provide an analysis of the main elements of the data management policy that will be used by the applicants with regard to all the datasets that will be generated by the project.

The DMP is not a fixed document, but evolves during the lifespan of the project.

The DMP should address the points below on a dataset by dataset basis and should reflect the current status of reflection within the consortium about the data that will be produced.

- **Data set reference and name**

Identifier for the data set to be produced.

- **Data set description**

Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.

- **Standards and metadata**

Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.

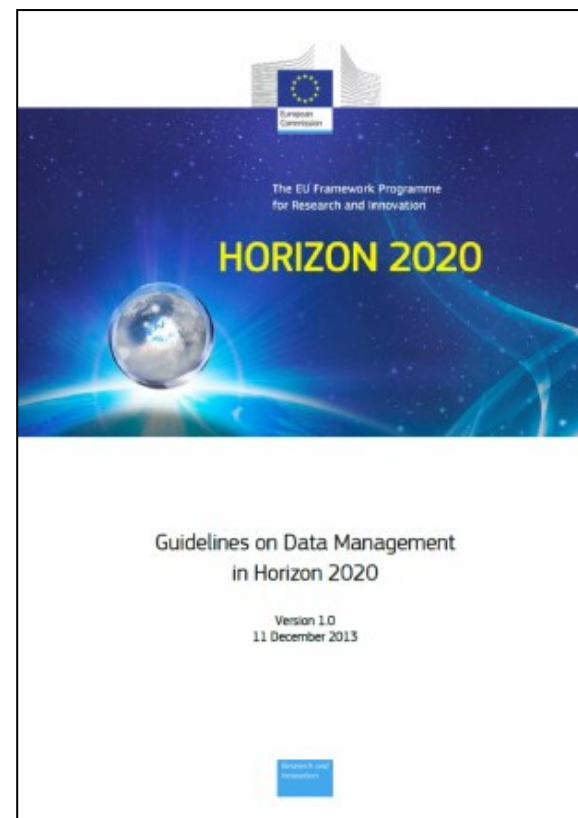
- **Data sharing**

Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).

In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

- **Archiving and preservation (including storage and backup)**

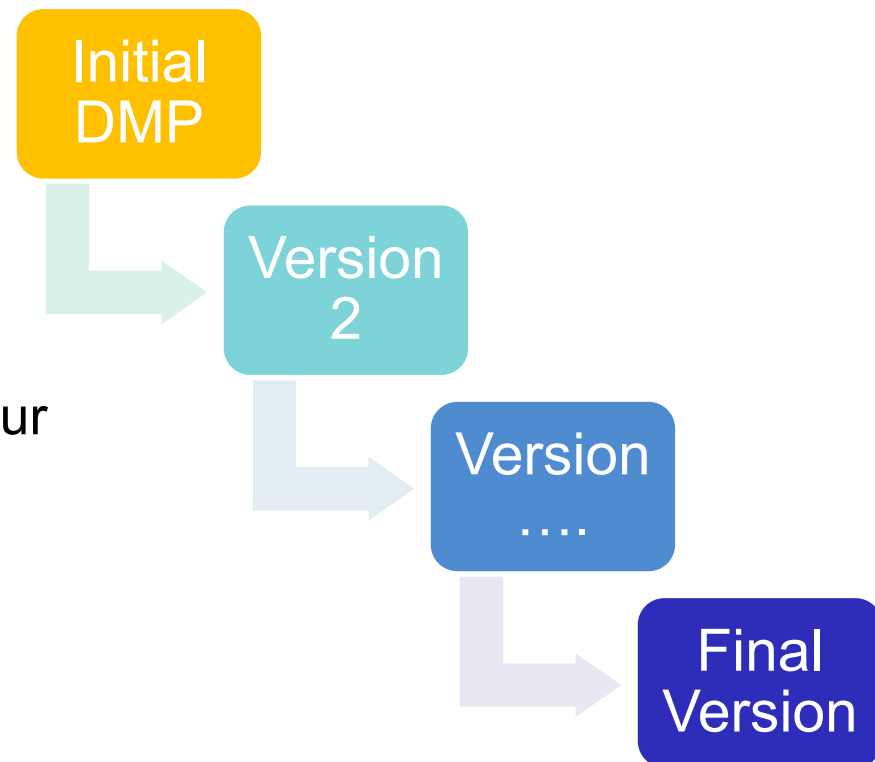
Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.



[http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-data-mgt\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf)



- 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
  - results can be exported or directly submitted



# DCC Checklist

- 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	
 <b>Checklist for a Data Management Plan, v4.0</b>	
Please cite as: DCC. (2013). <i>Checklist for a Data Management Plan. v.4.0</i> . Edinburgh: Digital Curation Centre. Available online: <a href="http://www.dcc.ac.uk/resources/data-management-plans">http://www.dcc.ac.uk/resources/data-management-plans</a>	
DCC Checklist	DCC Guidance and questions to consider
<b>Administrative Data</b>	
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	<b>Questions to consider:</b> <ul style="list-style-type: none"> <li>- What is the nature of your research project?</li> <li>- What research questions are you addressing?</li> <li>- For what purpose are the data being collected or created?</li> </ul> <b>Guidance:</b> <p>Briefly summarise the type of study (or studies) to help others understand the purposes for which the data are being collected or created.</p>
PI / Researcher	Name of Principal Investigator(s) or main researcher(s) on the project.
PI / Researcher ID	E.g. ORCID <a href="http://orcid.org/">http://orcid.org/</a>
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	<b>Questions to consider:</b> <ul style="list-style-type: none"> <li>- Are there any existing procedures that you will base your approach on?</li> <li>- Does your department/group have data management guidelines?</li> <li>- Does your institution have a data protection or security policy that you will follow?</li> <li>- Does your institution have a Research Data Management (RDM) policy?</li> <li>- Does your funder have a Research Data Management policy?</li> <li>- Are there any formal standards that you will adopt?</li> </ul> <b>Guidance:</b> <p>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.</p>
<b>Data Collection</b>	
What data will you collect or create?	<b>Questions to consider:</b> <ul style="list-style-type: none"> <li>- What type, format and volume of data?</li> <li>- Do your chosen formats and software enable sharing and long-term access to the data?</li> <li>- Are there any existing data that you can reuse?</li> </ul> <b>Guidance:</b> <p>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.</p>
How will the data be collected or created?	<b>Questions to Consider:</b> <ul style="list-style-type: none"> <li>- What standards or methodologies will you use?</li> <li>- How will you structure and name your folders and files?</li> <li>- How will you handle versioning?</li> <li>- What quality assurance processes will you adopt?</li> </ul> <b>Guidance:</b> <p>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</p>

# My plan (Horizon 2020 DMP)

0/5 questions answered

approx. 20% of available space used

[Plan details](#)
[Initial DMP](#)
[Mid-term Review DMP](#)
[Final review DMP](#)
[Share](#)
[Export](#)

## Scientific research data should be easily: 1. Discoverable (1 question, 0 answered)

Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier)?

B
I
☰
☷
🔗
📄

Save

Not answered yet

[Guidance](#)
[Share note](#)

### DCC guidance on Discovery by Users

#### Questions to consider:

- How will potential users find out about your data?
- Will you provide metadata online to aid discovery and reuse?

#### Guidance:

Indicate how potential new users can find out about your data and identify whether they could be suitable for their research purposes. For example, you may provide basic discovery metadata online (i.e. the title, author, subjects, keywords and publisher).

## 2. Accessible (1 question, 0 answered)

## 3. Assessable and intelligible (1 question, 0 answered)

## 4. Usable beyond the original purpose for which it was collected (1 question, 0 answered)

## 5. Interoperable to specific quality standards (1 question, 0 answered)

# PLOS – Ten Simple Rules



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](mailto:william.michener@gmail.com)



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

**Copyright:** © 2015 William K. Michener. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**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

### 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, data-bases), 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

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

# How to create a Horizon 2020 DMP?

- H2020 template consists of
  - Data set reference and name
  - Data set description
  - Standards and metadata
  - Data sharing
  - Archiving and preservation



- **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



- 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
  - Dublin Core
  - PREMIS
  - ...

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

- Which data will be shared?
  - final result?
  - intermediate data?
- Where will the data be deposited?
  - not all of the data must be shared in the same way
- Are there any embargo periods?
- Who will have access?



# H2020 DMP - Archiving and preservation

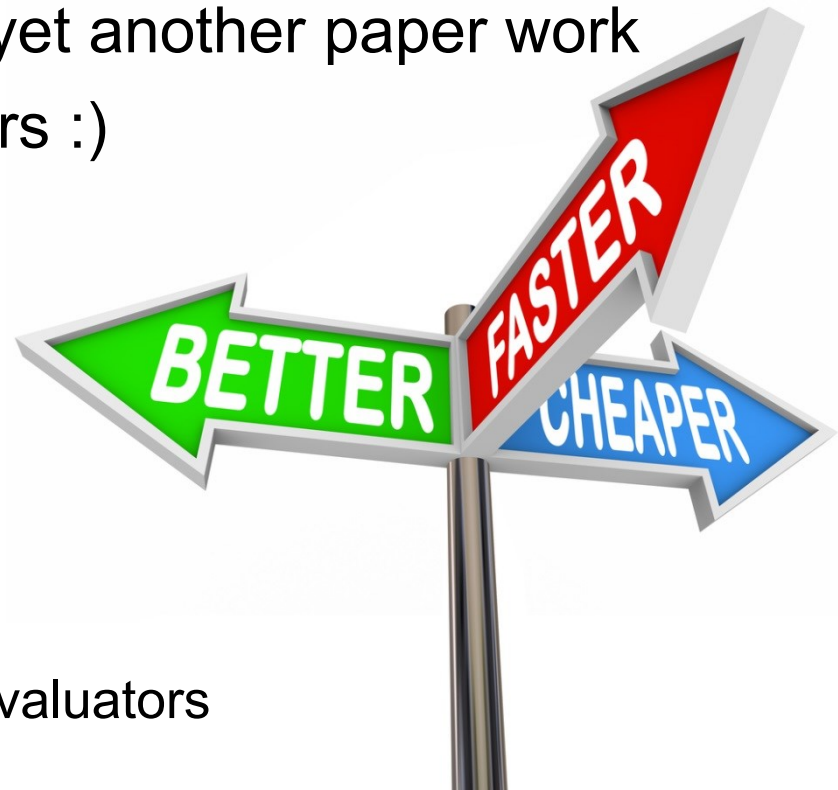
- Which data needs to be preserved?
  - What has to be kept e.g. data underlying publications?
  - What can't be recreated e.g. environmental recordings?
  - What is potentially useful to others?
  - What has scientific, cultural or historical value?
  - What legally must be destroyed?
- For how long?
- What is the cost and who will pay for it?

## Some good advice

- DMP can reveal how solid your research proposal is
- When answering questions from checklists write coherent text
- Be specific when referring to tools and standards
- Assign responsibilities and name responsible personnel
- Discuss any technical issues with the IT personnel
- Manage your data correctly from the very beginning
  - backups, file naming conventions, access restrictions, metadata collection

# Conclusions

- DMPs are NOT meant to be yet another paper work imposed on (poor) researchers :)
- Beneficiaries
  - Researchers
  - Projects applicants
  - Funding bodies and research evaluators
  - Repositories
  - Reusing parties and society as a whole



- 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)
- Registry of Research Data Repositories
  - <http://www.re3data.org>
- EUDAT webinars on data management
  - <https://www.eudat.eu/events/webinar/research-data-management-an-introductory-webinar-from-openaire-and-eudat>