Seminar on Responsible Research and Innovation: Open Science

Isabel Bernal
DIGITAL.CSIC

Achieve ITN Project Check Meeting, CNM-IMS
Seville, March 6, 2019
OPEN SCIENCE DEFINITIONS, COMPONENTS, CHANNELS
Open Science represents a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools.

(European Commission, 2016b:33)
Software Infrastructure and Environments for Reproducible and Extensible Research

- **Open licensing** should be used for data and code
- **Workflow tracking** should be carried out during the research process
- **Data** must be available and accessible
- **Code and methods** must be available and accessible
- **All 3rd party data and software** should be cited

Open Science in Practice

You can make your workflow more open by...

- adding alternative evaluation, e.g. with altmetrics
- communicating through social media, e.g. Twitter
- sharing posters & presentations, e.g. at FigShare
- using open licenses, e.g. CC0 or CC-BY
- publishing open access, ‘green’ or ‘gold’
- using open peer review, e.g. at journals or PubPeer
- sharing preprints, e.g. at OSF, arXiv or bioRxiv
- using actionable formats, e.g. with Jupyter or CoCalc
- open XML-drafting, e.g. at Overleaf or Authorea
- sharing protocols & workfl., e.g. at Protocols.io
- sharing notebooks, e.g. at OpenNotebookScience
- sharing code, e.g. at GitHub with GNU/MIT license
- sharing data, e.g. at Dryad, Zenodo or Dataverse
- pre-registering, e.g. at OSF or AsPredicted
- commenting openly, e.g. with Hypothes.is
- using shared reference libraries, e.g. with Zotero
- sharing (grant) proposals, e.g. at RIO

Bianca Kramer & Jeroen Bosman  https://101innovations.wordpress.com

DOI: 10.5281/zenodo.1147025
# Growth of preprint repositories

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Faster dissemination of results</td>
<td>+ Not all journals accept papers that have been submitted to a preprint server</td>
</tr>
<tr>
<td>+ Establishing primacy</td>
<td>+ Risk of 'preprint wars'</td>
</tr>
<tr>
<td>+ Preprint publishing fosters open science</td>
<td>+ Risk of embargo violations</td>
</tr>
<tr>
<td>+ Evidence of productivity and accomplishment</td>
<td>+ Poor quality and irreproducible data will be posted in preprint form</td>
</tr>
<tr>
<td>+ Visibility of work promotes invitation to meetings</td>
<td>+ Scientists will rush out data prematurely to claim priority and get credit</td>
</tr>
<tr>
<td>+ Feedback on your work and potential for improvements, enhanced quality</td>
<td>+ Scientists will try to &quot;scoop my work&quot; if I post as a preprint</td>
</tr>
<tr>
<td>+ Establishing priority of discoveries and ideas</td>
<td></td>
</tr>
<tr>
<td>+ Potential for developing new collaborations earlier</td>
<td></td>
</tr>
<tr>
<td>+ OA to your work across the globe</td>
<td></td>
</tr>
</tbody>
</table>

- **AgriXiv** (agriculture)
- **bioRxiv** (biology)
- **EarthArXiv** (Earth sciences)
- **engrXiv** (Engineering),
- **LawArXiv** (law)
- **LISSA** (library and information science)
- **MarXiv** (ocean and marine-climate sciences)
- **Mathematics Preprint Servers** (mathematics)
- **NutriXiv** (nutritional sciences)
- **paleorXiv** (palaeontology)
- **PsyArXiv** (psychology)
- **SocArXiv** (social sciences)

[https://www.fosteropenscience.eu/learning/sharing-preprints/#/id/5ac23bcbd1827131b90e79d](https://www.fosteropenscience.eu/learning/sharing-preprints/#/id/5ac23bcbd1827131b90e79d)
Growth of journals that publish data and software tool articles
Open Science services at DIGITAL.CSIC

- Open Access to publications and other traditional outputs
- Open Data (DOI minting)
- Preprints
  - Open Research Software
  - Open Lab Notebooks (DOI minting)
- Open Peer Review and Commentary Modules
- Altmetrics and other indicators
  - Public Profiles (researchers, groups, projects)
A glimpse of research outputs types in DIGITAL.CSIC
OPEN ACCESS AND OPEN DATA AT H2020. ISSUES AT STAKE, COMPLIANCE VIA DIGITAL.CSIC
Basics of H2020 Open Access mandate

**SELF-ARCHIVING**

*‘GREEN’ OPEN ACCESS*

- Deposit the final peer-reviewed manuscript in a repository of your choice.
- *Researchers must ensure open access to the publication within at most 6 months (12 months for publications in the social sciences and humanities).*

**OPEN ACCESS PUBLISHING**

*‘GOLD’ OPEN ACCESS*

- Publish in open access journals or in hybrid journals.
- *Article processing charges are eligible for reimbursement during the duration of your project. Hybrid journals sell subscriptions (i.e. closed access) AND offer the option of making some individual articles open access.*

**BOTH OPTIONS ARE POSSIBLE**

If the gold route is chosen the article must also be deposited in a repository to comply with Article 29.2.
1. **Publication of project’s results in a peer reviewed journal:**
   - Check journal self-archiving policy for repositories
   - Check APCs for Open Access publishing and budget them in a realistic way

2. **Upload into DIGITAL.CSIC:**
   - Through Mediated Archiving Service
   - Description and labelling according to H2020 requirements (Project tracking, vocabularies)
   - Generation of persistent identifier (handle)

3. **Aggregation:**
   - [OpenAIRE logo]
CSIC Support to Open Access Publishing

http://digital.csic.es/dc/accesoAbierto.jsp

http://bibliotecas.csic.es/publicacion-en-acceso-abierto
Basics of H2020 Open data policy

**STEP 1**
WRITE A DMP
dmponline.dcc.ac.uk

**STEP 2**
FIND REPOSITORY
Matches data needs

**STEP 3**
DEPOSIT DATA
(Open) Data
Metadata
Other tools

**SUPPORT**
Supporting infrastructure and information

Deliverable at
- 6 months
- Mid-term review
- Final review

Data Repositories
- discipline/institutional
- www.re3data.org
- Zenodo

- Standard File Formats
- Standards metadata schema
- Open Licences

- EC guidelines
- OpenAIRE.eu
- www.dcc.ac.uk

Designed by Freeok
H2020/ERC Templates for DMPs

ANEX 1
Horizon 2020 FAIR Data Management Plan (DMP) template
Version: 29 July 2016

Introduction

This Horizon 2020 FAIR DMP template has been designed to be applicable to any Horizon 2020 project that produces, collects or processes research data. You should develop a single DMP for your project to cover its overall approach. However, where there are specific issues for individual datasets (e.g. regarding openness), you should clearly spell this out.

FAIR data management

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.

This template is not intended as a strict technical implementation of the FAIR principles, it is rather inspired by FAIR as a general concept.

More information about FAIR:
- FAIR data principles (FORCE11 discussion forum)
- FAIR principles (article in Nature)

Structure of the template

The template is a set of questions that you should answer with a level of detail appropriate to the project.

It is not required to provide detailed answers to all the questions in the first version of the DMP that needs to be submitted by month 6 of the project. Rather, the DMP is intended to be a living document in which information can be made available on a finer level of granularity through updates as the implementation of the project progresses and when significant changes occur. Therefore, DMPs should have a clear version number and include a timetable for updates. As a minimum, the DMP should be updated in the context of the periodic evaluation/assessment of the project. If there are no other periodic reviews envisaged within the grant agreement, an update needs to be made in time for the final review at the latest.

In the following sections, the main sections to be covered by the DMP are outlined. At the end of the document, Table 1 contains a summary of these elements in bullet form. This template itself may be updated as the policy evolves.

ERC OPEN RESEARCH DATA MANAGEMENT PLAN (DMP)

Several datasets may be included into a single DMP.
# Common elements in DMPs

<table>
<thead>
<tr>
<th>Theme</th>
<th>DCC &amp; UC3 Guidance</th>
</tr>
</thead>
</table>
| DATA DESCRIPTION          | • Give a summary of the data you will collect or create, noting the content, coverage and data type, e.g., tabular data, survey data, experimental measurements, models, software, audiovisual data, physical samples, etc.  
• Consider how your data could complement and integrate with existing data, or whether there are any existing data or methods that you could reuse.  
• Indicate which data are of long-term value and should be shared and/or preserved.  
• If purchasing or reusing existing data, explain how issues such as copyright and IPR have been addressed. You should aim to minimise any restrictions on the reuse (and subsequent sharing) of third-party data. |
| DATA FORMAT               | • Clearly note what format(s) your data will be in, e.g., plain text (.txt), comma-separated values (.csv), geo-referenced TIFF (.tiff, .tif).  
• Explain why you have chosen certain formats. Decisions may be based on staff expertise, a preference for open formats, the standards accepted by data centres or widespread usage within a given community.  
• Using standardised, interchangeable or open formats ensures the long-term usability of data: these are recommended for sharing and archiving.  
• See UK Data Service guidance on recommended formats or DataONE Best Practices for file formats. |
| DATA VOLUME               | • Note what volume of data you will create in MB/GB/TB. Indicate the proportions of raw data, processed data, and other secondary outputs (e.g., reports).  
• Consider the implications of data volumes in terms of storage, access and preservation. Do you need to include additional costs?  
• Consider whether the scale of the data will pose challenges when sharing or transferring data between sites; if so, how will you address these challenges? |

- Roles and responsibilities
- Description, volume, structure, standards, formats of data
- Data collection and processing
- Metadata
- Intellectual Property Rights
- General Data Protection Regulation
- Ethical considerations
- Data Access, publishing and reuse
- Repository selection
- Security and storage
- Long term preservation
- Budget
CORE REQUIREMENTS FOR DATA MANAGEMENT PLANS

When developing solid data management plans, researchers are required to deal with the following topics and answer the following questions:

1. Data description and collection or re-use of existing data
   a. How will new data be collected or produced and/or how will existing data be re-used?
   b. What data (for example the kinds, formats, and volumes) will be collected or produced?

2. Documentation and data quality
   a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?
   b. What data quality control measures will be used?

3. Storage and backup during the research process
   a. How will data and metadata be stored and backed up during the research process?
   b. How will data security and protection of sensitive data be taken care of during the research?

4. Legal and ethical requirements, codes of conduct
   a. If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?
   b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?
   c. How will possible ethical issues be taken into account, and codes of conduct followed?

CRITERIA FOR THE SELECTION OF TRUSTWORTHY REPOSITORIES

Trustworthy repositories should meet the following minimum criteria:

1. Provision of Persistent and Unique Identifiers (PIDs)
   a. Allow data discovery and identification
   b. Enable searching, citing, and retrieval of data
   c. Provide support for data versioning

2. Metadata
   a. Enable finding of data
   b. Enable referencing to related relevant information, such as other data and publications
   c. Provide information that is publicly available and maintained, even for non-published, protected, retracted, or deleted data
   d. Use metadata standards that are broadly accepted by the scientific community
   e. Ensure that metadata are machine-readable

3. Data access and usage licences
   a. Enable access to data under well-specified conditions
   b. Ensure data authenticity and integrity
   c. Enable retrieval of data
   d. Provide information about licensing and permissions (in ideally machine-readable form)
   e. Ensure confidentiality and respect rights of data subjects and creators
Finding real data management plans by discipline and funders templates
Compliance via DIGITAL.CSIC

Inform DIGITAL.CSIC as soon as possible
No files limit on the repository

Get ready all basic information: format, size, structure of data, expectations about data management and publishing through DIGITAL.CSIC

DIGITAL.CSIC will ask data creators about nature, processing and other characteristics of datasets. Good documentation and metadata are as important as the data itself.
DIGITAL.CSIC does not impose any specific usage license but preference for Creative Commons/Open data Commons.
Preference for open/embargoed access data but exceptions are accepted.
Data curation: description and license
Research data repositories and disciplines representation

Figure 1: Disciplinary provision of research data repositories
(source: re3data.org – CC-BY, last accessed on 02/01/2018)
FAIR Data Principles

What is FAIR DATA?

- **Findable**
  - Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

- **Accessible**
  - Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

- **Interoperable**
  - Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

- **Reusable**
  - Data and collections have a clear usage licenses and provide accurate information on provenance.

- Guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention)

https://www.nature.com/articles/sdata201618
Measuring FAIRNESS of your data

Self-assessment tool with questions related to the principles underpinning Findable, Accessible, Interoperable and Reusable. You will be given a ‘green bar’ indicator based on your answers in that section, and when all sections are completed, an overall 'FAIRness' indicator is provided.
FAIR data and European Open Science Cloud (EOSC)

- EOSC formal launch on November 22, 2018
- EOSC has a mandate to enable Europe’s 1.7 million researchers to easily share scientific data and software tools in an open and collaborative environment.

- EOSC shared resources need to be developed to cover all the aspects of FAIR data:
  - Findable, through e.g. catalogues of data/services and metadata;
  - Accessible, through e.g. Persistent Unique Identifiers, Data Management Plans;
  - Interoperable, through e.g. interoperable standards and common metadata;
  - Reusable, through e.g. common IPR and legal provisions (e.g. Creative Commons).

https://sciencebusiness.net/report/priorities-european-open-science-cloud
CSIC will lead an international consortium to make headways on EOSC infrastructure, enhance computational capabilities, contribute datasets from DIGITAL.CSIC and train on data management related skills.

CSIC participates in EPOS, European research infrastructure on Solid Earth and contributes data via DIGITAL.CSIC.
RESEARCH SOFTWARE AND HARDWARE
Software Management Plan: guidance and template

- What software will you develop?
- Who are the intended users of your software?
- How will you make your software available to your users?
- How will you support those who use your software?
- How will your software contribute to research?
- How will your software relate to other research objects?
- How will you measure your software's contribution to research?
- Where will you deposit your software to guarantee its long-term availability?
- https://zenodo.org/record/2159713#.XHl7AvlKi1s
How to describe research software: international standards

Description of research software should be detailed enough so that software may be qualified as “TRIFID”: Testable, Robust, Installable, Findable, Identifiable, Documented

Choosing appropriate file formats

File formats best suited for long term sustainability and accessibility:

- Are frequently used.
- Have open specifications.
- Are independent of specific software, developers or vendors.
Tools to choose a software license

Rights in Copyright

<table>
<thead>
<tr>
<th>Public Domain</th>
<th>Non-Protective FOSS License</th>
<th>Protective FOSS License</th>
<th>Proprietary License</th>
<th>Trade Secret</th>
</tr>
</thead>
<tbody>
<tr>
<td>All rights relinquished</td>
<td>more rights granted</td>
<td>more rights retained</td>
<td>All rights retained</td>
<td></td>
</tr>
</tbody>
</table>

Software licenses in context of copyright according to Mark Webbink
Finding resources of interest

Data standards, databases, policies
https://fairsharing.org/

Training materials, workflows and events
https://tess.elixir-europe.org/
Open Source Hardware

- Open source hardware is hardware whose design is made publicly available so that anyone can study, modify, distribute, make, and sell the design or hardware based on that design. The hardware’s source, the design from which it is made, is available in the preferred format for making modifications to it. Ideally, open source hardware uses readily-available components and materials, standard processes, open infrastructure, unrestricted content, and open-source design tools to maximize the ability of individuals to make and use hardware.

The distribution terms of Open Source Hardware must comply with the following criteria:

1. Documentation
2. Scope
3. Necessary Software
4. Derived Works
5. Free redistribution
6. Attribution
7. No Discrimination Against Persons or Groups
8. No Discrimination Against Fields of Endeavor
9. Distribution of License
10. License Must Not Be Specific to a Product
11. License Must Not Restrict Other Hardware or Software
12. License Must Be Technology-Neutral

Examples:
OpenDrop
CERN Open Hardware Repository

https://www.oshwa.org/definition/
Towards Open Science Hardware

Open Source Hardware Checklist

Make sure your project is heading in the right direction!

This checklist is made up from key points of the Open Source Hardware Definition and Best Practices for a quick reference guide to ensure your project is being properly labeled as Open Source Hardware.

- Does your hardware comply with the open source hardware definition at http://www.oshwa.org/definition/
- Have you allowed anyone to study, modify, distribute, make, and sell the hardware?
- If you used a Creative Commons license for your source files (documentation), did you choose options compatible with the definition? Non-Commercial and No Derivatives are not open source.
- Did you put the oshw logo on your hardware so people can easily identify it as open source hardware? (Strongly recommended)
- Do all company logos on the hardware belong to you? Do not infringe on trademarks!
- Are your source files in an easily attainable format?
- Are the source files publicly available online?
- Are your source files easy to find, for example, linked to from the product page?
- Have you documented your project in a way that people will be able to copy?
- Is your documentation free of charge?
- Have you included images in your documentation? (Strongly recommended)
- Are you emotionally prepared to allow your project to be copied?
- If not all parts/versions are open, have you clearly specified which portions of the design are being released as open source hardware and which are not?
Open hardware licensing

Open hardware licensing is complex, multifaceted, and heterogeneous.

Copyright (on which most licenses are based) doesn’t apply to hardware itself, only to the design files for it – and, then, only to the elements which constitute “original works of authorship” and not the underlying functionality or ideas.

Hardware is mostly beyond the scope of copyright, and open content licenses do not address problems with distribution and redistribution of the physical hardware.

Popular copyleft licenses include:
- Creative Commons Attribution, Share-Alike (BY-SA)
- GNU General Public License (GPL)
- Hardware-Specific Licenses: TAPR OHL, CERN OHL

Permissive licenses include:
- FreeBSD license
- MIT license
- Creative Commons Attribution (BY)
- Hardware-Specific License: Solderpad Hardware License
Facilitating transfer, access, credit and reuse of lab materials

• OpenMTA enables broader sharing and use of biological materials by biotechnology practitioners working within the practical realities of technology transfer
• Materials available under the OpenMTA are free of any royalty or fees, other than appropriate and nominal fees for preparation and distribution.
• Providers may request attribution and reporting for materials distributed under the OpenMTA.
• Materials available under the OpenMTA may be modified or used to create new substances.
• The OpenMTA does not restrict any party from selling or giving away the materials.
• The OpenMTA supports the transfer of material between researchers at all types of institutions, including academic, industry, government and community laboratories.
• Article Opening options for material transfer
• https://www.youtube.com/watch?v=zt6b8_mY-i4
THANKS FOR LISTENING!