Datamanagment, OpenData

In

Horizon 2020

Björn Usadel
RWTH Aachen University and Forschungszentrum Jülich

Originally a trained biochemist, moved to Bioinformatics early coordindator of the German Plant Primary Database (>12 year project funding, then Helmholtz)

FW7 DROPS: WP Leader
FW7 EPPN: WP Leader
H2020 Goodberry: WP Leader

H2020 EPPN2020 (provisional): Co-Leader IT
ESFRI EMPHASIS : Co –Leader IT

Co leading Germany +France

Why data management ?

Why data management ?

Olivier Voinnet

From Wikipedia, the free encyclopedia

Olivier Voinnet (*1973)[1] is a French biologist and (currently) professor of RNA biology at the ETH Zurich.[2][3] Voinnet obtained his PhD in 2001 in England in the group of David Baulcombe and later obtained a position as independent group leader at the CNRS in Strasbourg where he was promoted to *Directeur de Recherche* in 2005. In 2010, he moved to ETH Zurich where he was appointed full professor of RNA Biology.[1][2]

Contents [hide]

"**The original lab book data** provided to the editors of *Science* showed that these errors did not alter the data in any material way…"

Manipulation investigated [edit]

At 2015, his work was investigated for manipulation.[4] The investigation at ETH Zurich found that the scientist "breached his duty of care in the handling of figures as well as in his supervisory duties as a research director... and will receive an admonition in relation to his conduct" but also concluded that "this is not a case of scientific misconduct as defined in ETH Zurich's Rules of Procedure".[5] Another, independent, investigation by CNRS established "the existence of deliberate chart/diagram manipulations, in breach of the ethical standards applicable to the presentation of scientific results" and went on to say that such "inappropriate presentation of experimental data, however, does not amount to fabrication."[6]

; Literal citation Erratum; *Science* 22 Jan 2016: Vol. 351, Issue 6271, DOI: 10.1126/science.aaf2336

Why data management ?

- Make best use of data and share data

- Make sure all partners have access to the data

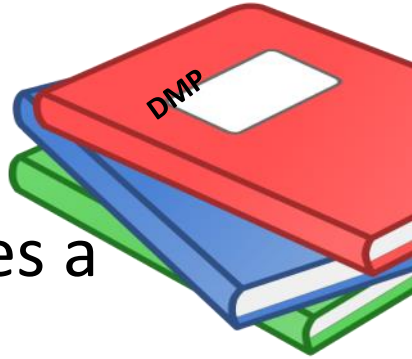
  (Usually EU projects feature large consortia and last a long time, one has to make sure that everyone can use and re-use data)

- Documentation and archival

- Make sure data is not lost, and bring new lab members up to speed

- Good Laboratory Praxis !

EU H2020 The data management plan



- The EU makes things easy, as it [usually] requires a data management plan (DMP)

> "Horizon 2020 have produced guidelines for Research Data Management [PDF 151KB]. All project proposals submitted to 'Research and Innovation actions' as well as 'Innovation actions' should include a section on research data management."

- If properly written and thought through the plan will actually help you structure your thoughts and potentially make the science of your proposal better

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

EU H2020 The data management plan

- What you need to do

"When to write and revise your the Data Management Plan
The first version of the DMP is expected to be delivered **within the first 6 months** of the project. More elaborated versions of the DMP can be delivered at later stages of the project.
The **DMP should be updated as a minimum in time with the periodic evaluation/assessment of the project**. […] **the consortium can define a timetable for review in the DMP itself.**
New versions of the DMP should be created whenever important changes to the project occur due to inclusion of new data sets, changes in consortium policies or external factors."

- What you probably should do

Think through at least some points before grant submission!

https://www.openaire.eu/opendatapilot-dmp

Data Management plan at submission stage

However, **good research data management as such should be addressed under the impact criterion**, as relevant to the project. Your application should address the following issues:

- What standards will be applied?
- How will data be exploited and/or shared/made accessible for verification and reuse? If data cannot be made available, why?
- How will data be curated and preserved?
- Your policy should also:
  reflect the current state of consortium agreements on data management

Participating in the ORD Pilot does **not necessarily mean opening up all your research** data. […]

be consistent with exploitation and Intellectual Property Rights (IPR) requirements

Use common sense!

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

EU H2020 Writing the proposal

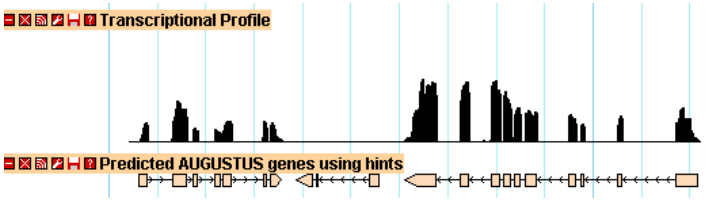
In section "2.2 - Measures to maximise impact" one should already mention how data is treated. This covers

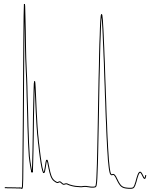
- What types of data will the project generate/collect?
- What standards will be used?
- How will this data be exploited and/or shared/made accessible for verification and re-use. If data cannot be made available, explain why.
- How will this data be curated and preserved?

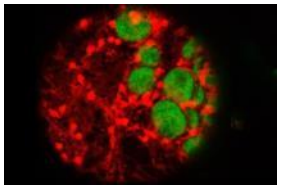
Section 2.2. Measure to maximise impact

- **What types of data will the project generate/collect?**

  Possible items can be e.g. qRT PCR results, microarray studies, chromotographics traces, marker enzymes, next generation sequencing data*, health profiles*, pictures and images….
  But also textual data such as patient interviews





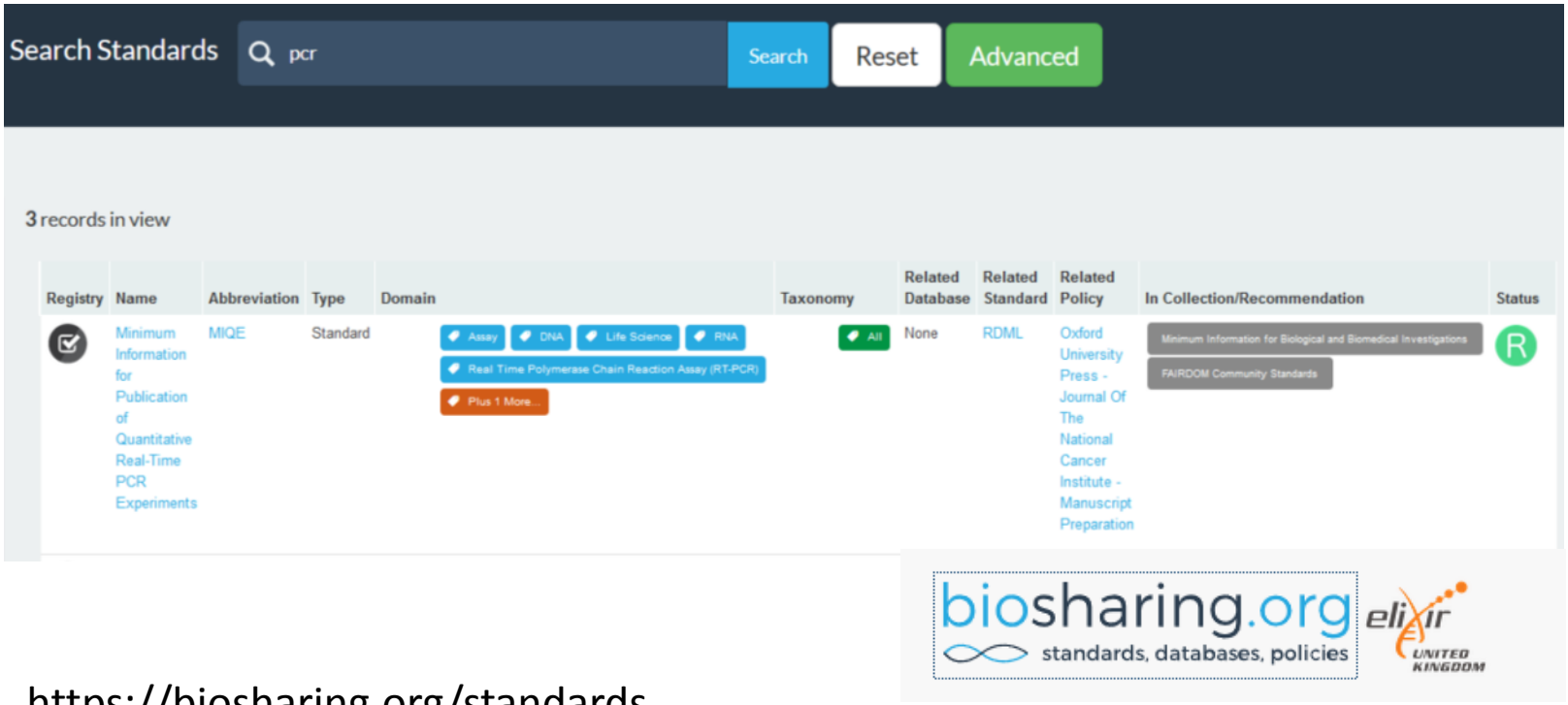


* Be careful about the ethical implications

Section 2.2. Measure to maximise impact

- **What standards will be used?**

For many molecular biology derived datasets standards exist! For example you might have heard about the whole set of MIAXXXX (Minimal information about ….) Many standards can be found on biosharing.org

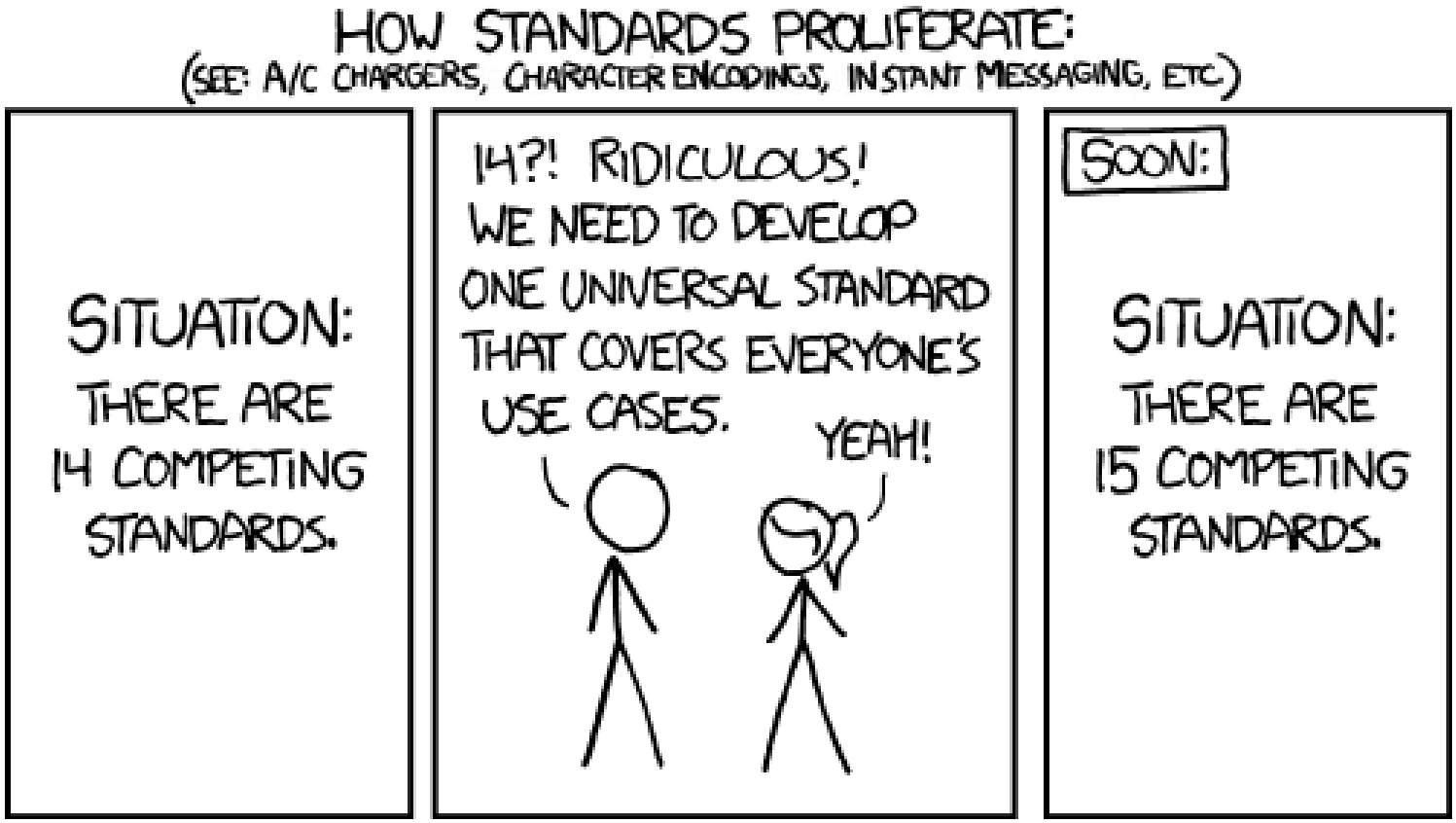


https://biosharing.org/standards

Section 2.2. Measure to maximise impact

- **What standards will be used?**

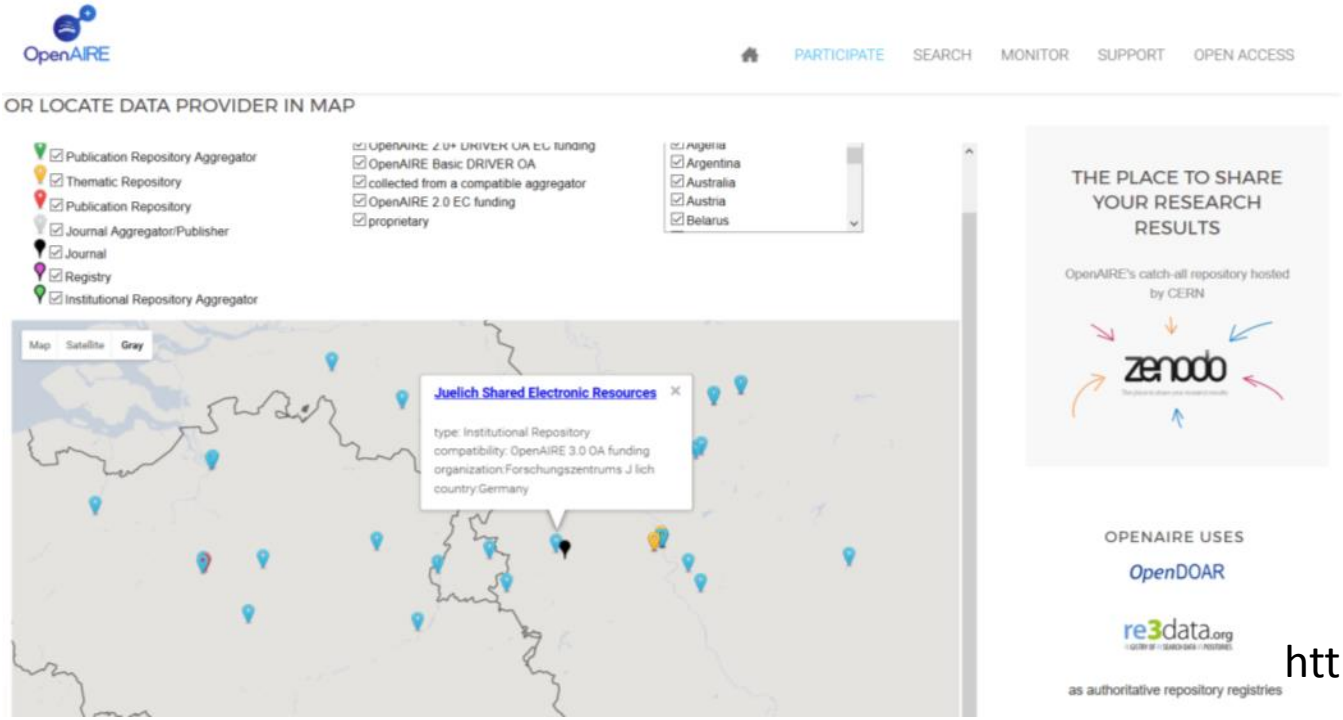
Make sure the standard is suitable for your research! Say “standards like MIQE” “MIAPPE 1.0 or later” especially if the field is still evolving



Section 2.2. Measure to maximise impact

- **How will this data be exploited and/or shared/made accessible for verification and re-use. If data cannot be made available, explain why.**

  Likely you will share the data (➔OpenData), think of ways how you can make the data available to others. The best is to make use of public, established resources. Check openaire.eu



https://www.openaire.eu

Section 2.2. Measure to maximise impact

- **How will this data be exploited and/or shared/made accessible for verification and re-use. If data cannot be made available, explain why.**

  Nature Scientific Data has a brief overview of some important repositories

urages authors to archive data to one of the above data-type specific reposit
specific repository is not available, we recommend the following generalist re
y of data. Generalist repositories may also be appropriate for archiving assoc
ol data, supplementing the primary data in a data-type specific repository.

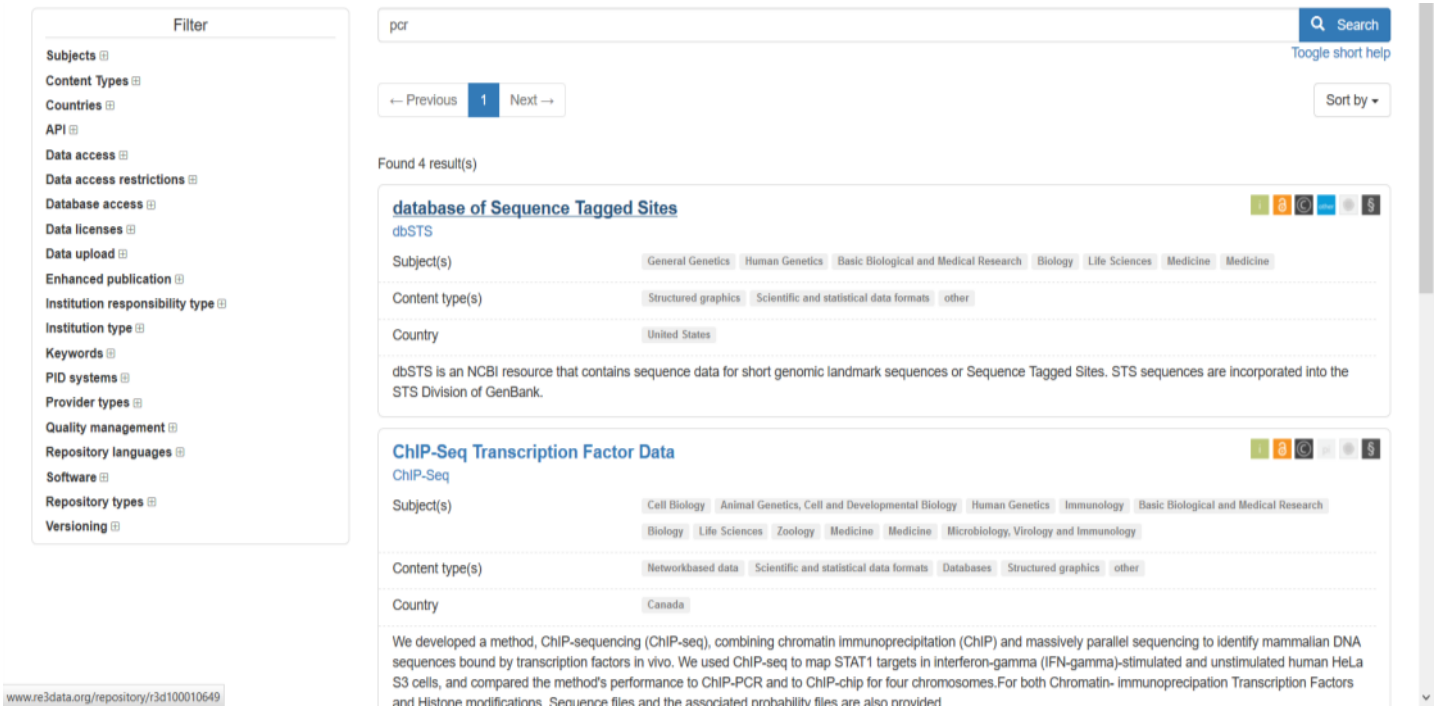
formation on es/costs	Size limits	Integrated with *Scientific Data* manuscript submission system
20 USD for first 20 GB, nd $50 USD for each lditional 10 GB	None stated	Yes ✓
0 GB free per *Scientific ata* manuscript. lditional fees apply for rger datasets	1 TB per dataset	Yes ✓ - To qualify for the 100 G free storage, data must be uploaded to figshare via our submission system. Download instructions.
ontact repository for tasets over 1 TB	2.5 GB per file, 10 GB per dataset	No

http://www.nature.com/sdata/policies/repositories

Section 2.2. Measure to maximise impact

- **How will this data be exploited and/or shared/made accessible for verification and re-use. If data cannot be made available, explain why.**

Re3data allows you search and find repositories



Zenodo a generalist repository

zenodo | Search | Upload | Communities | Log in | Sign up

February 27, 2017

Dataset | Open Access

Dataset for "Soil fluxes of carbonyl sulfide (COS), carbon monoxide, and carbon dioxide in a boreal forest in southern Finland"

Sun, W.; Kooijmans, L. M. J.; Maseyk, K.; Chen, H.; Mammarella, I.; Vesala, T.; Levula, J.; Keskinen, H.; Seibt, U.

This is the dataset (ver. 2017.02.13) for the manuscript "Soil fluxes of carbonyl sulfide (COS), carbon monoxide, and carbon dioxide in a boreal forest in southern Finland" submitted to the journal *Atmospheric Chemistry and Physics*.

Preview

hyy15_chflux_20170213.zip

hyy15_blank.csv	1.3 kB
hyy15_chflux_release.csv	1.3 MB
hyy15_moss.csv	2.9 kB
readme.md	4.4 kB
readme.pdf	462.1 kB

Publication date:
February 27, 2017

DOI:
DOI 10.5281/zenodo.322936

Keyword(s):
carbonyl sulfide | carbon monoxide | soil-atmosphere gas exchange | boreal forest

Grants:
European Commission:
- INGOS - Integrated non-CO2 Greenhouse gas Observing System (284274)

Related identifiers:
Cited by:
10.5194/acp-2017-180

Communities:
European Commission Funded Research (OpenAIRE)
Zenodo

Share

Cite as

Sun, W., Kooijmans, L. M. J., Maseyk, K., Chen, H., Mammarella, I., Vesala, T., … Seibt, U. (2017).

Section 2.2. Measure to maximise impact

- **How will this data be curated and preserved?**
  If you can use one of the previously mentioned providers ➔easy use them.

- Is my data safe with you / What will happen to my uploads in the unlikely event that Zenodo has to close?

  Yes, your data is stored in CERN Data Center. Both data files and metadata are kept in multiple online replicas and independent replicas. CERN has considerable knowledge and experience in building and operating large scale digital repositories and a commitment to maintain this data centre to collect and store 100s of PBs of LHC data as it grows over the next 20 years. In the highly unlikely event that Zenodo will have to close operations, we guarantee that we will migrate all content to other suitable repositories, and since all uploads have DOIs, all citations and links to Zenodo resources (such as your data) will not be affected.

From the zenodo website

EU H2020 Writing the proposal

Add (milestones and) deliverables for the DMP! By default you have to have a DMP.

If you have a workpackage about data management already, it makes a lot of sense to add it there and like this you can reiterate some points in the project summary

Task 5.4 Data Management Plan
To safeguard open access and sustainable access to GoodBerry data, a data management plan will be elaborated. This will include SOPs to measure traits as detailed in D1.1 and D4.5 but will also deal with ontologies, general best practices, minimal requirements such as MIAMET, MINSEQe etc. data storage, backups and data accessibility (see also Task 5.1)(**D5.1** and **D5.5**).
Partners involved: RWTH AACHEN
Duration: month 1 – month 48

Deliverables

N°	Brief description	Month of delivery
D5.1	Data Management Plan first version R1.0	6

EU H2020 Writing the proposal

- Check your environment! Maybe your institution /library can help you! As they are already involved in data management schemes.

- Make sure you also allocate necessary resources!

Summary: Writing the proposal

- Think about data requirements, plan what kind of data is going to be important and how much you will have
- Familiarize yourself with standards (probably you know them already anyway)
- See if there are already data repositories/standards tackling exactly what you need

Writing the DMP

H2020 Programme

Guidelines on

FAIR Data Management in Horizon 2020

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

This also includes a „template“

DMP EU Template

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')?

Most questions you can answer from the grant proposal.

For some sets to whom data might be useful: it might be just the consortium. If you explain this properly this is a just answer.
But be creative there is often many more uses

DMP EU Template

2. The FAIR principle

The FAIR Guiding Principles

- **To be Findable**
- **To be Accessible**
- **To be Interoperable**
- **To be Reusable**

https://www.force11.org/group/fairgroup/fairprinciples

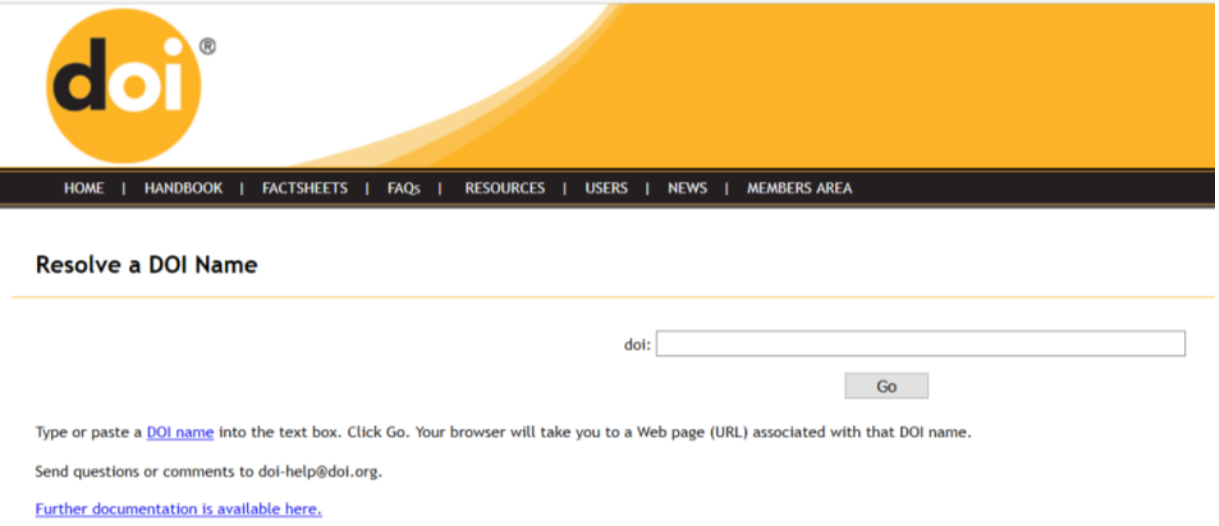
To be Findable

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.



What is this?
What was done?
Who did it ?
How was it done?
Who submitted?

https://guidelines.openaire.eu/

https://www.force11.org/group/fairgroup/fairprinciples

To be Accessible

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

To be Interoperable

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

"Metadata being machine readable is a *conditio sine qua non* for FAIRness."

Use e.g. ontologies or controlled vocabularies

To be Reusable

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

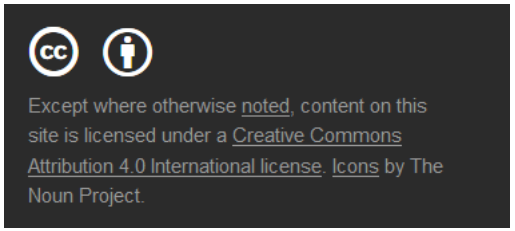
R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.

Try out our simple *License Chooser*.

Three "Layers" Of Licenses

From https://creativecommons.org/licenses/



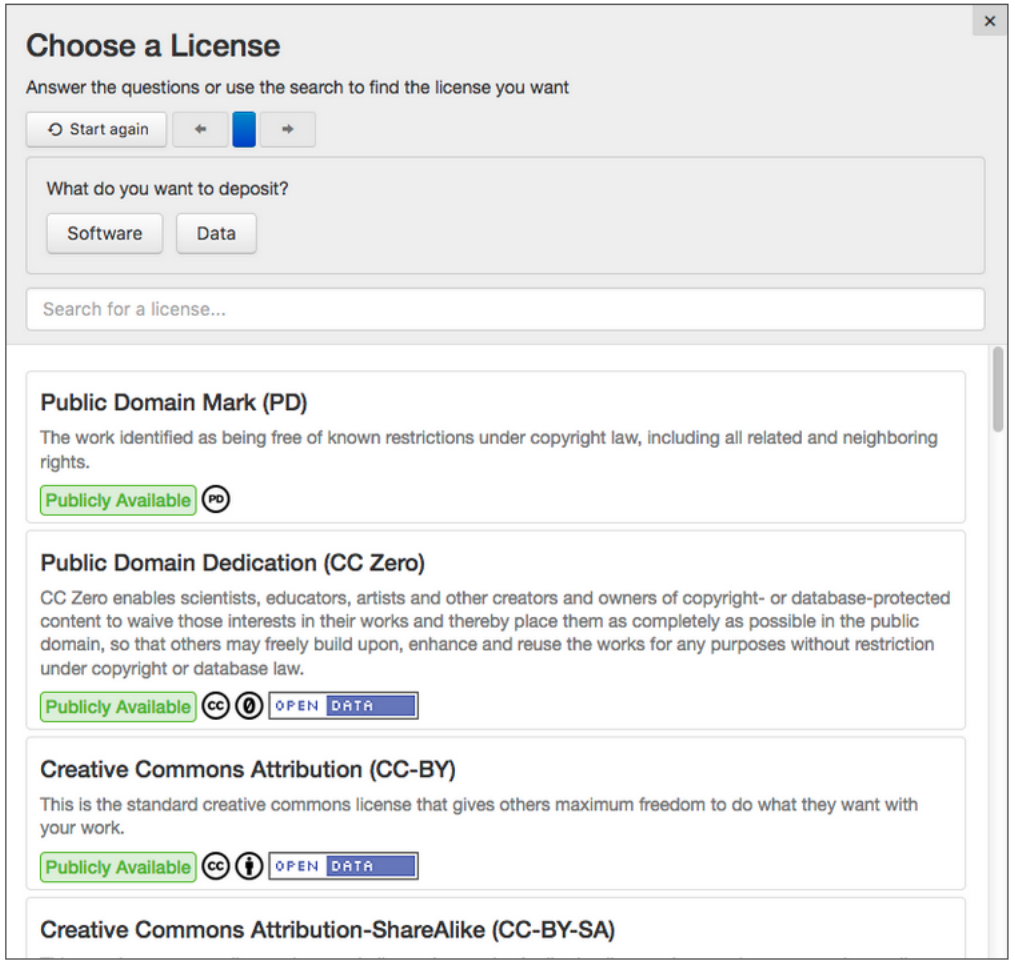
https://www.force11.org/group/fairgroup/fairprinciples

Selecting a license

Figure 1: Locating the license selector in the B2SHARE service.

Usage

During usage, the tool will ask the user some questions which will narrow down the possibilities of license choice.

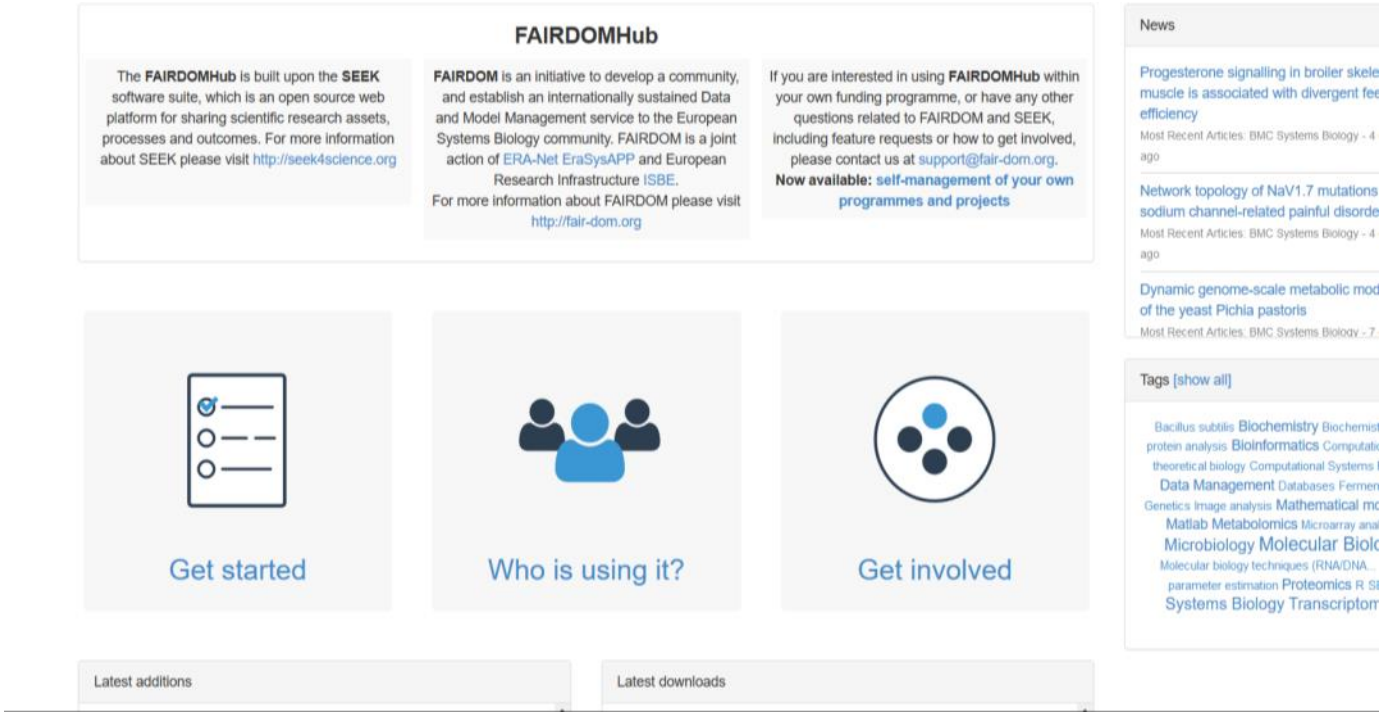


https://eudat.eu/services/userdoc/license-selector

FAIR: sounds unfairly difficult and technical?

Many repositories are FAIR already (or come close)

Don‘t reinvent the wheel!



An example plant phenotyping: Discussions started a long time ago

transPLANT

Standardisation of Plant Phenotyping Experiment Description

DOWNLOADS | ISA-TAB FOR PHENOTYPING | MIAPPE | STANDARDISATION GROUP

working group

Search ...

PAGES

Downloads
ISA-Tab for Phenotyping
MIAPPE
Standardisation group

RECENT POSTS

A new paper published
November 24, 2016

Configuration change
April 1, 2016

Configuration change
September 1, 2015

Working on MIAPPE
August 24, 2015

Opinion paper published
June 15, 2015

RECENT COMMENTS

MINIMUM INFORMATION

OPINION PAPER PUBLISHED

JUNE 15, 2015 HCWI

Following discussions of transPLANT partners and EPPN representatives, an opinion paper on the need of standardisation in the field of plant phenotyping has appeared:

Pawel Krajewski*, Dijun Chen, Hanna Cwiek, Aalt D.J. van Dijk, Fabio Fiorani, Paul Kersey, Christian Klukas, Matthias Lange, Augustyn Markiewicz, Jan Peter Nap, Jan van Oeveren, Cyril Pommier, Uwe Scholz, Marco van Schriek, Bjoern Usadel, and Stephan Weise

Towards recommendations for metadata and data handling in plant phenotyping

Journal of Experimental Botany, 2015. doi:10.1093/jxb/erv271

GOOD PRACTICE | MINIMUM INFORMATION | STANDARDISATION

PREVIOUS POST

Describing the file format is necessary

transPLANT

Standardisation of Plant Phenotyping Experiment Description

DOWNLOADS | ISA-TAB FOR PHENOTYPING | MIAPPE | STANDARDISATION GROUP

working group

Search ...

PAGES

Downloads
ISA-Tab for Phenotyping
MIAPPE
Standardisation group

RECENT POSTS

A new paper published
November 24, 2016
Configuration change
April 1, 2016
Configuration change
September 1, 2015
Working on MIAPPE
August 24, 2015
Opinion paper published
June 15, 2015

RECENT COMMENTS

FORMAT, MINIMUM INFORMATION

A NEW PAPER PUBLISHED

NOVEMBER 24, 2016 HCWI

Following the opinions expressed in Krajewski et al., we are presenting a new paper summarising the solutions proposed for the improvement of phenotypic data description:

Hanna Ćwiek-Kupczyńska, Thomas Altmann, Daniel Arend, Elizabeth Arnaud, Dijun Chen, Guillaume Cornut, Fabio Fiorani, Wojciech Frohmberg, Astrid Junker, Christian Klukas, Matthias Lange, Cezary Mazurek, Anahita Nafissi, Pascal Neveu, Jan van Oeveren, Cyril Pommier, Hendrik Poorter, Philippe Rocca-Serra, Susanna-Assunta Sansone, Uwe Scholz, Marco van Schriek, Ümit Seren, Björn Usadel, Stephan Weise, Paul Kersey and Paweł Krajewski

Measures for interoperability of phenotypic data: minimum information requirements and formatting

Plant Methods, 2016. DOI 10.1186/s13007-016-0144-4

BEST PRACTICES | STANDARDISATION

Next steps new workshops, (re) defining ontologies and needs

Minimum Information About a Plant Phenotyping Experiment (MIAPPE)

Attributes (concepts, subconcepts - in terms of ontology) marked by asterisk (*) are essential for a description of experiment (e.g. by Poorter et al. [26]); the rest forms an extended description. For some attributes examples of possible values are listed.

Checklist section	Attributes	Source list / Biosharing ID / Reference	Recommended ontologies
General metadata	Unique identifier* Title* Description* Submission date Public release date Publications Laboratory address and contact details	Default ISA-Tab configuration [1]	OBI, Ontology for Biomedical Investigations [2] CRO, Crop Research Ontology [3]
Timing and location	Timing: Start of experiment (date)* Duration (days/months/years)* Experiment location: Geographic location* Latitude and longitude Altitude Inclination and aspect Habitat	Poorter et al. [4] Morrison et al. [5] CIMR [6]: Environmental Analysis Context [7]	OBI, Ontology for Biomedical Investigations [2] GAZ, Gazetteer [9]
Biosource	Organism (taxon)* Infraspecific_name* Infraspecific_rank Common name	MIxS Plant-associated environmetal package [10] Yilmaz et al. [11]	UNIPROT Taxonomy [13] NCBI Taxonomy [14]

Last Published: 2017-01-12 | Version: 2.4.8

INTRODUCTION

- Welcome
- Download & Source Code

APPLICATIONS

- Running Repositories
- Demos and Samples

SCIENTIFIC PUBLICATIONS

- Paper - Talks - Poster

DOCUMENTATION

- Development
- Run

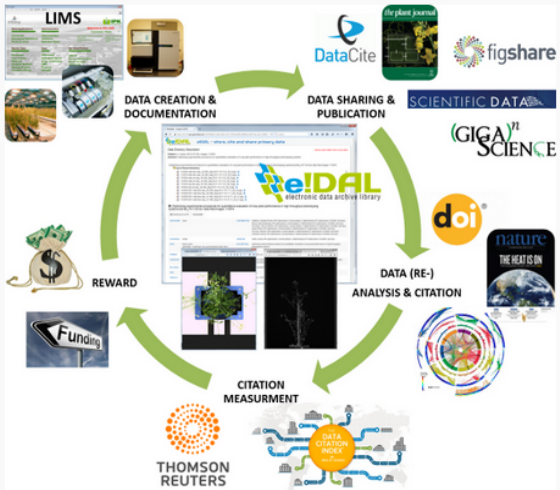
ABOUT

- License and

e!DAL-MetaData-API - store, cite and share primary data

e!DAL is a lightweight software framework for publishing and sharing research data. Its main features are version tracking, metadata management, information retrieval, journal and founding agency proven registration of persistent identifiers (DOI), an embedded HTTP(S) server for public data access, access as a network file system, and a scalable storage backend.

In any research that make use of the e!DAL components **please honour the author's work** and cite:
Arend et al. e!DAL - a framework to store, share and publish research data



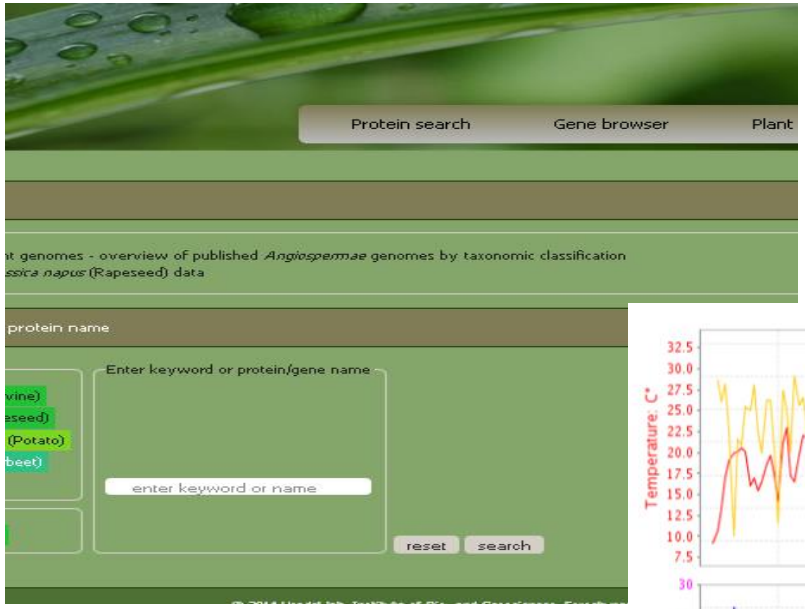
Public Data Repositories

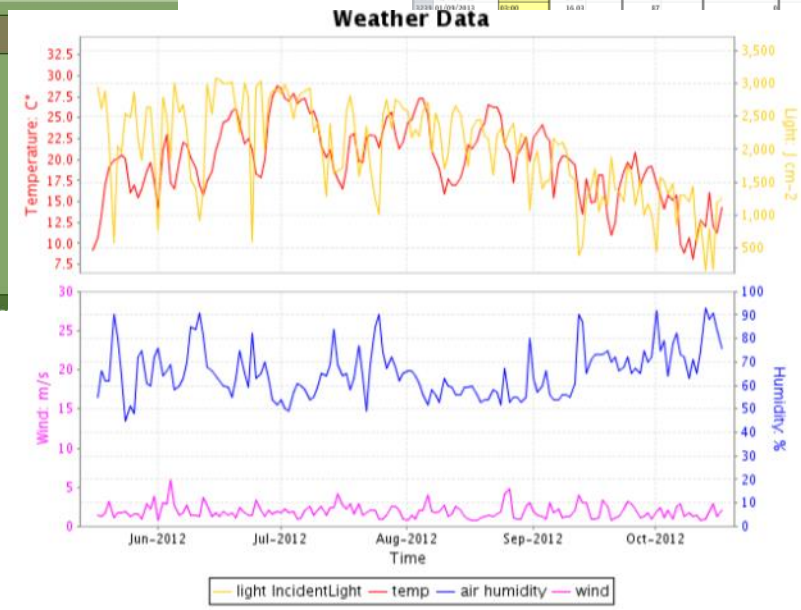
e!DAL driven repositories publish already a high number of scientific citable and Data Cite registered research data. Registered users may submit data sets to the related repository.

Personal Research Data Repository

Packaged as e!DAL server, all required API components are compiled as executable JAR archive. This can be executed at any platform to operate an own data publication infrastructure. Please follow the instructions to set-up and execute an e!DAL server. Java projects may access a e!DAL Server using remote client-API.

You might still use your own database analysis tools to integrate data





Plabipd already covers the genomic base, within PROGRESS additional data

Summary DMP: FAIR

- Making data FAIR without any knowledge in the field can be difficult (but it is not impossible!)
- If you can, use acknowledged repositories and existing standards.
- Potentially liase/explore with repositories

(Maybe EGI, EUDAT or the Research data alliance RDA can help)

DMP template

3. Allocation of resources

- What are the costs for making data FAIR in your project?

- How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

- Who will be responsible for data management in your project?

- Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Be aware you can claim costs, but data should be preserved long term also when the project is over! Make sure to be able to keep websites and domain names as well. Do this in your own interest. (In the future somebody might search your EU project website and find …. something unexpected)

$$$ Who pays?

Plant genomes / Solanum pennellii (new cultivar)

WEB CONTENT DISPLAY

Sequencing the gigabase plant genome of the wild tomato species *Solanum pennellii* using Oxford Nanopore single molecule sequencing

Contributors

Maximilian Schmidt[1], Alexander Vogel[1], Alexandra Wormit[1], Alisandra Denton[1], Anthony Bolger[1], Henri van de Geest[2], Benjamin Istace[6], Marie E. Bolger[3], Saleh Alseekh[4], Janina Maß[3], Christian Pfaff[3], Ulrich Schurr[3], Jean-Marc Aury[6], Alisdair R. Fernie[4], Dani Zamir[5], Björn Usadel[1,3].

[1]Institute for Botany and Molecular Genetics, BioEconomy Science Center, RWTH Aachen University, Aachen, Germany.

[2]Wageningen Plant Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands

[3]Institute for Bio- and Geosciences (IBG-2: Plant Sciences), Forschungszentrum Jülich, Jülich, Germany.

[4]Department of Molecular Physiology, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany.

[5]Faculty of Agriculture, Hebrew University of Jerusalem, Rehovot, Israel.

[6]Genoscope (CEA) and UMR 8030 CNRS-Genoscope-Université d'Evry, 2 rue Gaston Crémieux, BP5706, 91057 Evry, France.

Background

One dataset 6 weeks of lab work… more than 50TB of data.

By current measures and standards

ALL THIS IS RAW DATA THAT NEEDS TO BE PRESERVED

In this case it is 3rd generation sequencing data, so it should be no problem

DMP template

4. Data security

- What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

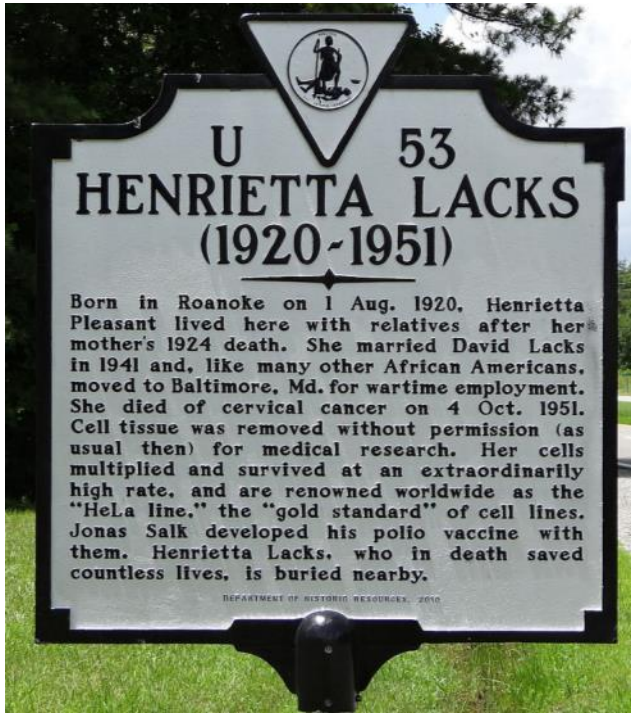
- Is the data safely stored in certified repositories for long term preservation and curation?

DMP template

5. Ethical aspects

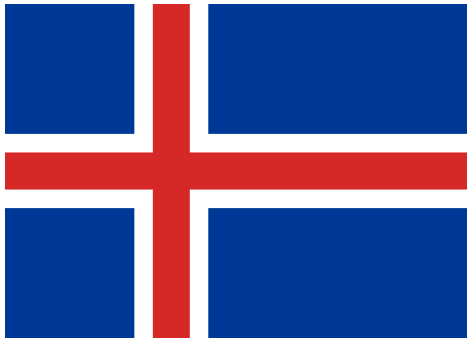
- Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

- Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?



Large scale DNA is never anonymous!

HeLa cell line data published
The family was not amused!



According to an interview with technologyreview Decode Genetics can infer almost all Icelander's Genomic make up . But can not warn due to Ethics rules

DMP template

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

News | Projects | Structures | Services for Researchers | Research Data Management

Research › Research Data Management › Research Data Management at RWTH Aachen

Research Data Management at RWTH Aachen

- Leitlinie zum Forschungsdatenmanagement an der RWTH Aachen
- What Are Research Data?
- Why Manage Data?
- What Does My Funder or Publisher Require?
- How Do I Get Help?
- Contact
- Governance

Research Data Management at RWTH Aachen

Digital data is both a product and an object of research. Researchers have the responsibility to protect and secure this research data and to safeguard its accurate processing. The appropriate management of research data is of central importance for the attractiveness and competitiveness of a university.

For this reason, RWTH Aachen seeks to support its researchers in their research data management activities, based on binding standards and processes. The University offers advising and training sessions as well as infrastructural services in order to ensure efficient and legally informed reseach work.

CONTACT

ServiceDesk Research Data Management

+49 241 80 24680

Send Email

What is Research Data Management?

Research Data Management is a form of project management and workflow control focusing on the handling of research data in science and academia. It is a uniform process that implements methods and processes to ensure the long-term re-usability of research data as well as to organize and control work processes as efficiently as possible. In order to document the methods and processes to be used, it is advisable to create a data management plan right

DMP Examples and Resources

Checklist for a Data Management Plan

The DCC synthesises requirements for Data Management Plans and best practice within the wider community. This allows us to provide a Checklist that presents the main questions or themes that researchers may want to cover when writing a DMP.

In 2013 the DCC reviewed and shortened its Checklist. The current version is available to download below.

Checklist for a Data Management Plan (v.4.0, 2014)

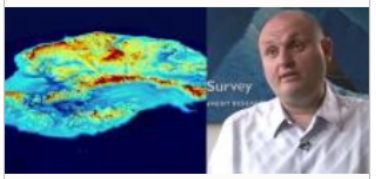
Also see the DMP Checklist flyer, a handy foldout version of the Checklist. Hard copies are available if you would like some for events.

**** This publication is available in print and can be ordered from our online store ****

Earlier versions

The DCC first created a Content Checklist for a Data Management Plan in 2009. This was put out for consultation and over the years it developed into a comprehensive list of questions that researchers may wish to cover.

Managing Research Data (video)



This short documentary, Digital Curation Centre: Managing Research Data, offers a unique insight into the importance of providing access to research data and the risks of not managing data effectively.

Watch this video

http://www.dcc.ac.uk/resources/data-management-plans Attribution to DCC

In this section

DMPonline

Research funders and organisations increasingly require data management plans, both during the bid-preparation stage and after funding has been secured.

DMPonline is the DCC's data management planning tool. It provides tailored guidance and examples to help researchers write data management plans.

The tool includes a number of templates for funders in the UK and overseas so researchers can write DMPs according to the specific requirements they need to meet. It can also be customised by institutions so they can add their own templates and guidance.

A screencast provides an overview of how the tool works.

Try the tool for yourself at http://dmponline.dcc.ac.uk

Anyone can use DMPonline. If your organisation is not listed, just select 'other organisation' or ask for it to be added.

If you would like to create a foreign language version of DMPonline, please contact us on dmponline@dcc.ac.uk

To keep up with DMPonline news, subscribe to the RSS feed or watch GitHub for updates.

All our code is open and available for you to redistribute and/or modify it under the terms of the GNU Affero General Public License.

Useful links

DMPonline

IDCC

The 12th International Digital Curation Conference (IDCC) will take place at The Royal College of Surgeons of Edinburgh, UK

20 - 23 February 2017.

Read more

http://www.dcc.ac.uk/dmponline Attribution to DCC CC

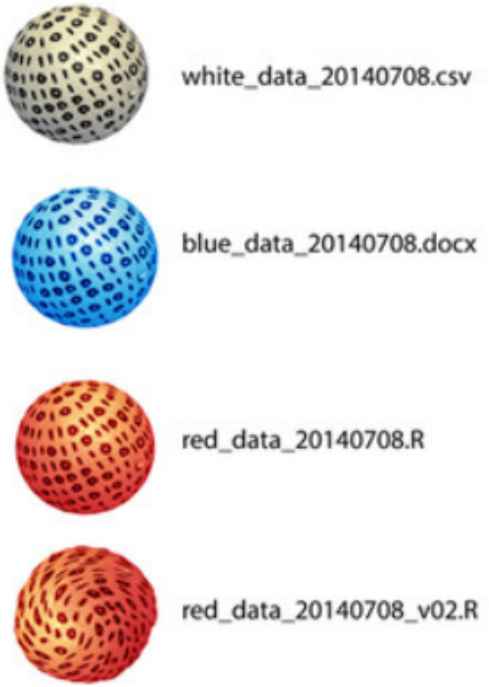
research
data
netherlands

Meaningful file names

Below are tips on meaningful and consistent file names. Read more in *'Naming files and folders'*. (2)

- Make sure to use consistent file names. When you use a date in the file name, choose a notation (for instance, YYYYMMDD of yymmdd).
- Do not use strange characters like ?\!@*%{[<> in the file name.
- Use traceable file names, such as Project_Instrument_locatie_YYYYMMDD.ext.
- Make sure to only use each file once in the folder structure. If you store a file in more than one place, several versions of the same file can unwillingly be created.
- See also version management.

It is good practice to note the file naming and its meaning in a readme.txt.



File naming and version management

Even if a researcher is well underway with his project consistent file naming is still an option by using a *bulk file rename utility*. (3) It is important, however, to check if this bulk renamer delivers on its promises.

The Open Data Pilot

van Panhuis et al. BMC Public Health 2014, 14:1144
http://www.biomedcentral.com/1471-2458/14/1144

BMC Public Health

RESEARCH ARTICLE Open Access

A systematic review of barriers to data sharing in public health

Willem G van Panhuis[1*], Proma Paul[1], Claudia Emerson[2], John Grefenstette[1], Richard Wilder[3], Abraham J Herbst[4,5], David Heymann[6] and Donald S Burke[1]

Abstract

Background: In the current information age, the use of data has become essential for decision making in public health at the local, national, and global level. Despite a global commitment to the use and sharing of public health data, this can be challenging in reality. No systematic framework or global operational guidelines have been created for data sharing in public health. Barriers at different levels have limited data sharing but have only been anecdotally discussed or in the context of specific case studies. Incomplete systematic evidence on the scope and variety of these barriers has limited opportunities to maximize the value and use of public health data for science and policy.

Methods: We conducted a systematic literature review of potential barriers to public health data sharing. Documents that described barriers to sharing of routinely collected public health data were eligible for inclusion and reviewed independently by a team of experts. We grouped identified barriers in a tax international dialogue on solutions.

Results: Twenty potential barriers were identified and classified in six categories: technical political, legal and ethical. The first three categories are deeply rooted in well-known challeng systems for which structural solutions have yet to be found; the last three have solutions tha dialogue aimed at generating consensus on policies and instruments for data sharing.

Table 1 Evidence for barriers to sharing of routinely collected public health data

Category	Barrier	Peer-reviewed		Non peer-reviewed
		Empirical data	Non-empirical*	
Technical	1. Data not collected	[6,21,24,31]	[2,4,7,18,22,14,26-28,30]	[3,23,25]
	2. Data not preserved		[33]	[3,32,34,35]
	3. Data not found		[45]	[3,34]
	4. Language barrier			[36]
	5. Restrictive data format		[40]	[3,34,36-39,41]
	6. Technical solutions not available		[42]	[37]
	7. Lack of metadata and standards	[21,24,43]	[40,44,45]	[1,35-37,39,41,46]
Motivational	8. No incentives		[27,45,49]	[35]
	9. Opportunity cost	[51,52]	[13,33,50,53]	[35]
	10. Possible criticism		[33]	[32]
	11. Disagreement on data use	[21]	[49]	
Economic	12. Possible economic damage		[7,26,27,30]	[55]
	13. Lack of resources	[56,21]	[13,27,28,30,42,53,57]	[3,23,34-36,39,37]
Political	14. Lack of trust	[19,59,60]	[33,61]	[34-37]
	15. Restrictive policies		[30]	
	16. Lack of guidelines		[45,62,65]	[37,41,63,64]
Legal	17. Ownership and copyright		[62,65,66,69]	[37,63,64,67]
	18. Protection of privacy	[12,19,59,73,75]	[44,57,62,66,72,74]	[36,37,64,67,68,70,71]
Ethical	19. Lack of proportionality			[76]
	20. Lack of reciprocity	[51,52]	[50,77,78]	
Number of unique documents (% of total)		14 (21.5%)	30 (46.2%)	21 (32.3%)

*No or little original data presented.

OpenAIRE

PARTICIPATE SEARCH MONITOR

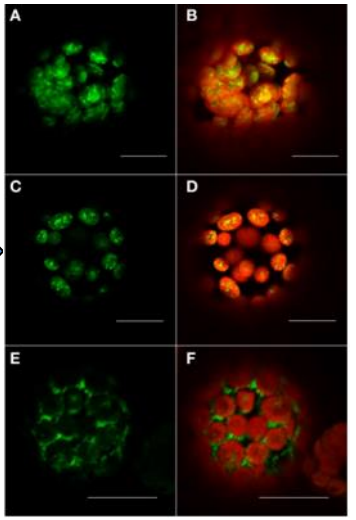
What is the Open Research Data Pilot?

Updated on 15 November 2016

"Don't panic - **you are not expected to share sensitive data or breach any IPR agreements** with industrial partners. You **do not need to deposit all the data** you generate during the project either – only that which underpins published research findings and/or has longer-term value. In addition to supporting your research's integrity, openness has many other benefits. **Improved visibility** means your research will reach more people and have a greater impact – for science, society and your own career. "

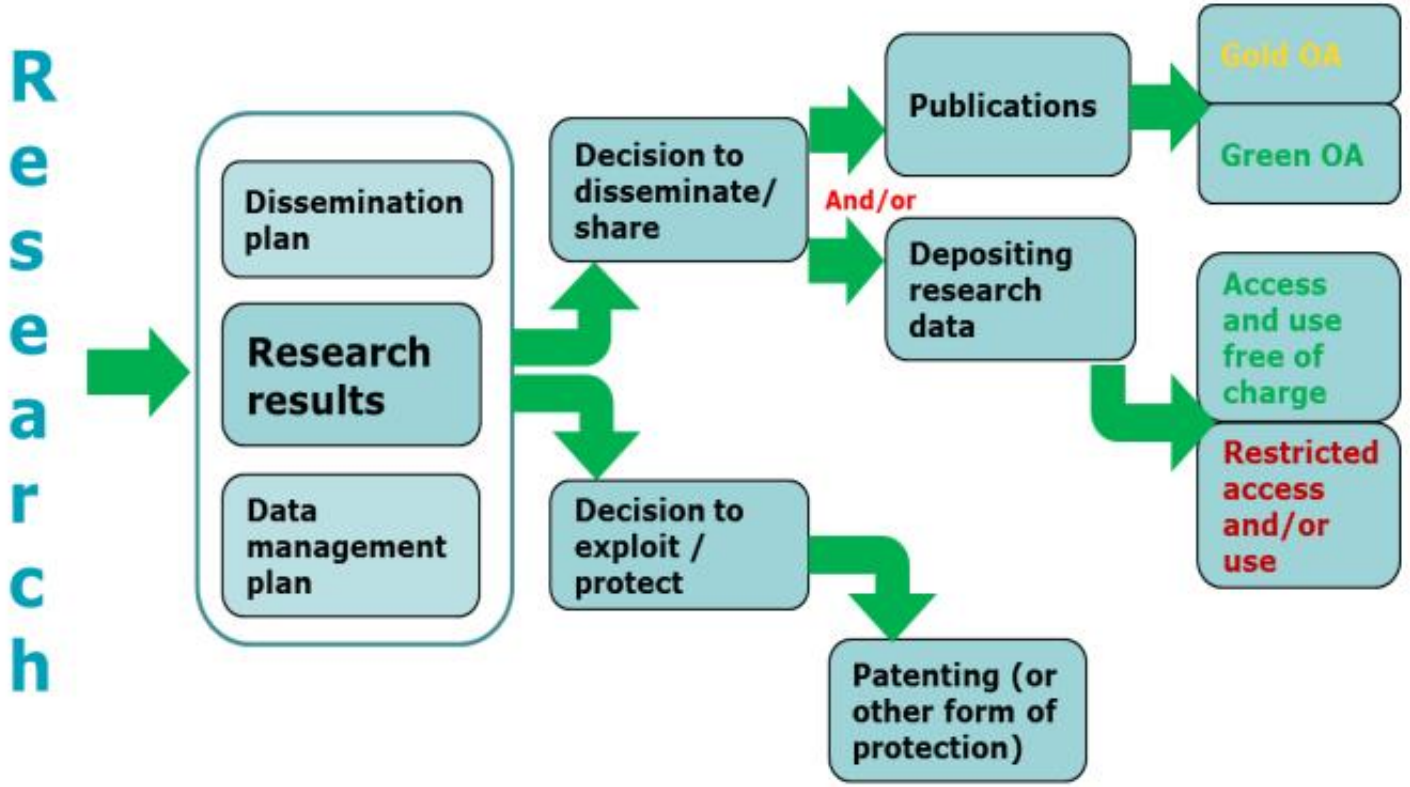
	XP_1	XP_2	…	XP_n
Gene 1	1.1	1.2	23.1	22.1
Gene 2	1.5	5.3	12.1	5.6
Gene 3	9.1	4.2	9.2	4.3
..				
Gene m	8.1	4.3	1.3	1.5

ML, combine



Machine Learning, data mining

https://www.openaire.eu/opendatapilot



Graph: Open access to scientific publication and research data in the wider context of dissemination and exploitation

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Opting OUT

- participation is incompatible with the obligation to protect results that can
- reasonably be expected to be commercially or industrially exploited
- participation is incompatible with the need for confidentiality in connection with security issues
- participation is incompatible with rules on protecting personal data
- participation would mean that the project's main aim might not be achieved
- the project will not generate / collect any research data or
- there are other legitimate reasons (you can enter these in a free-text box at the proposal stage).

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Thank you

Questions?