Tools for data management planning

Sarah Jones

Digital Curation Centre,
University of Glasgow

Rob Hooft

DTL
ELIXIR NL

Webinar series

26 July 2018

EMBL-EBI

DMPonline / DMPRoadmap

https://dmponline.dcc.ac.uk

Sarah Jones
Digital Curation Centre, Glasgow
sarah.jones@glasgow.ac.uk
Twitter: @sjDCC @DMPonline

EBI webinar series, 26 July 2018

What is a DMP and why do one?

A short plan explaining what data will be created and how it will be managed and shared

- Make informed decisions to anticipate and avoid problems
- Avoid duplication, data loss and security breaches
- Develop procedures early on for consistency
- Ensure data are accurate, complete, reliable and secure
- Save time and effort to make their lives easier!

DMPs should evolve as you conduct research. Plans change!

How does DMPonline help?

Online tool to help researchers develop Data Management Plans, tailored to their context

Requirements from funders, institutions and others

Guidance and examples from your funder, uni, research community and others

What is DMPRoadmap?

An open source platform for all things DMP

The codebase used to deliver DMPonline, the DMPTool, DMPOPIDoR, DMPTuuli, DMPAssistant and many other services

Managed by DCC and UC3 with community input:
https://github.com/DMPRoadmap

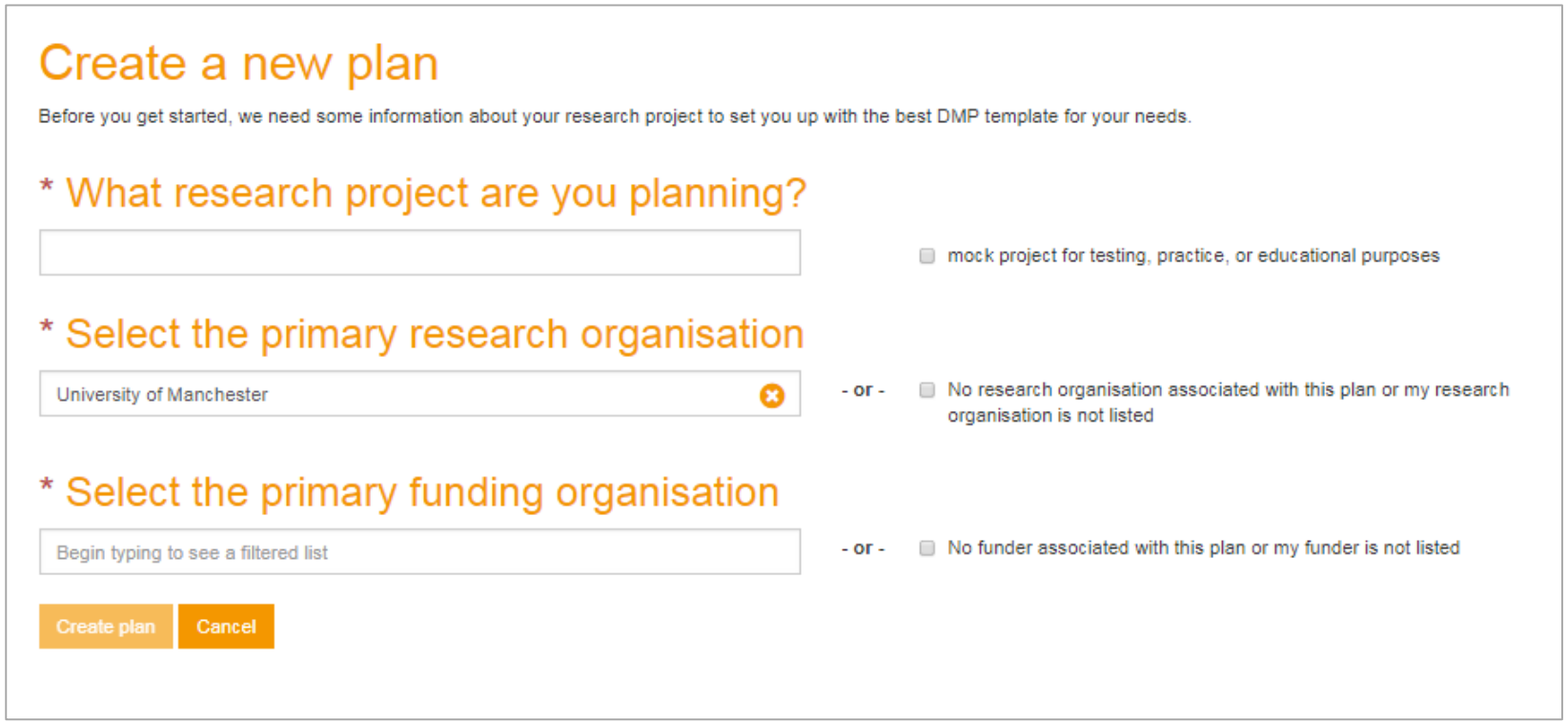
DMP guidelines and content

Theme	DCC & UC3 Guidance
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 (.tif, .tfw). • 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?

https://github.com/DMPRoadmap/roadmap/wiki/Themes

Create plan wizard

- Simplified set of questions to get started
- Easier to point to institutional guidance
- Accessible dropdown boxes



Configurable guidance

Centre for the Observation and Modelling of Earthquakes and Tectonics (COMET)

Project Details | Plan overview | Outline DMP | Full DMP | Share | Download

* Project title

Centre for the Observation and Modelling of Earthquakes and Tectonics (

mock project for testing, practice, or educational purposes

Funder

Grant number

Project abstract

The Centre for the Observation and Modelling of Earthquakes and Tectonics (COMET) is a collaborative project funded by NERC, involving scientists from the University of Oxford, University of Cambridge, and University College London.

ID

Principal Investigator

Name

Sarah Jones

Plan Guidance Configuration

To help you write your plan, DMPonline can show you guidance from a variety of organisations.

Select up to 6 organisations to see their guidance.

- Digital Curation Centre
- University of Edinburgh
  - Edinburgh
  - Roslin Institute
- University of Manchester

Find guidance from additional organisations below

See the full list

Submit

- Choose different organisations e.g. research partners
- Turn guidance on/off as you write plan

Write plan with examples and guidance

Project Details | Plan overview | Write Plan | Share | Download

expand all | collapse all

0/13 answered

Data Collection (0 / 2)

What data will you collect or create?

Save

St Andrews example answer

Existing data

We have examined datasets produced by authors such as Knopfler (1949) and Gabriel (1975) which tested a similar hypothesis but with a different sample. This data is freely available at (LINK) and there are no licensing or access issues. The size of the data file is approximately 50MB and there is no charge for accessing or reusing the data. The original data will be acknowledged in the research paper, together with its DOI.

New data

In this study we will investigate the effects of therapeutic agents on cells from children with acute lymphoblastic leukaemia. We will also characterise the cells that may be responsible for initiating and maintaining this malignancy.

Quantitative data derived from our experimental approaches and statistical analyses of these data will be managed along with Genotypic data from results of our microarray studies. There will also be administrative data on linked anonymised tissue samples and storage locations of the samples.

Data will be initially be collected in a variety of file formats mainly Microsoft Excel for databases, MS Word and equipment specific software such as FlowJo for flow cytometry data. Graph Pad prism will be used to generate graphs and conduct statistical analyses. Presentations will usually be compiled using MS Powerpoint. These files will also be stored in Open Document Format or as CSV files, for spreadsheets, to enable sharing. Digital images will be saved as TIFF files for this purpose.

Guidance | Comments

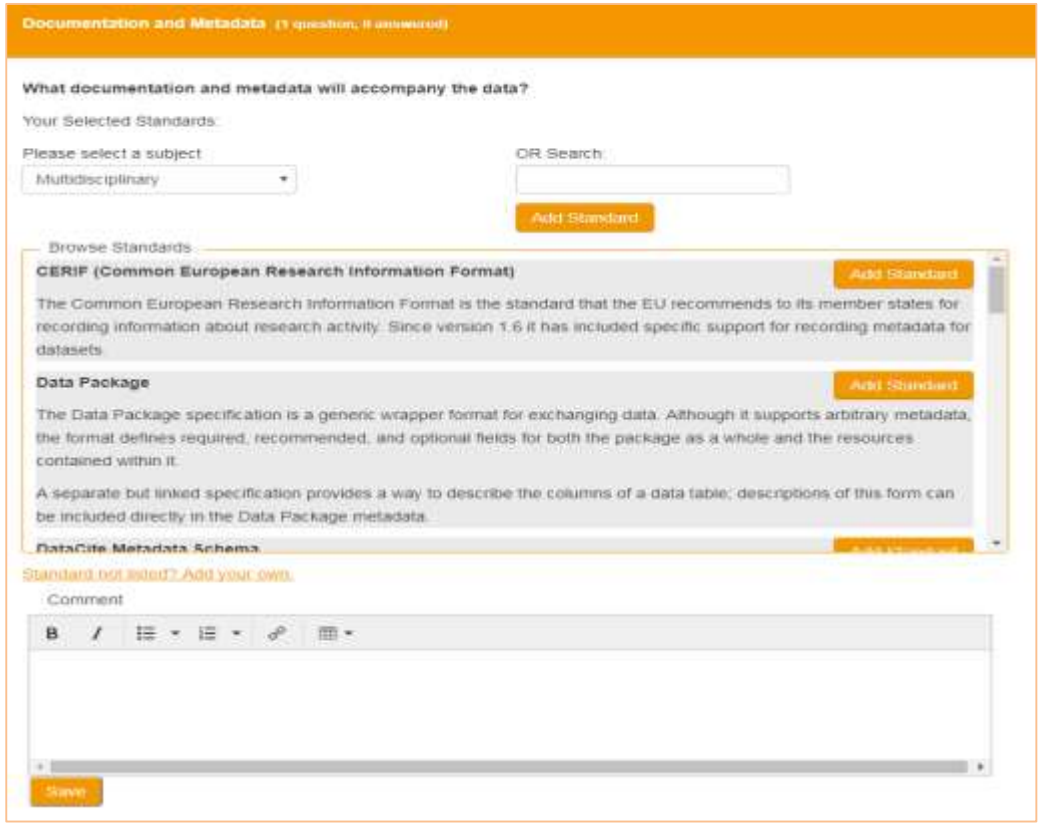
EPSRC | St Andrews | DCC

expand all | collapse all

Data volume

The University policy is to provide each PI with 0.5TB of centrally managed storage for research data. Click here to request your allocation.

Integrate catalogues to guide answers



Have integrated RDA metadata standards directory

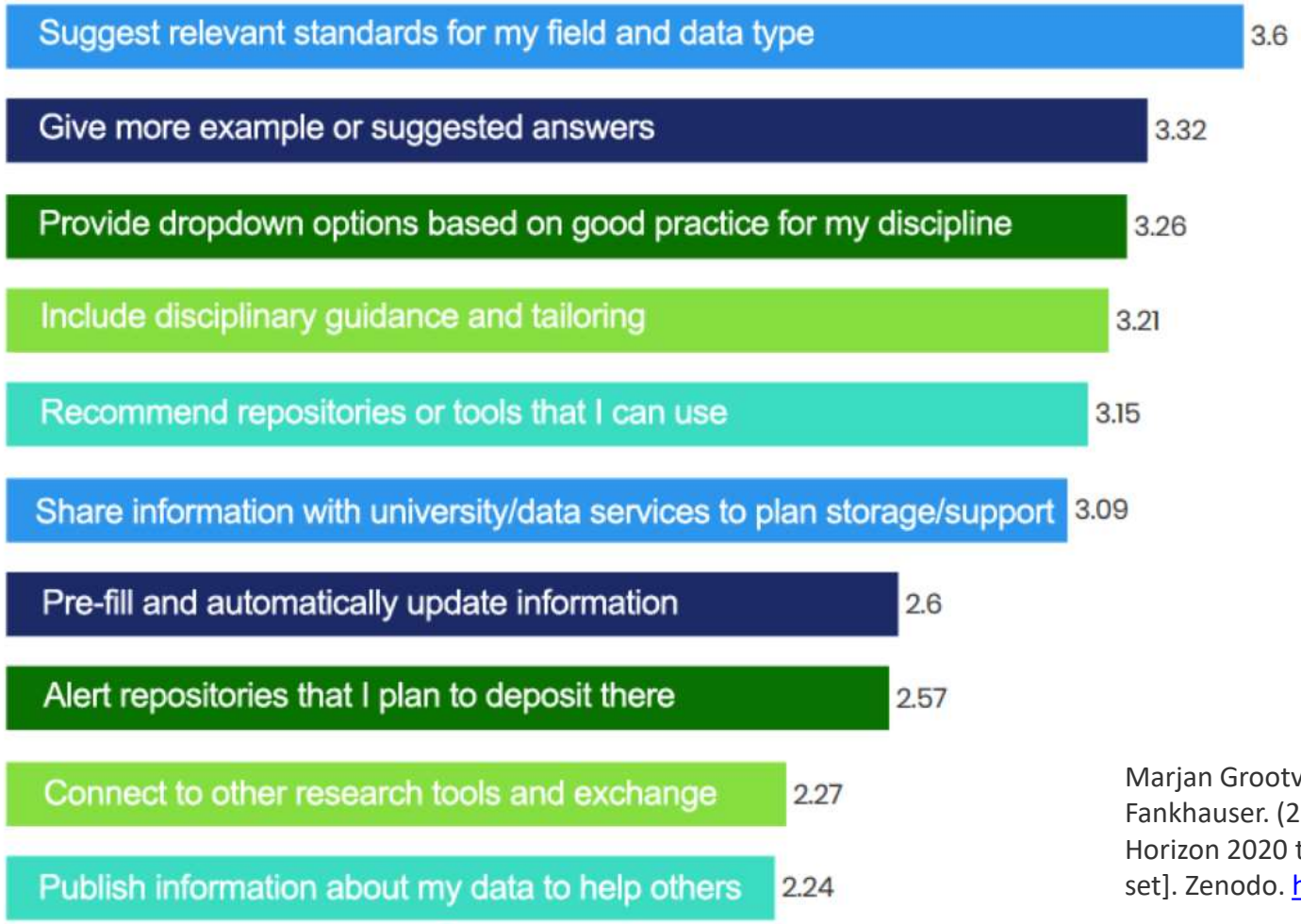
Possibilities for recommending repositories, pulling in grant numbers from funder databases etc

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

OpenAIRE

FAIRsharing.org
standards, databases, policies

What H2020 DMP writers prioritised



Marjan Grootveld, Ellen Leenarts, Sarah Jones, Emilie Hermans, & Eliane Fankhauser. (2018). OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans (Version 1.0.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1120245

Sharing DMPs

Manage collaborators

Invite specific people to read, edit, or administer your plan. Invitees will receive an email notification that they have access to this plan.

Email address	Permissions	
jimmy.angelakos@ed.ac.uk	Editor ▼	Remove
s.jones@arts.gla.ac.uk	Owner	

Set plan visibility

Public or organisational visibility is intended for finished plans. You must answer at least 50% of the questions to enable these options. Note: test plans are set to private visibility by default.

- (•) Private: visible to me, specified collaborators and administrators at my organisation
- ( ) Organisation: anyone at my organisation can view
- ( ) Public: anyone can view

Request expert feedback

Click below to give data management staff at your organisation access to read and comment on your plan.

You can continue to edit and download the plan in the interim.

Request feedback

Plans are private by default, but can be shared organisationally or made public.

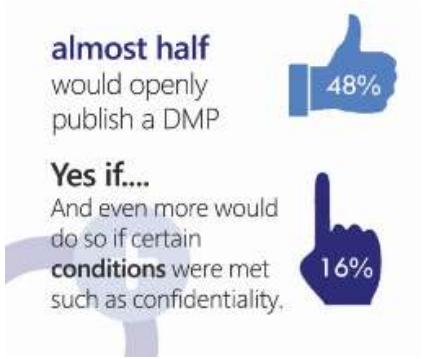
Plan publishing

Public DMPs

Public DMPs are plans created using the DMPonline service and shared publicly by their owners. They are not vetted for quality, completeness, or adherence to funder guidelines.

Project Title	Template	Organisation	Owner	Download
Students Behaviour towards Research	Population Research Committee Template	Other	Dimakatso Pilusa	PDF
Max-plus switching systems and long max-plus matrix products	EPSRC Data Management Plan	University of Birmingham	Arthur Kennedy-Cochran-Patrick	PDF
RELIEF: Reducing Environmental impact of the Leather-tanning industry with Electron beam Facilities	STFC Template	Lancaster University	Robert Apsimon	PDF
LOOPER: Learning Loops in the Public Realm	ESRC Template	University of Manchester	James Evans	PDF
Mitigation of Waxing in Oil and Gas Transport Pipelines	DCC Template	Cranfield University	Israel Adeferni	PDF
Vergleich der Kriminalität von London und New South Wales von 2001-2012	DMP University of Vienna Deutsch V2	Other	Alexander Gallauner	PDF
Structured Risk-based Peer Evaluation System Study (StRIPES Study)	University of Bristol General Template	University of Bristol	Ian Bennett-Britton	PDF
Brains on Board	EPSRC Data Management Plan	University of Sheffield	James Marshall	PDF

https://dmponline.dcc.ac.uk/public_plans



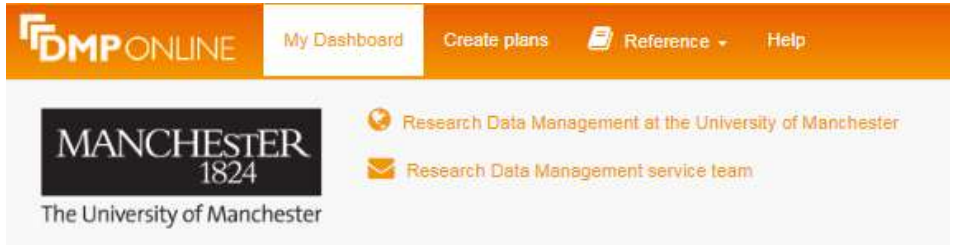
https://zenodo.org/record/1120245

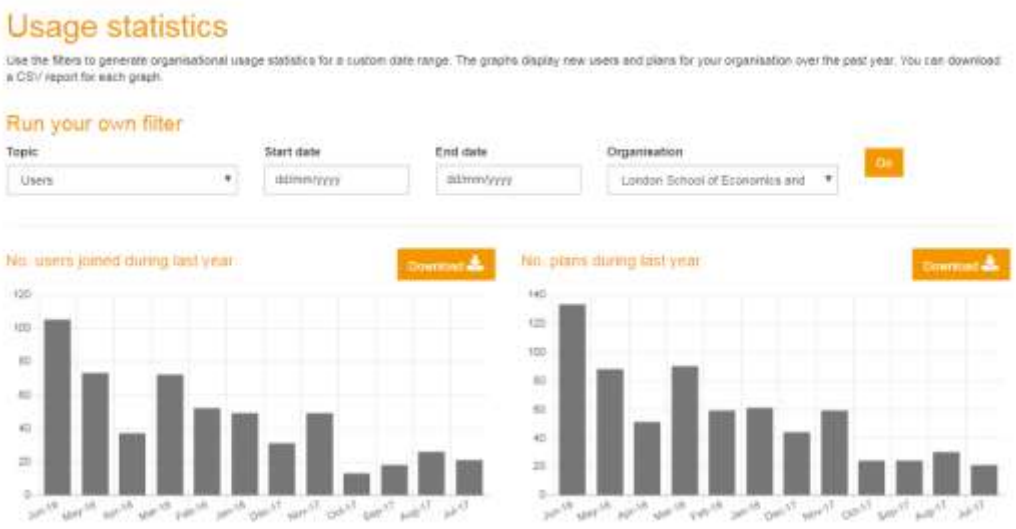
But also via:

- Repository deposit
- Journal publishing e.g. RIOjournal - https://riojournal.com
- LIBER catalogue - https://libereurope.eu/dmpcatalogue
- …..

Improved administrator controls

- Easier customisation workflow
- Plan review / feedback controls
- Access to all org plans
- Usage dashboard
- Branding
- …





Vision for future: machine-actionable DMPs

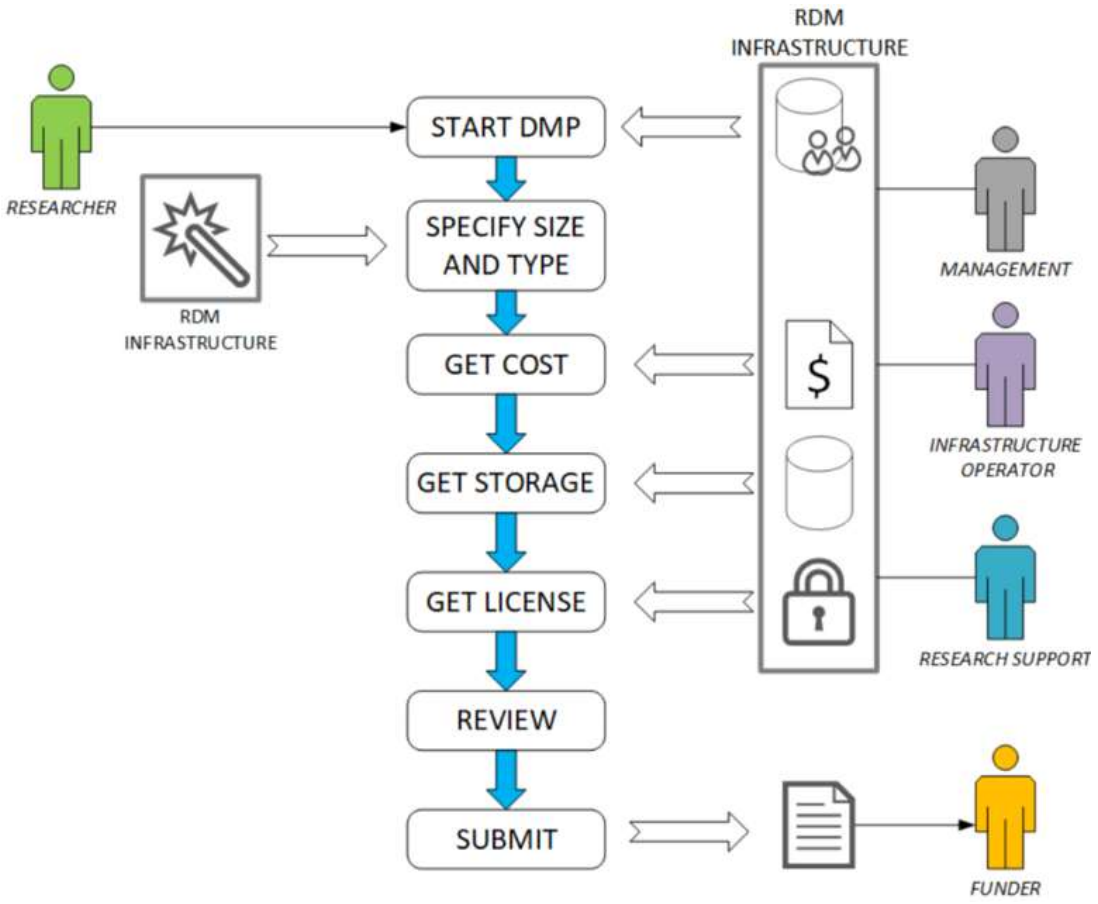
Transform static documents in active, machine-actionable DMPs that exchange data across systems to enable:

- Researchers to manage, share and discover data more easily
- Infrastructure providers to plan their resources
- Institutions to provide effective data services
- Funders to monitor data-related activities

RDA common standards for DMPs

Aim to develop a common data model to enable tools and systems involved in processing research data to read and write information to/from DMPs

https://www.rd-alliance.org/groups/dmp-common-standards-wg



Source: Tomasz Miksa, Peter Neish, Paul Walk, & Andreas Rauber. (2018). Defining requirements for machine-actionable Data Management Plans (Version preprint). Zenodo. http://doi.org/10.5281/zenodo.1266211

Where to find out more?

DMPRoadmap

https://github.com/DMPRoadmap

Follow us on twitter:
@DMPonline and #DMPonline
@digitalcuration and #ukdcc
https://dmponline.dcc.ac.uk

DTL

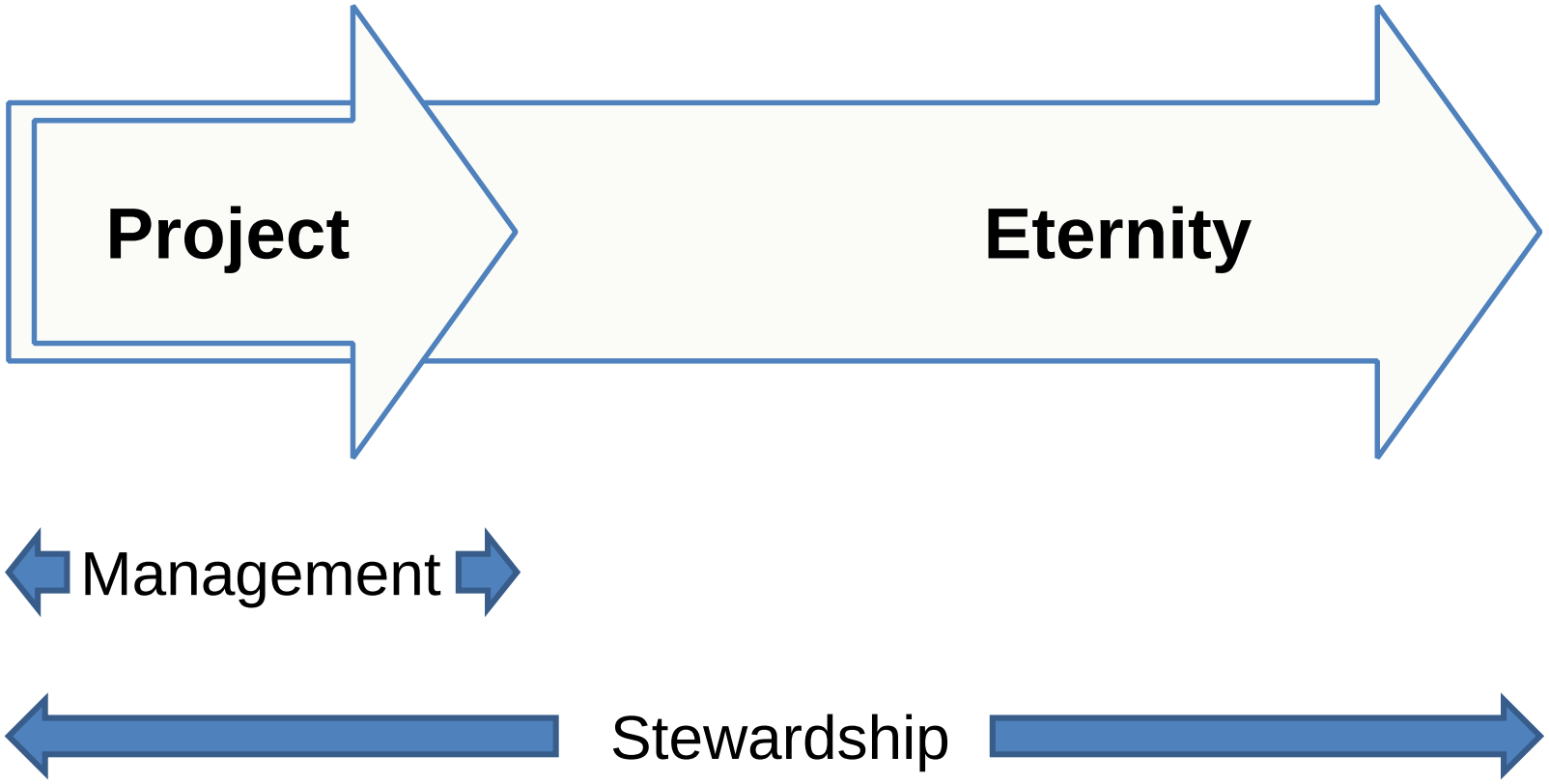
DUTCH TECHCENTRE FOR LIFE SCIENCES

elixir NETHERLANDS

DATA STEWARDSHIP WIZARD

ROB HOOFT

EBI Webinar, 2018-07-26


Project
Eternity
Management
Stewardship

Reusable

FAIR

Findable:

F1. (meta)data are assigned a globally unique and persistent identifier;
F2. data are described with rich metadata;
F3. metadata clearly and explicitly include the identifier of the data it describes;
F4. (meta)data are registered or indexed in a searchable resource;

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;

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;

Reusable:

R1. meta(data) are richly described with 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 detailed provenance;
R1.3. (meta)data meet domain-relevant community standards;

DMP for a ZonMw Project

expand all | collapse all

0/29 answered

1. General information (0 / 11) +

2. Legislation and regulations (0 / 2) +

3. Findable (0 / 4) +

4. Accessible (0 / 3) +

5. Interoperable (0 / 4) +

6. Reusable (0 / 0) +

7. Sustainable data storage (0 / 5) +

DTL

* What research project are you planning?

* Select the primary research organisation

Other

* Select the primary funding organisation

Begin typing to see a filtered list

Create plan Cancel

DTL

2. FAIR data (0 / 4)

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

Guidance | Comments

EC | DCC

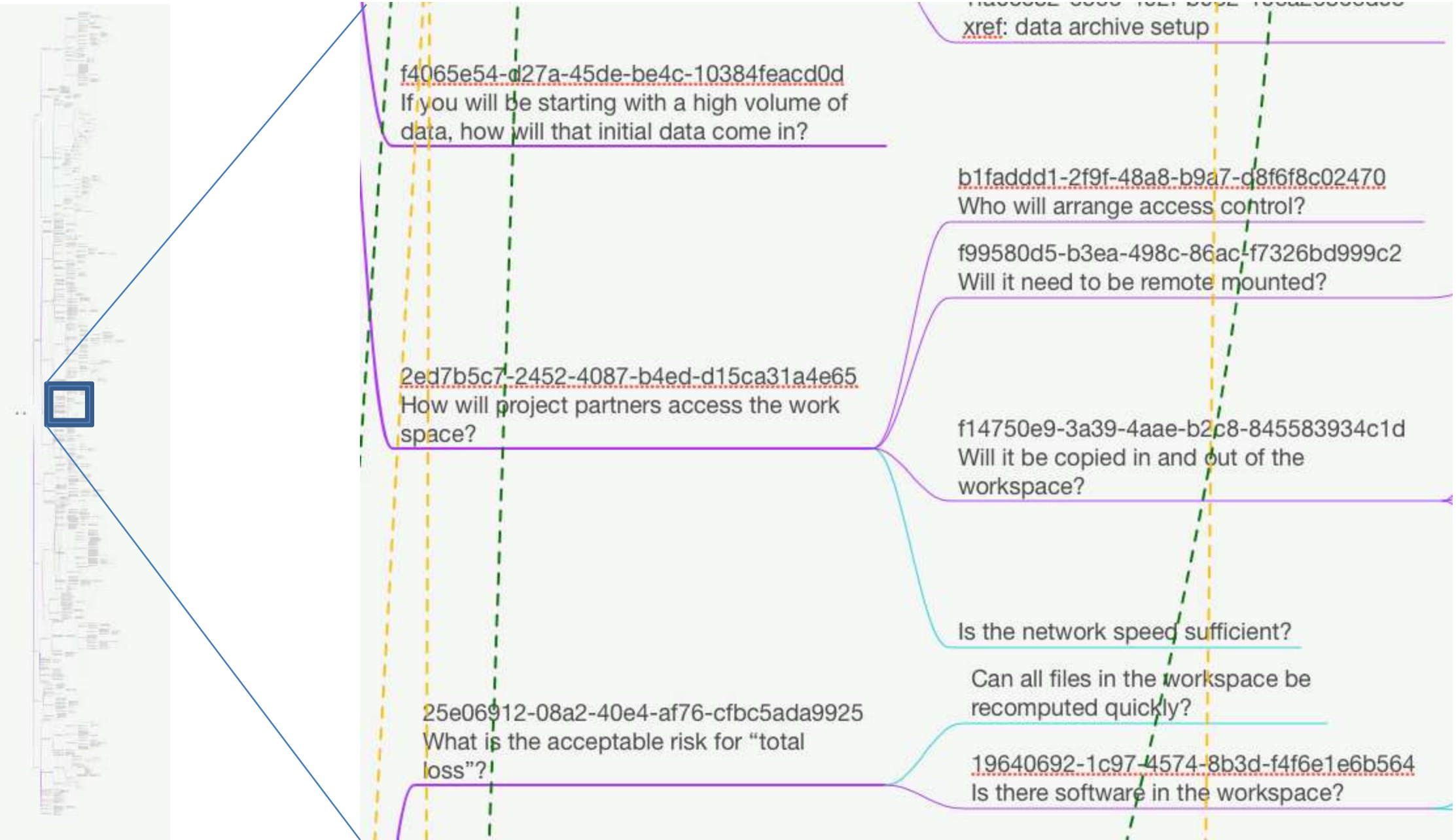
Question Specific Guidance

The Research Data Alliance provides a Metadata Standards Directory that can be searched for discipline-specific standards and associated tools.

Irritant
Painful
Dangerous

In preparing for battle I have always found that plans are useless, but planning is indispensable.

Dwight D. Eisenhower


xref: data archive setup
f4065e54-d27a-45de-be4c-10384feacd0d
If you will be starting with a high volume of data, how will that initial data come in?
b1faddd1-2f9f-48a8-b9a7-d8f6f8c02470
Who will arrange access control?
f99580d5-b3ea-498c-86ac-f7326bd999c2
Will it need to be remote mounted?
2ed7b5c7-2452-4087-b4ed-d15ca31a4e65
How will project partners access the work space?
f14750e9-3a39-4aae-b2c8-845583934c1d
Will it be copied in and out of the workspace?
Is the network speed sufficient?
Can all files in the workspace be recomputed quickly?
25e06912-08a2-40e4-af76-cfbc5ada9925
What is the acceptable risk for "total loss"?
19640692-1c97-4574-8b3d-f4f6e1e6b564
Is there software in the workspace?

Design of experiment

Before you decide to embark on any new study, it is nowadays good practice to consider all options to keep the data generation part of your study as limited as possible. It is not because we can generate massive amounts of data that we always need to do so. Creating data with public money is bringing with it the responsibility to treat those data well and (if potentially useful) make them available for re-use by others.

Is there any pre-existing data?

Are there any data sets available in the world that are relevant to your planned research?

- ○ No
- ◉ Yes

Will you be using any pre-existing data (including other people's data)?

Will you be referring to any earlier measured data, reference data, or data that should be mined from existing literature? Your own data as well as data from others?

- ○ No
- ◉ Yes

What reference data will you use?

Does this data format enable sharing and long term archiving?

Complicated (binary) file formats tend to change over time, and software may not stay compatible with older versions. Also, some formats hamper long term usability by making use of patents or being hampered by restrictive licensing

- (●) No
- ( ) Yes

Will you convert to a file format more suitable for archiving later?

- ( ) No
- (●) Yes

You may need to reserve time and budget for this

What data formats/types will you be using?

Have you identified types of data that you will use that are used by others too? Some types of data (e.g. genetic variants in the life sciences) are used by many different projects. For such data, often common standards exist that help to make these data reusable. Are you using such common data formats?

Item

Is this a standard data format used by others too?

- No
- Yes

Does this data format enable sharing and long term archiving?

Complicated (binary) file formats tend to change over time, and software may not stay compatible with older versions. Also, some formats hamper long term usability by making use of patents or being hampered by restrictive licensing

- No
- Yes

Data Stewardship Wizard

- Organization
- Users
- KM Editor
- KM Packages
- DS Planner

Rob Hooft

Logout

Knowledge Model Editor

Create

Name	Knowledge Model ID	Parent Package ID	Actions
ELIXIR mod	sdf	elixir:root:1.0.0	
FAIR Metrics	fair-metrics	-	
FIT CTU Customization	root	elixir:root:1.0.0	
Ivan	root2	elixir:root2:1.0.0	
KM with FAIR Metrics questions	km-metrics	elixir:root:1.0.0	
SdRTest	SdRTestID	-	Publish
Test	test	-	Publish
Test 001	test001	-	
Test 002 outdated	test002	elixir:test001:1.0.1	Upgrade
anothermodel	this	-	Publish

https://dsw.fairdata.solutions/

Thanks to:

And all the real experts
(most of whom are in ELIXIR)