# Data Management Planning

*Pinar Alper*

*Training on Research Data Management*
*25-26 June 2019*
*Luxembourg Learning Centre*

# Why DMPs?

## - Funder requirement

- “Ensure data management via DMPs.”
- “Researchers are accountable for how data is treated during and after the project.”

- timely release of data - once patents are filed or on (acceptance for) publication
- (open) data sharing - minimal or no restrictions if possible
- preservation of data - typically 5-10+ years

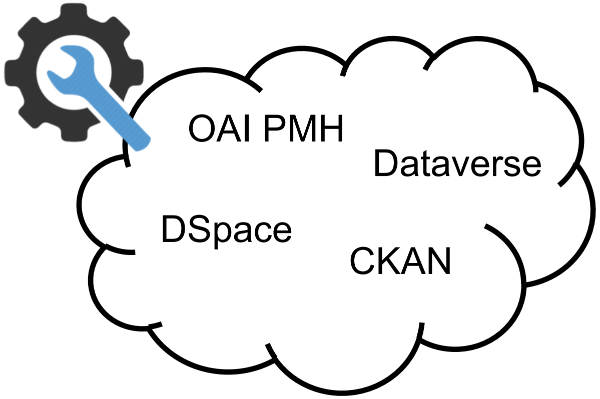
*Introduction to Data Management.* Joy Davidson. UK Digital Curation centre. 2015

*Turning FAIR into reality.* Final Report and Action Plan from the EC Expert Group on FAIR Data. November 2018

*Funding research data management and related infrastructures.* Knowledge Exchange and Science Europe briefing paper. May 2016

# Making data a first class-citizen in research

— decades of efforts, varying levels of “FAIR” ness

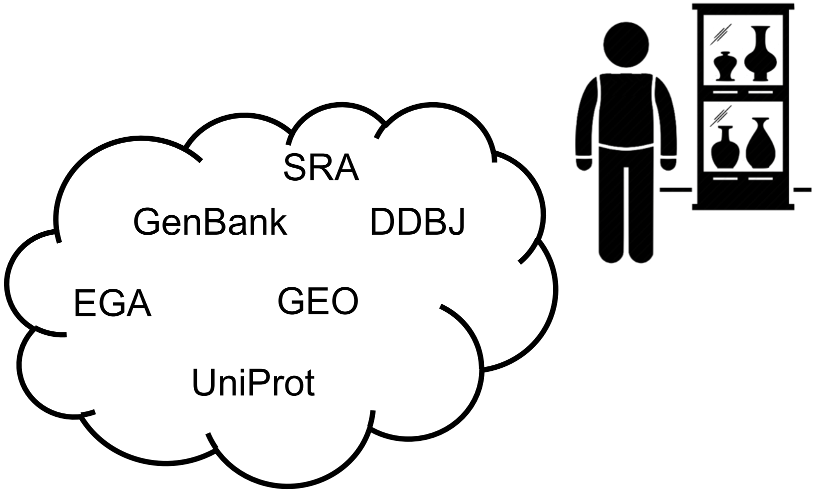


OAI PMH

Dataverse

DSpace

CKAN



SRA

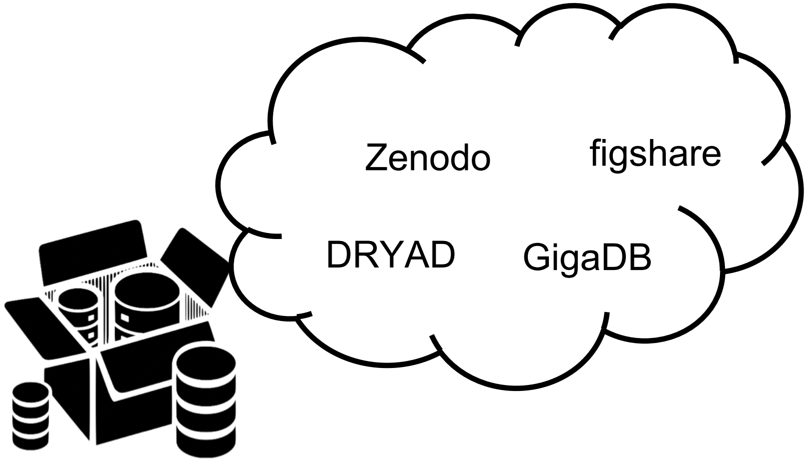
GenBank

DDBJ

EGA

GEO

UniProt



Zenodo

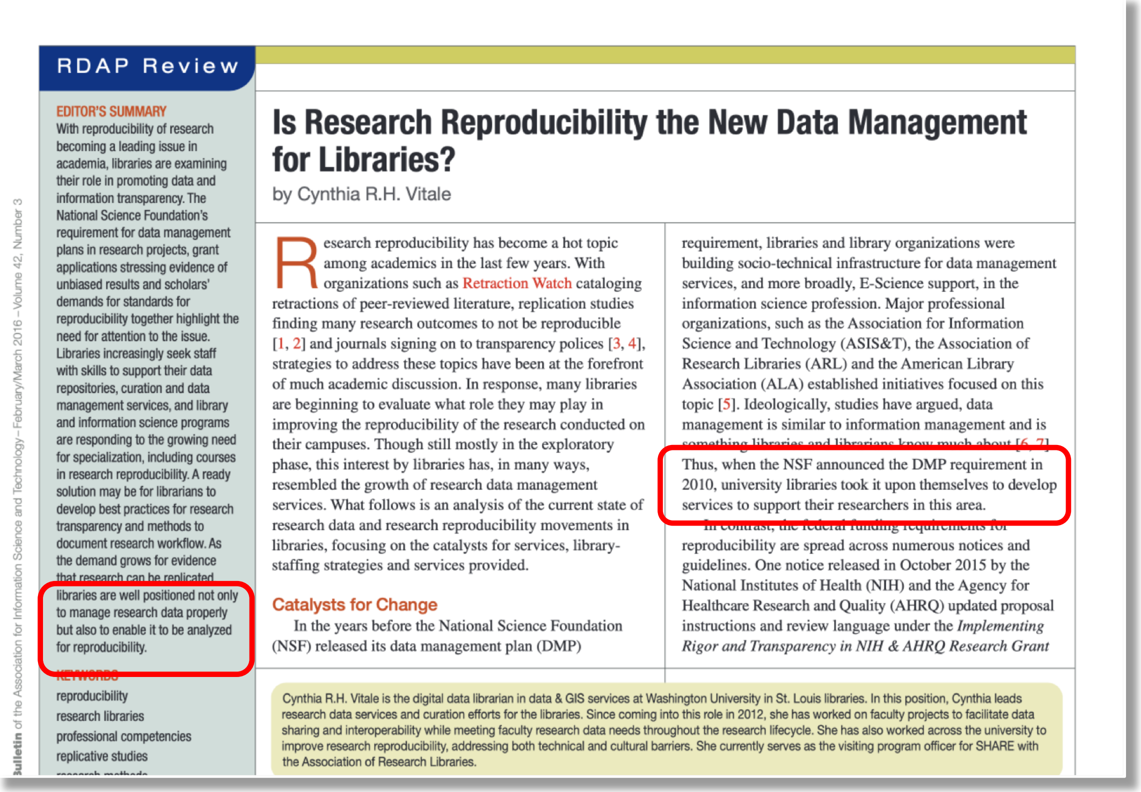
figshare

DRYAD

GigaDB

# Making data a first class-citizen in research

- A change in research culture and funding
- DMP is one such intervention
- Future interventions; reporting standards, code, computational reproducibility?



Bulletin of the Association for Information Science and Technology – February/March 2016 – Volume 42, Number 3

RDAP Review

EDITOR'S SUMMARY

With reproducibility of research becoming a leading issue in academia, libraries are examining their role in promoting data and information transparency. The National Science Foundation's requirement for data management plans in research projects, grant applications stressing evidence of unbiased results and scholars' demands for standards for reproducibility together highlight the need for attention to the issue. Libraries increasingly seek staff with skills to support their data repositories, curation and data management services, and library and information science programs are responding to the growing need for specialization, including courses in research reproducibility. A ready solution may be for librarians to develop best practices for research transparency and methods to document research workflow. As the demand grows for evidence that research can be replicated, libraries are well positioned not only to manage research data properly but also to enable it to be analyzed for reproducibility.

KEYWORDS

reproducibility
research libraries
professional competencies
replicative studies
research methods

## Is Research Reproducibility the New Data Management for Libraries?

by Cynthia R.H. Vitale

Research reproducibility has become a hot topic among academics in the last few years. With organizations such as Retraction Watch cataloging retractions of peer-reviewed literature, replication studies finding many research outcomes to not be reproducible [1, 2] and journals signing on to transparency polices [3, 4], strategies to address these topics have been at the forefront of much academic discussion. In response, many libraries are beginning to evaluate what role they may play in improving the reproducibility of the research conducted on their campuses. Though still mostly in the exploratory phase, this interest by libraries has, in many ways, resembled the growth of research data management services. What follows is an analysis of the current state of research data and research reproducibility movements in libraries, focusing on the catalysts for services, library-staffing strategies and services provided.

### Catalysts for Change

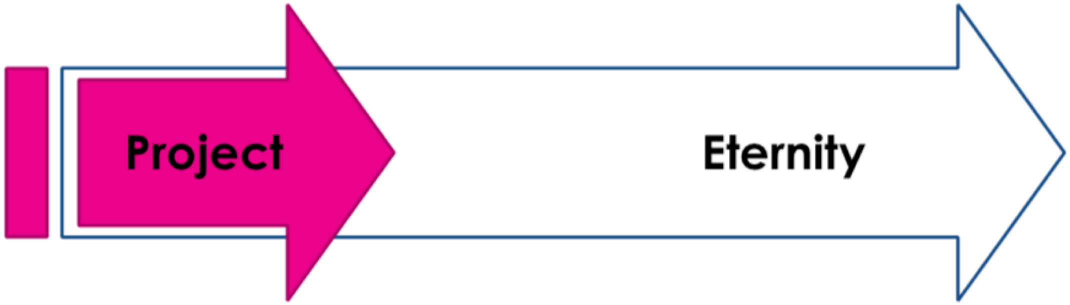
In the years before the National Science Foundation (NSF) released its data management plan (DMP) requirement, libraries and library organizations were building socio-technical infrastructure for data management services, and more broadly, E-Science support, in the information science profession. Major professional organizations, such as the Association for Information Science and Technology (ASIS&T), the Association of Research Libraries (ARL) and the American Library Association (ALA) established initiatives focused on this topic [5]. Ideologically, studies have argued, data management is similar to information management and is something libraries and librarians know much about [6, 7]. Thus, when the NSF announced the DMP requirement in 2010, university libraries took it upon themselves to develop services to support their researchers in this area.

In contrast, the federal funding requirements for reproducibility are spread across numerous notices and guidelines. One notice released in October 2015 by the National Institutes of Health (NIH) and the Agency for Healthcare Research and Quality (AHRQ) updated proposal instructions and review language under the *Implementing Rigor and Transparency in NIH & AHRQ Research Grant*

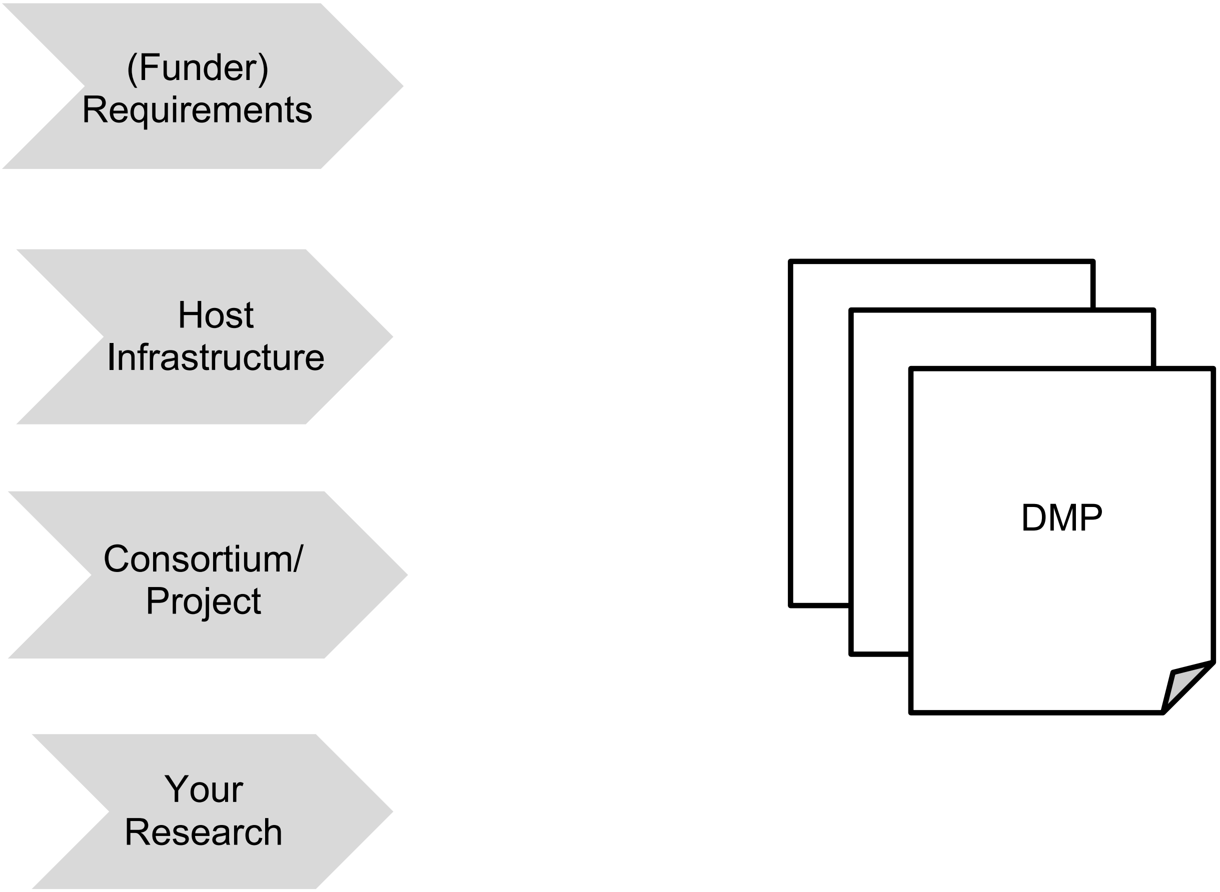
Cynthia R.H. Vitale is the digital data librarian in data & GIS services at Washington University in St. Louis libraries. In this position, Cynthia leads research data services and curation efforts for the libraries. Since coming into this role in 2012, she has worked on faculty projects to facilitate data sharing and interoperability while meeting faculty research data needs throughout the research lifecycle. She has also worked across the university to improve research reproducibility, addressing both technical and cultural barriers. She currently serves as the visiting program officer for SHARE with the Association of Research Libraries.

# What is a DMP?

- formal document
- awareness tool



# A DMP is shaped by



(Funder) Requirements

Host Infrastructure

Consortium/ Project

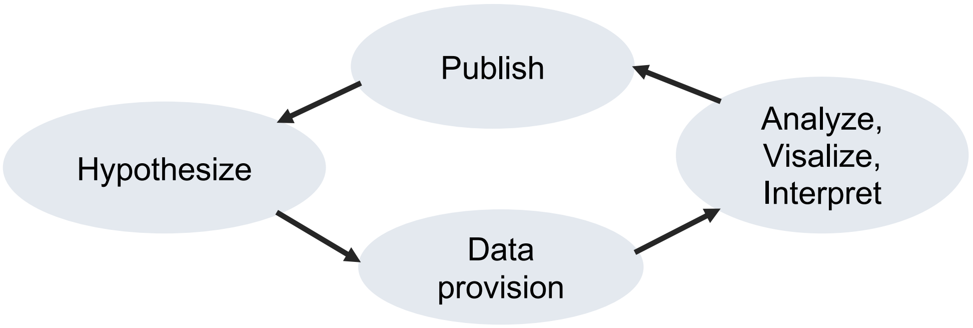
Your Research

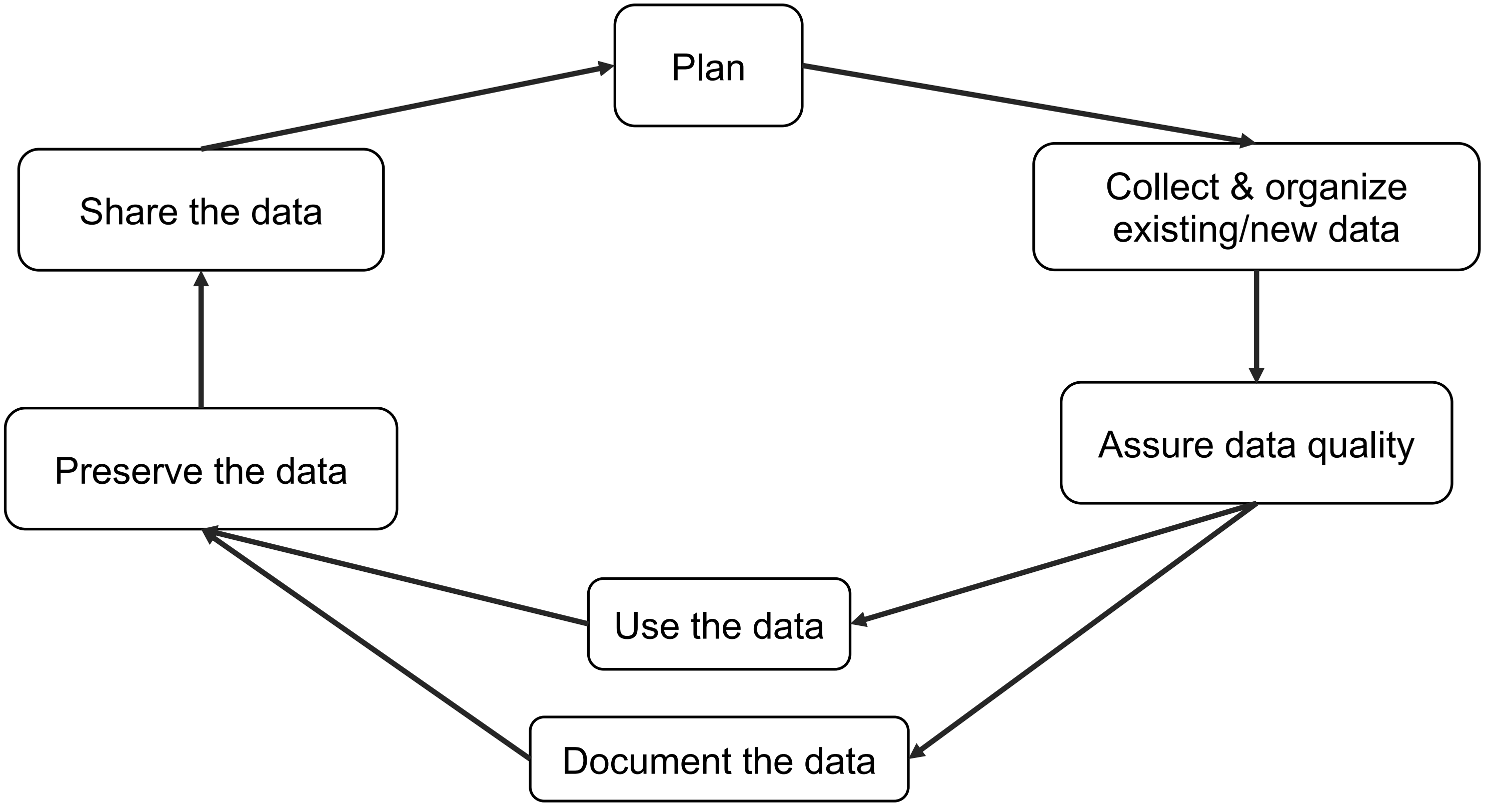
DMP

tip

An example goes a long way.

# The DMP worldview



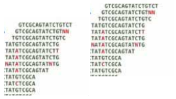


*Adapted from “Ten Simple Rules for Creating a Good Data Management Plan”. W Michener*

# Identification

- What constitutes " data "

Primary/Derived data



Research Record

Accompanying documents

tip

Be thorough when identifying data.

- Type, structure, format, size

tip

Identify data as early as possible, ideally during consortium setup.

Use of proprietary formats must be justified. Normally, a big no-no for preservation.

# Collection or generation

- Reference data
- Newly generated data
  - Capture instrument, software, method
  - Period of capture and update

tip

Highlight re-use.
Justify generation.

- Sources, Subjects
- Cohort/Study/Site description

tip

Identify data utility during and after project, potential re-use.

High value data, e.g. one-time events, costly collection, validation studies

https://instr.iastate.libguides.com/dmp/step6

# Storage and Preservation

- Platforms
  - Institutional, project-specific, other
- Transfer channels
- Backup policy
- Which data will be retained during and after the project? For how long?

Storing data on hard-drive/server/labbook is not preservation.

Estimate time and effort to prepare data for preservation e.g. export from DB

*tip*

Not all data may need to be retained, e.g. re-computable results.

Check whether institutional resources will suffice for your backup and recovery requirements.

# Organization

- Data Management
- Data Organization
  - File organization/naming
  - Version management
  - Identifiers; accession, doi
  - Searchability

LabCollector

Galaxy PROJECT

REDCap

Research Electronic Data Capture

PostgreSQL

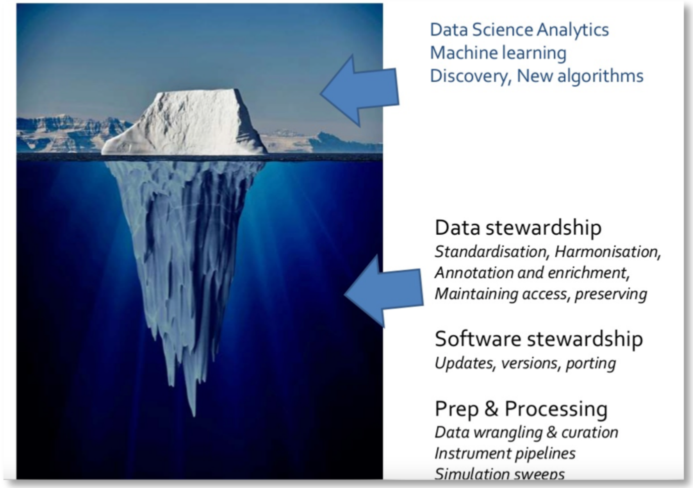
git

*tip*

Paid data management software also goes in to the DMP budget.

# Quality Assurance and Control

- Automated or manual QA/QC measures for validity and integrity
  - Tool/pipeline
  - Training
  - Standards
  - Instrument Calibration
  - Visualisation/Analytics platform



Data Science Analytics
Machine learning
Discovery, New algorithms

Data stewardship
*Standardisation, Harmonisation, Annotation and enrichment, Maintaining access, preserving*

Software stewardship
*Updates, versions, porting*

Prep & Processing
*Data wrangling & curation*
*Instrument pipelines*
*Simulation sweeps*

The New York Times

*For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights*

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

*tip*

Be very conservative in your assumptions of source data quality.

Building the FAIR Research Commons: A Data Driven Society of Scientists. C A Goble.

# Ethics , Legal Complaince

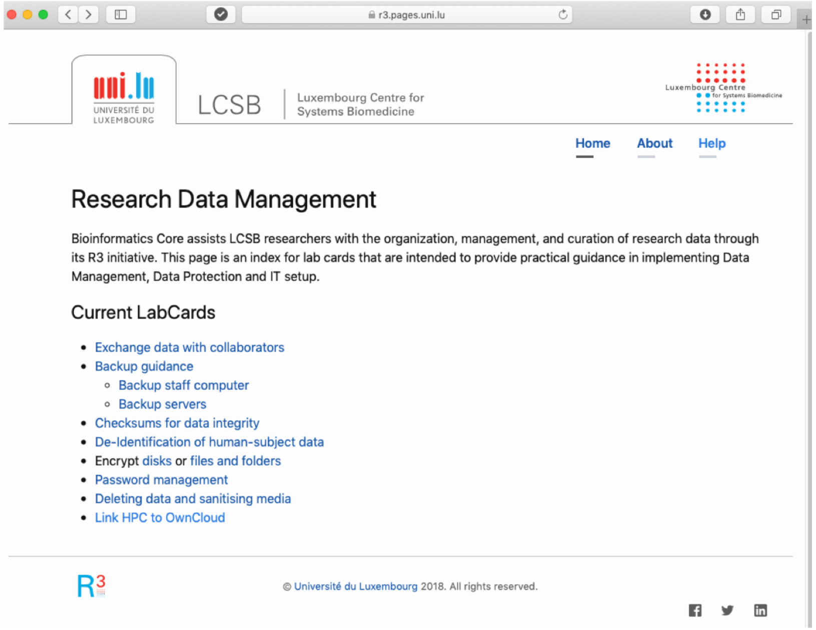
- What are the applicable regulations to your data?
- Is ethics/irb approval required?
- For Sensitive Human data, what is your Data Protection Concept?
  - Have you gained consent also for data sharing and preservation?
  - What type of agreements are needed during and after the project. – consortium agreement should cover sharing!
  - What are the protective measures for data and subject privacy?

*tip*

Seek support from Legal Office and Ethics Board. …Standard clauses..

# Security

- What are the security measures?
    - Encryption, Access Control, Password Management

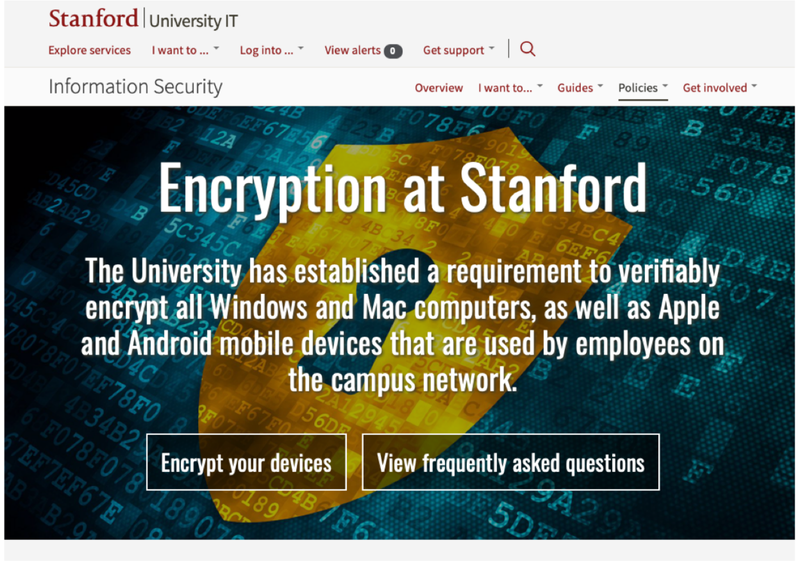

r3.pages.uni.lu
uni.lu
UNIVERSITÉ DU LUXEMBOURG
LCSB
Luxembourg Centre for Systems Biomedicine
Luxembourg Centre for Systems Biomedicine
Home
About
Help

## Research Data Management

Bioinformatics Core assists LCSB researchers with the organization, management, and curation of research data through its R3 initiative. This page is an index for lab cards that are intended to provide practical guidance in implementing Data Management, Data Protection and IT setup.

### Current LabCards

- Exchange data with collaborators
- Backup guidance
    - Backup staff computer
    - Backup servers
- Checksums for data integrity
- De-Identification of human-subject data
- Encrypt disks or files and folders
- Password management
- Deleting data and sanitising media
- Link HPC to OwnCloud



## Encryption at Stanford

The University has established a requirement to verifiably encrypt all Windows and Mac computers, as well as Apple and Android mobile devices that are used by employees on the campus network.

*tip*

Refer to standards, institutional policy/procedures.

e.g. ISO 27001, SAML SSO

# Documentation

- **Metadata**: Information enabling the read and interpretation of data
- What documentation and metadata will accompany the data
- Required for publicly shared data
- How will you capture/create it
  - Basics e.g. title, source, license
  - Standards e.g. Minimum information guidelines

*tip*

Estimate effort for documentation before commiting to it in the DMP.

# Ownership, Intellectual rights, Access

- Will data stay in the public domain?
- Or are there IP restrictions for ownership (consortium owns the data), copyright commercialisation.
- What will be the license for the data?
- May be added during DMP update

*tip*

https://eudat.eu/services/userdoc/b2share-usage
http://www.dcc.ac.uk/resources/how-guides/license-research-data

# Dissemination

- If, when and how will you share data?

  Data Paper, Repository, Database

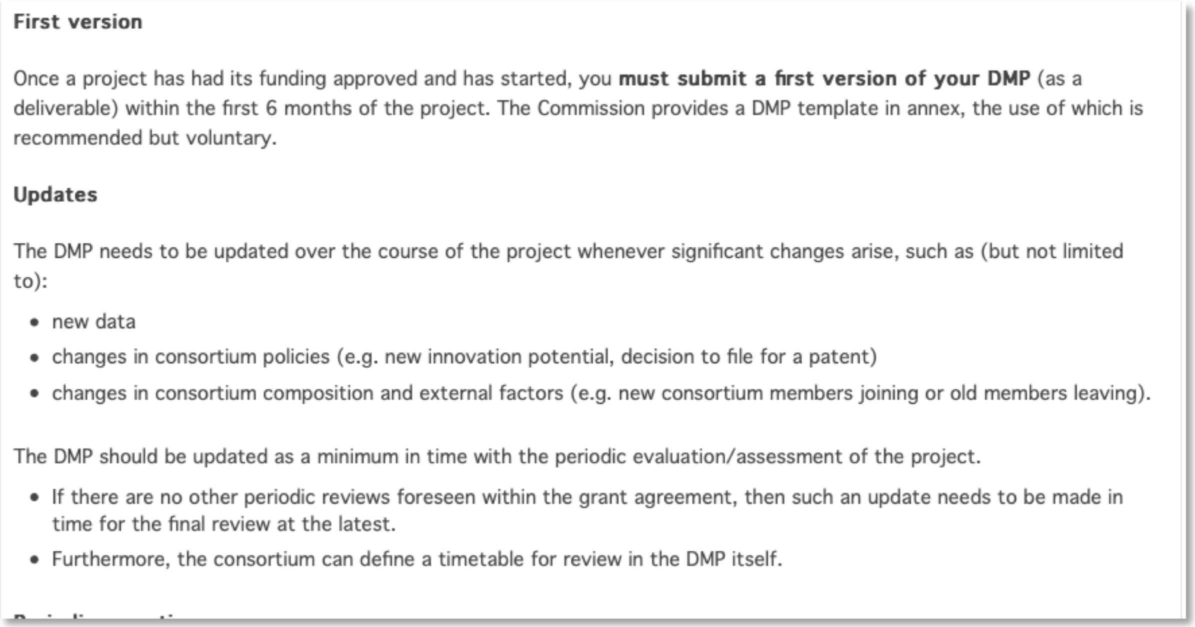
- For sensitive human data what is the access orchestration process?
- May be added during DMP update

*tip*

Prefer repositories that handle preservation and access-orchestration.

# DMP as a living document

- You will be expected to update your DMP



**First version**

Once a project has had its funding approved and has started, you **must submit a first version of your DMP** (as a deliverable) within the first 6 months of the project. The Commission provides a DMP template in annex, the use of which is recommended but voluntary.

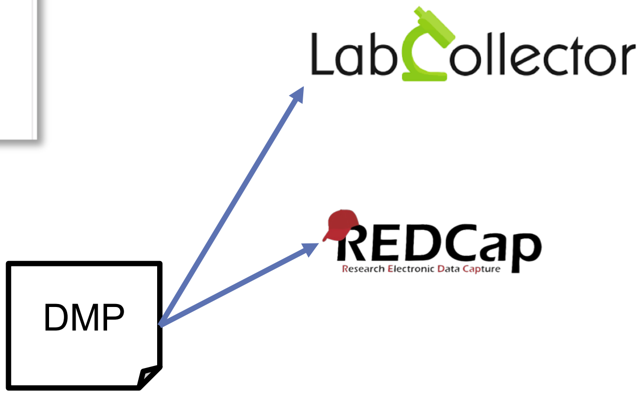
**Updates**

The DMP needs to be updated over the course of the project whenever significant changes arise, such as (but not limited to):

- new data
- changes in consortium policies (e.g. new innovation potential, decision to file for a patent)
- changes in consortium composition and external factors (e.g. new consortium members joining or old members leaving).

The DMP should be updated as a minimum in time with the periodic evaluation/assessment of the project.

- If there are no other periodic reviews foreseen within the grant agreement, then such an update needs to be made in time for the final review at the latest.
- Furthermore, the consortium can define a timetable for review in the DMP itself.



*tip*

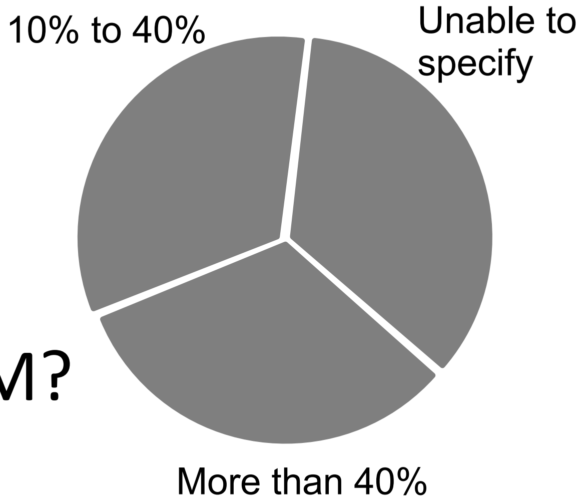
Annex DMP with executable specifications rather than static listings.

# Budgeting for Data Management

## — 5% of total project costs

**An 2016 survey:**

- Infra providers, libraries, universities
- What percentage of total budget of your organization is allocated for RDM?



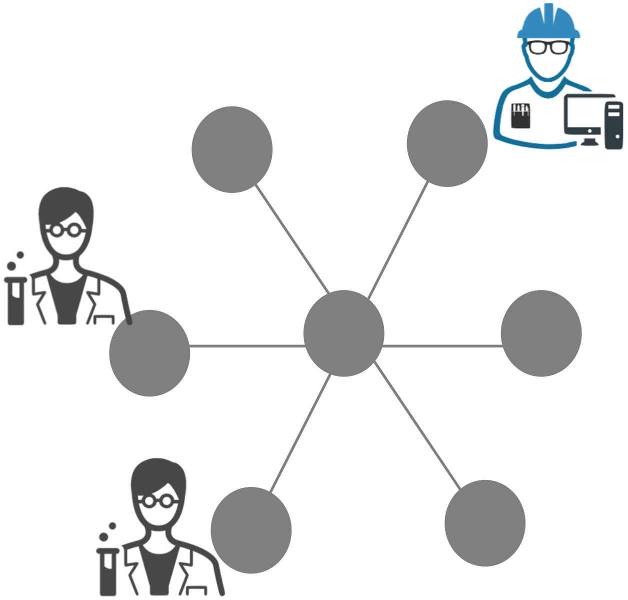
**A commonly cited recommendation:**

- An overall average of **5% of the total project costs** ……to sustain and share data”

** *Funding research data management and related infrastructures.* Knowledge Exchange and Science Europe briefing paper. May 2016

# Roles and responsibilities

## — DMP via “data specialist” partner



- Main author of the DMP
- Task planning and follow up.

*tip*

DMP tools may be of help to the “data special” partner
Or to researchers when writing their DMP for the very first time.

# Administrative information

- DMP ID
- Funder, Grant ID
- Project name description
- PI , Data Contact name and ID
- Dates
- Related policies

iD ORCiD
Connecting Research
and Researchers

doi DIGITAL
OBJECT
IDENTIFIER

# DMP tools

- UK DCC

DMPONLINE

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

https://dmponline.be

roadmap

https://github.com/DMPRoadmap/

- University of California DCC

DMPTool
Build your Data Management Plan

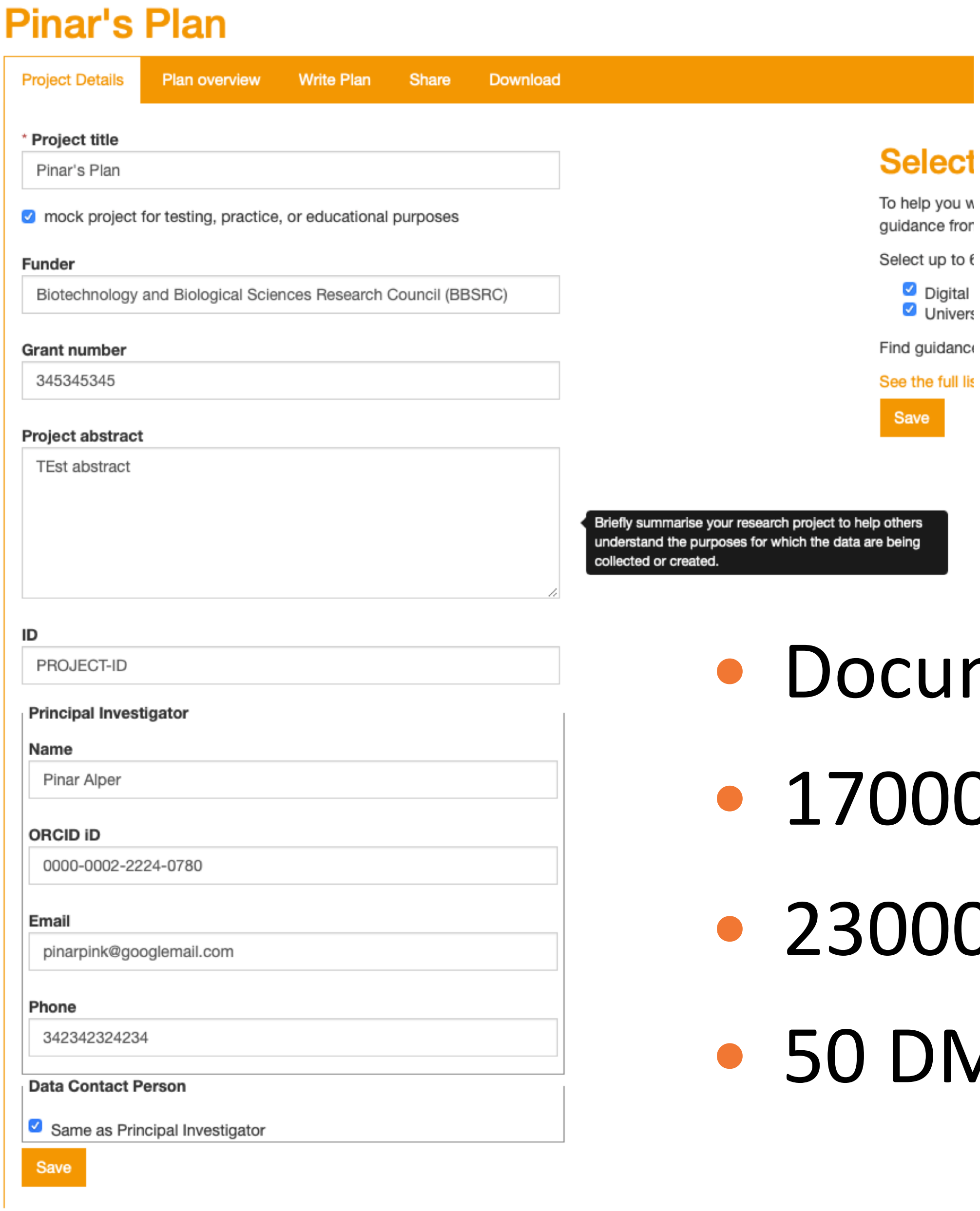
https://dmptool.org

- Dutch Techcenter for Life Sciences, ELIXIR-NL

DS Wizard

https://app.ds-wizard.org/welcome

# DMP Online



## Pinar's Plan

* Project title

Pinar's Plan

☑ mock project for testing, practice, or educational purposes

Funder

Biotechnology and Biological Sciences Research Council (BBSRC)

Grant number

345345345

Project abstract

TEst abstract

Briefly summarise your research project to help others understand the purposes for which the data are being collected or created.

ID

PROJECT-ID

Principal Investigator

Name

Pinar Alper

ORCID iD

0000-0002-2224-0780

Email

pinarpink@googlemail.com

Phone

342342324234

Data Contact Person

☑ Same as Principal Investigator

Select

To help you w
guidance fror

Select up to 6

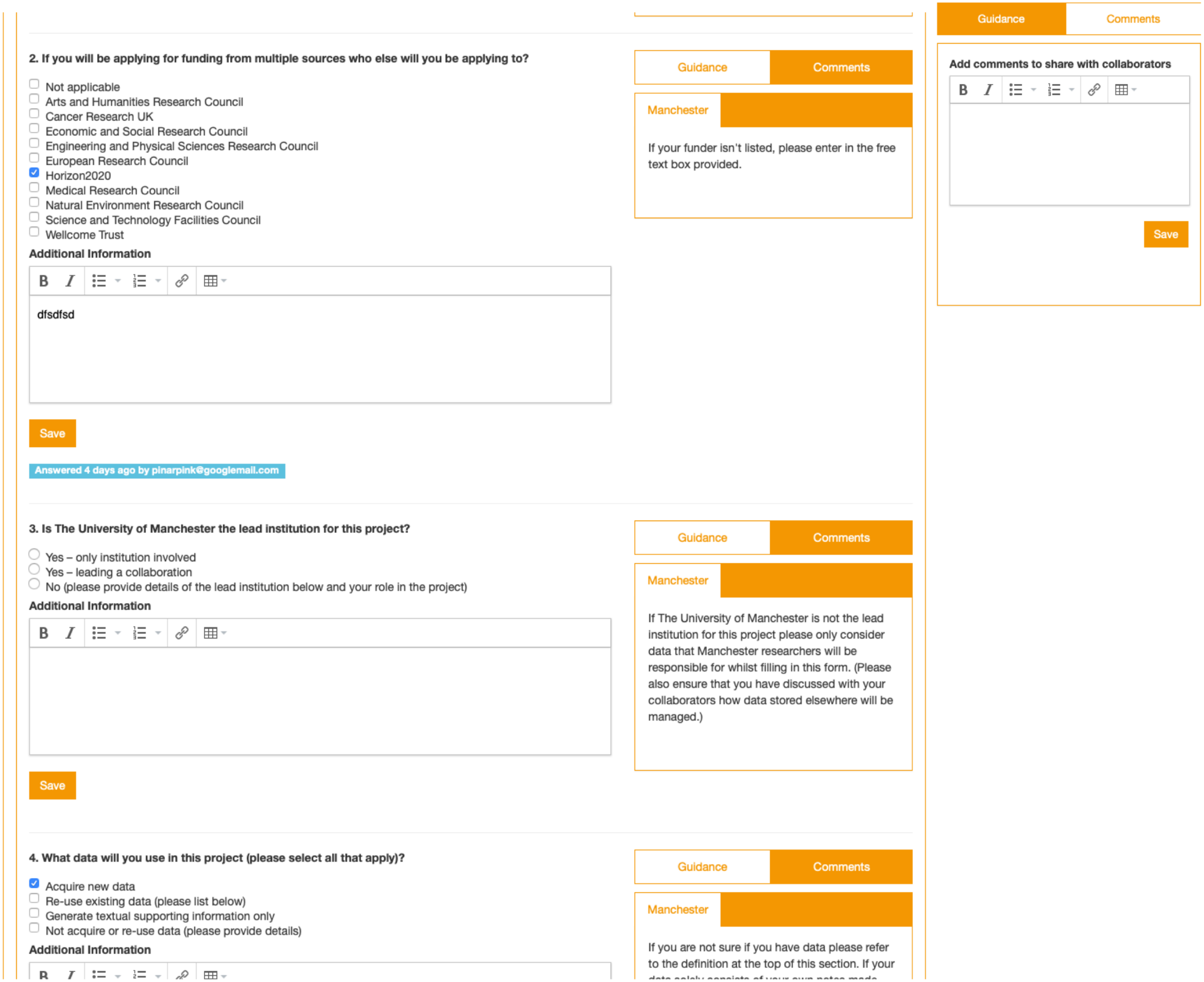
☑ Digital
☑ Univers

Find guidanc

See the full lis

- Document generator app
- 17000 + Users
- 23000 + Plans
- 50 DMPs publicly available

# DMP Online



2. If you will be applying for funding from multiple sources who else will you be applying to?

□ Not applicable
□ Arts and Humanities Research Council
□ Cancer Research UK
□ Economic and Social Research Council
□ Engineering and Physical Sciences Research Council
□ European Research Council
☑ Horizon2020
□ Medical Research Council
□ Natural Environment Research Council
□ Science and Technology Facilities Council
□ Wellcome Trust

Additional Information

dfsdfsd

Answered 4 days ago by pinarpink@googlemail.com

If your funder isn't listed, please enter in the free text box provided.

Add comments to share with collaborators

3. Is The University of Manchester the lead institution for this project?

○ Yes – only institution involved
○ Yes – leading a collaboration
○ No (please provide details of the lead institution below and your role in the project)

Additional Information

If The University of Manchester is not the lead institution for this project please only consider data that Manchester researchers will be responsible for whilst filling in this form. (Please also ensure that you have discussed with your collaborators how data stored elsewhere will be managed.)

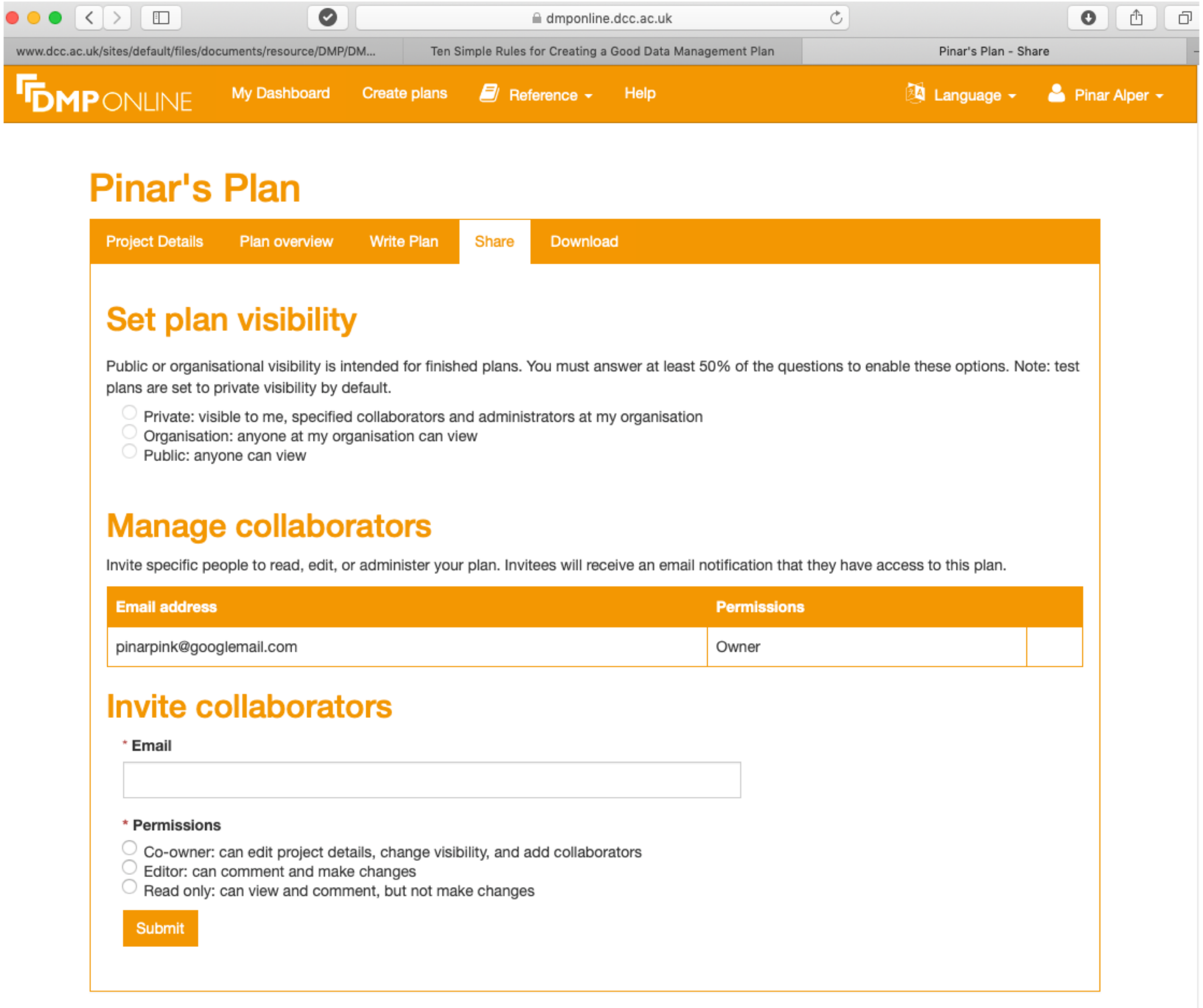
4. What data will you use in this project (please select all that apply)?

☑ Acquire new data
□ Re-use existing data (please list below)
□ Generate textual supporting information only
□ Not acquire or re-use data (please provide details)

Additional Information

If you are not sure if you have data please refer to the definition at the top of this section. If your data solely consists of your own notes made

# DMP Online



dmponline.dcc.ac.uk
www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DM...
Ten Simple Rules for Creating a Good Data Management Plan
Pinar's Plan - Share
DMPONLINE
My Dashboard
Create plans
Reference
Help
Language
Pinar Alper

## Pinar's Plan

## 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

## 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 | |
|---|---|---|
| pinarpink@googlemail.com | Owner | |

## Invite collaborators

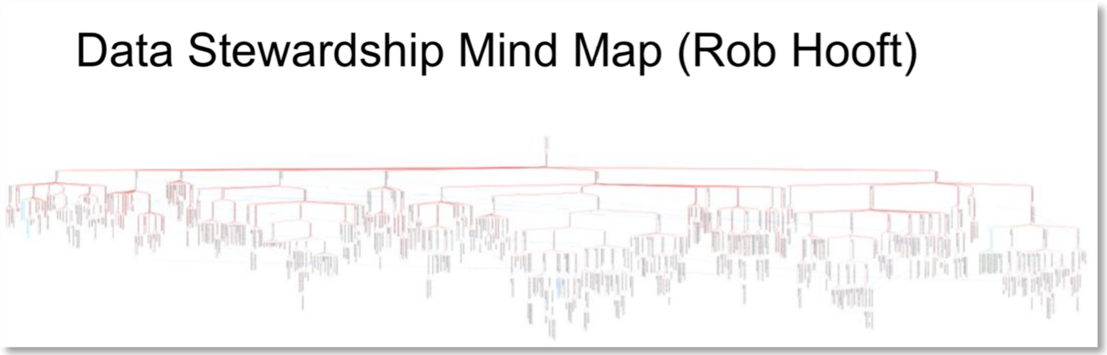
* Email

* Permissions

○ Co-owner: can edit project details, change visibility, and add collaborators
○ Editor: can comment and make changes
○ Read only: can view and comment, but not make changes

Submit

# ELIXIR DS Wizard



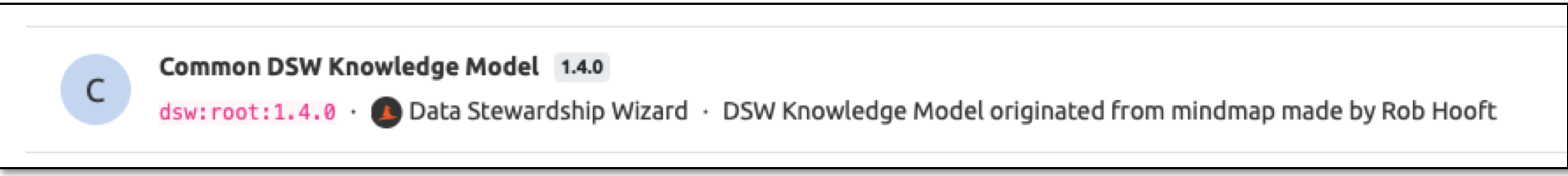
Data Stewardship Mind Map (Rob Hooft)

- “Insights collected over 4 years”
- “Guidance over 300+items ”
- Awareness tool
- Under active development

  https://github.com/ds-wizard/ds-wizard/issues

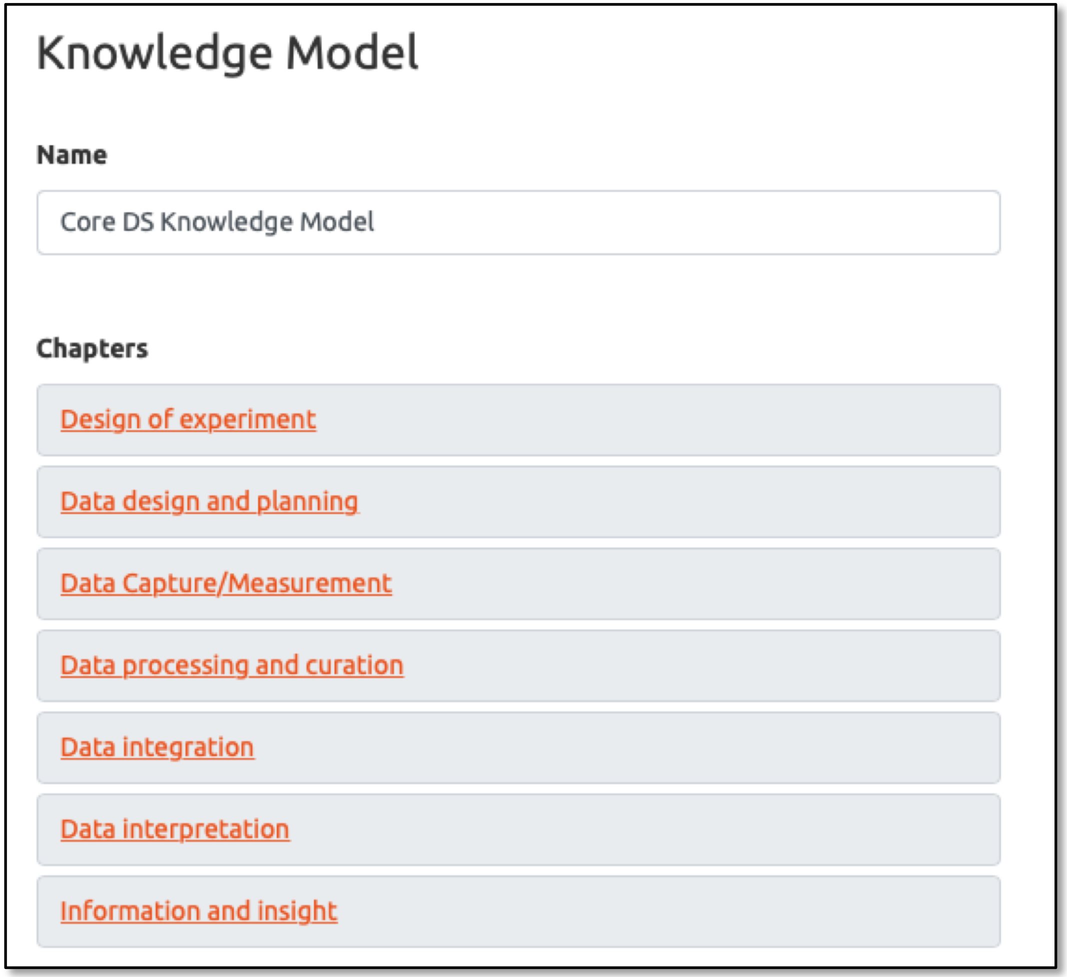
  info@ds-wizard.org

# Common DSW Knowledge Model



Common DSW Knowledge Model 1.4.0

dsw:root:1.4.0 · Data Stewardship Wizard · DSW Knowledge Model originated from mindmap made by Rob Hooft



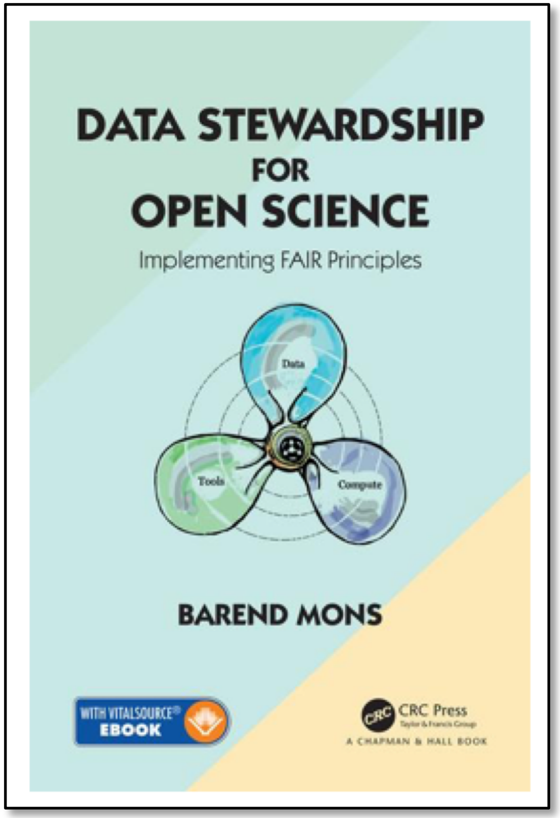
Knowledge Model

Name

Core DS Knowledge Model

Chapters

- Design of experiment
- Data design and planning
- Data Capture/Measurement
- Data processing and curation
- Data integration
- Data interpretation
- Information and insight



DATA STEWARDSHIP FOR OPEN SCIENCE

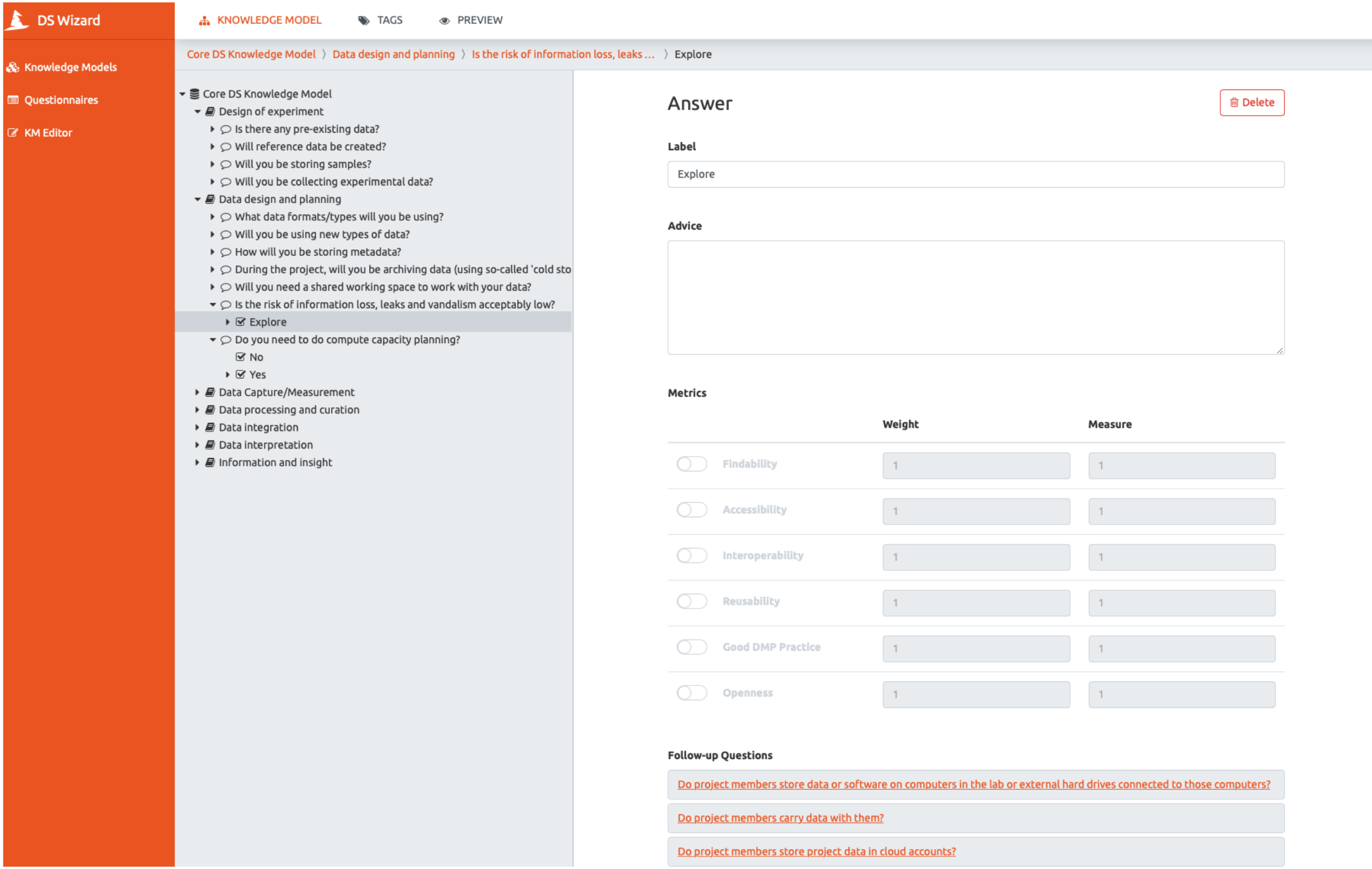
Implementing FAIR Principles

BAREND MONS

“All models are wrong but some are useful.”

*George E. P. Box*

# ELIXIR DS Wizard



DS Wizard

KNOWLEDGE MODEL    TAGS    PREVIEW

- Knowledge Models
- Questionnaires
- KM Editor

Core DS Knowledge Model 〉 Data design and planning 〉 Is the risk of information loss, leaks ... 〉 Explore

- Core DS Knowledge Model
  - Design of experiment
    - Is there any pre-existing data?
    - Will reference data be created?
    - Will you be storing samples?
    - Will you be collecting experimental data?
  - Data design and planning
    - What data formats/types will you be using?
    - Will you be using new types of data?
    - How will you be storing metadata?
    - During the project, will you be archiving data (using so-called 'cold sto
    - Will you need a shared working space to work with your data?
    - Is the risk of information loss, leaks and vandalism acceptably low?
      - Explore
    - Do you need to do compute capacity planning?
      - No
      - Yes
  - Data Capture/Measurement
  - Data processing and curation
  - Data integration
  - Data interpretation
  - Information and insight

## Answer

Delete

**Label**

Explore

**Advice**

**Metrics**

| | Weight | Measure |
|---|---|---|
| ○ Findability | 1 | 1 |
| ○ Accessibility | 1 | 1 |
| ○ Interoperability | 1 | 1 |
| ○ Reusability | 1 | 1 |
| ○ Good DMP Practice | 1 | 1 |
| ○ Openness | 1 | 1 |

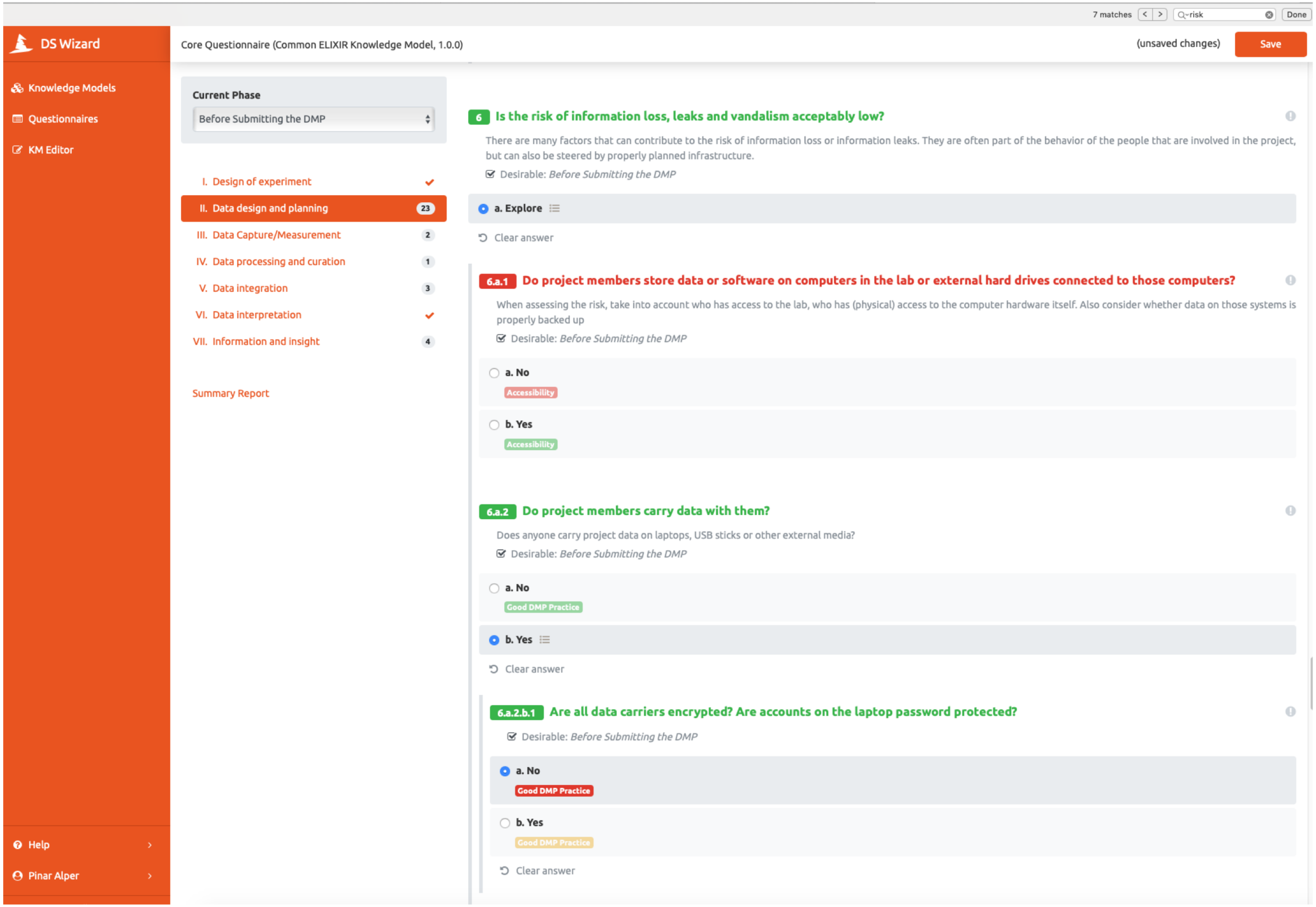
**Follow-up Questions**

Do project members store data or software on computers in the lab or external hard drives connected to those computers?

Do project members carry data with them?

Do project members store project data in cloud accounts?

# ELIXIR DS Wizard



Core Questionnaire (Common ELIXIR Knowledge Model, 1.0.0)

(unsaved changes)

**Current Phase**

Before Submitting the DMP

I. Design of experiment ✓
II. Data design and planning 23
III. Data Capture/Measurement 2
IV. Data processing and curation 1
V. Data integration 3
VI. Data interpretation ✓
VII. Information and insight 4

Summary Report

**6 Is the risk of information loss, leaks and vandalism acceptably low?**

There are many factors that can contribute to the risk of information loss or information leaks. They are often part of the behavior of the people that are involved in the project, but can also be steered by properly planned infrastructure.

☑ Desirable: *Before Submitting the DMP*

● a. Explore

Clear answer

**6.a.1 Do project members store data or software on computers in the lab or external hard drives connected to those computers?**

When assessing the risk, take into account who has access to the lab, who has (physical) access to the computer hardware itself. Also consider whether data on those systems is properly backed up

☑ Desirable: *Before Submitting the DMP*

○ a. No
Accessibility

○ b. Yes
Accessibility

**6.a.2 Do project members carry data with them?**

Does anyone carry project data on laptops, USB sticks or other external media?

☑ Desirable: *Before Submitting the DMP*

○ a. No
Good DMP Practice

● b. Yes

Clear answer

**6.a.2.b.1 Are all data carriers encrypted? Are accounts on the laptop password protected?**

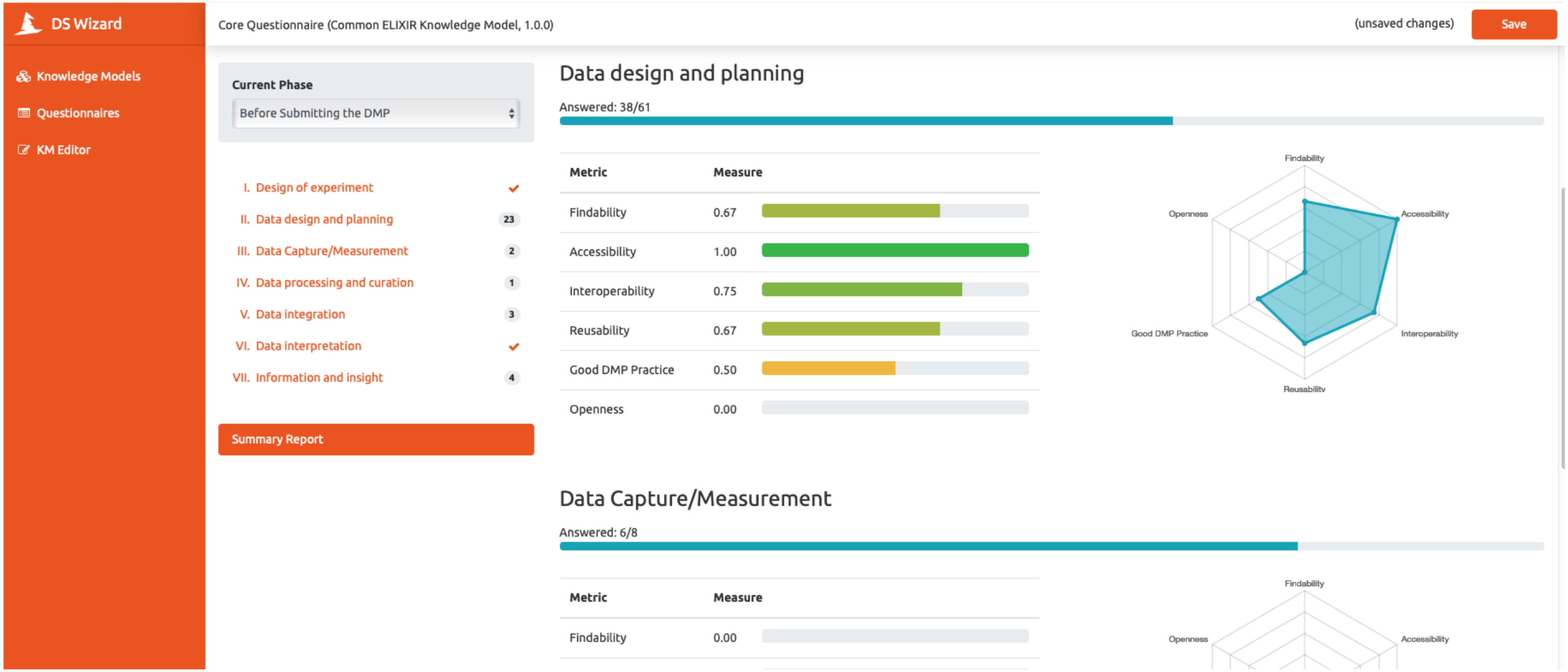
☑ Desirable: *Before Submitting the DMP*

● a. No
Good DMP Practice

○ b. Yes
Good DMP Practice

Clear answer

# ELIXIR DS Wizard



Core Questionnaire (Common ELIXIR Knowledge Model, 1.0.0)

**Current Phase**

Before Submitting the DMP

- I. Design of experiment ✓
- II. Data design and planning 23
- III. Data Capture/Measurement 2
- IV. Data processing and curation 1
- V. Data integration 3
- VI. Data interpretation ✓
- VII. Information and insight 4

Summary Report

## Data design and planning

Answered: 38/61

| Metric | Measure |
|---|---|
| Findability | 0.67 |
| Accessibility | 1.00 |
| Interoperability | 0.75 |
| Reusability | 0.67 |
| Good DMP Practice | 0.50 |
| Openness | 0.00 |

## Data Capture/Measurement

Answered: 6/8

| Metric | Measure |
|---|---|
| Findability | 0.00 |

# Should you use a DMP tool?

- DMP tools are not a match to state of collaborative document authoring tools.
- Embodies data management expertise.
- Not integrated to grant submission systems.

DMPONLINE

DMPTool

Build your Data Management Plan

DS Wizard

https://DMPOnline.be

Questionnaires to help the data specialist

Pooling DMP templates and instances

DMP tools may be of help when there is no “data expert” partner.

tip

Pooling DMP content has proved useful for some research communities.

Be mindful when re-using DMPs; document agreed-upon decisions not possibilities.

"Keep it short and specific"

# Data Stewardship Wizard
# -Practical-

elixir
LUXEMBOURG

# Scenario

You are the data specialist in project X, which works with human biosamples and data. In the DMP you had specified that the project will deposit its output data to relevant repositories. You had listed GigaDB and Nature Scientific Data as possible venues.

Project completion date is fast approaching and a few manuscripts are already under preparation. You check the submission requirements of above-mentioned venues, they require extra information in the case of human data. You need to collect this information from various consortium members, as data is acquired from various cohorts that collaborators have access.

You decide to encode Nature Scientific Data's submission checklist as a questionnaire for the consortium, and collect information this way.

# Practical

Application URL:

https://learning.ds-wizard.org

Usernames:

datasteward_XX@example.com

researcher_XX@example.com

*(see handout for your number)*

Password:

**<to-be-provided>**

Materials:

**<to-be-provided>**

# Acknowledgements

Jan Slifka, Vojtech Knaisl - ELIXIR Czech Republic

Celia van Gelder, Mateusz Kuzak - ELIXIR Netherlands

Wei Gu, Roland Krause - ELIXIR Luxembourg

Luxembourg Centre
for Systems Biomedicine

elixir
LUXEMBOURG