Data Management with Data Stewardship Wizard

Robert Pergl
robert.pergl@fit.cvut.cz
robert.pergl@codevence.com
Project Coordinator

DSW
DATA STEWARDSHIP WIZARD

FACULTY OF INFORMATION TECHNOLOGY
CTU IN PRAGUE

CODEVENICE
Contents

• Motivation
• Features
• Embedding in the FAIR ecosystem
• How to start
• Adoption
• Governance
• Learn more
Motivation

Data Management "from burden to benefit"

- A good "Fahrplan" - knowing the steps and needed resources
- Avoiding forgetting something important
- DMP = what, why and how to do
- DMP ≠ an essay exercise to satisfy a funder
- From beginners to data stewards

One size does not fit all

- Extensive customisation possibilities
- Openness (open-source code)
Challenge #1

How do you know your DMP is good?

- Unconsciously Incompetent
- Consciously Incompetent
- Consciously Competent
- Unconsciously Competent
Challenge #2  Making a good DMP encompasses a vast knowledge

- Rob Hooft (DTL, ELIXIR NL)
- Complex decision tree
- ~600 nodes

https://zenodo.org/record/2614820
Features

- Guidance
- Learning
- Effectiveness
Guidance: Pilot's Check List
1. Is there any pre-existing data?

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

- Desirable: Before Submitting the Proposal
- Data Stewardship for Open Science: [link]
- External links: [Google dataset search], [Datacite Search]

- a. No
- b. Yes

Clear answer

1.1. 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?

- Desirable: Before Submitting the Proposal
- Data Stewardship for Open Science: [link]

- a. No
- b. Yes

Clear answer
1.a.4.b.1.a.1 What repository will this data be stored in?

Domain repositories often have the best functionality to make the data findable and reusable; even though it may look like a database that could be reused in a completely different field would be better findable in a generic repository, the limited availability of domain-specific metadata make that less valuable.

Many repositories are listed in [https://fairsharing.org/](https://fairsharing.org/)

If a repository offers to give your data set a DOI or alternative persistent identifier it is a good idea to use that option.

- Desirable: Before Finishing the Project
- External links: [FAIRsharing](https://fairsharing.org), Registry of Research data Repositories

---

**Choice of Option**

- **a. A domain-specific repository**
- **b. Our national repository**
- **c. Our institutional repository**
- **d. A special-purpose repository for the project**

---

**Recommendation**

Disadvantage of a general purpose repository is the lack of data-specific features (e.g. 'play' instead of 'download' for an audio file) and limited findability.
Is there pre-existing data?

What's up?

For many decades if not centuries, virtually every experiment started with the collection or creation of "observations" and in fact data. In social sciences and humanities the tendency to "issue" data that had been created earlier, in all kinds of surveys and increasingly of course from social media maybe already somewhat more established. However, in many of the hard experimental sciences, the generation of new data specifically generated to answer a hypothetical question is still so commonplace that careful thinking about the actual need to generate new data may just not be on the radar screen. Obviously, data creation will need to continue, but increasingly we have to ask the question whether such new data are absolutely necessary to answer the question we want to answer. With more and more data becoming available in reusable formats, there may well be existing data collections "Other People's" Data and associated Services (OPEDAS) that without or with some extra effort needed, can answer at least part of the question or least may be crucial for the interpretation of your own data.

Do

- Search for data sets (OPEDAS) that may be re-usable and can help you to reduce the number of new data sets you may have to generate (and steward later on).
- Include annotated collections of data and curated databases in your search.
- Check the accessibility and license situation attached to the relevant data sets you found.
- Check their interoperability. They may be relevant but not interoperable with your analysis pipelines. In that case you may have to extract, transform and load (ETL) them or decide that although relevant they are not reusable for your purpose.
- Ensure that using OPEDAS will not restrict in any way the use of your results later on, including copyright and freedom to operate on the request of IPR.
- Check how to cite and acknowledge OPEDAS.
- Consider to actively involved OPEDAS owners in your research in order to make optimal use of their data.
- Speak to colleagues who did similar experiments before to find out about potential OPEDAS you may consider to use.

Don't

- Assume no OPEDAS exist without thorough checking using all your possibilities.
- Start an experiment without properly checking with colleagues about the best approach and OPEDAS out there.
- Budget for data generation in your study without justifying to the funder why the generation of the data is necessary.
- Move into actual experimentation without consulting a data expert.

Links

- DS Question GitHub resources repository: atq
Summary Report

Answered (current phase): 100/100
Answered: 253/305

<table>
<thead>
<tr>
<th>Metric</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Findability</td>
<td>0.76</td>
</tr>
<tr>
<td>Accessibility</td>
<td>0.90</td>
</tr>
<tr>
<td>Interoperability</td>
<td>0.67</td>
</tr>
<tr>
<td>Reusability</td>
<td>0.92</td>
</tr>
<tr>
<td>Good DMP Practice</td>
<td>0.84</td>
</tr>
<tr>
<td>Openness</td>
<td>0.72</td>
</tr>
</tbody>
</table>
Effectiveness: Autocomplete

2.7 Funding
Add all the funding that are part of this project.
Desirable: Before Submitting the Proposal

2.7.1.1 Funder
Czech
Ministerstvo Obrany České Republiky
Grantová Agentura České Republiky
České Vysoké Učení Technické v Praze
Parazitologický ústav, Akademie Věd České Republiky
Ministerstvo Zdravotnictví České Republiky
Effectiveness: Collaboration - Sharing

Share Project

Users
Add users

Vojtěch Knaiel
Owner

Jan Slířka
Editor

Marek Suchánek
Viewer

Visible by all other logged-in users
Other logged-in users can view the Project.

Public link

Cancel
Save
Effectiveness: Collaboration - Online

II. Re-using data

Before you decide to embark on any new study, it is nowadays good practice to check all options to re-use existing available data, either collected or generated by yourself in an earlier project, or data from others (Barend Mons calls this “Other People’s Data And Services” or OPEDAS). This can include reusable data that have been created for an earlier study, and also so-called “reference data” which is used by many projects.

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. And the circle is only complete if such data is actually re-used.

1 Is there any pre-existing data?

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

☐ Desirable: Before Submitting the Proposal
☐ Data Stewardship for Open Science: easy

a. No
b. Yes

Clear answer

1.b. 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?

☐ Desirable: Before Submitting the Proposal
☐ Data Stewardship for Open Science: easy

a. No
b. Yes
Effectiveness: TODOs

2. What existing encodings/terminologies/vocabularies/ontologies will you be using?

3. Will you be using new types of data?

Sometimes the type of data you collect cannot be stored in a commonly used data format. In such cases you may need to make your own, keeping interoperability as high as possible.

Desirable: Before Submitting the Proposal
Data Stewardship for Open Science: ikk

- a. No, all of my data will fit in common formats

- b. Yes, I will need to use custom formats for some of my data

Clear answer

Re-using data
- Is there any pre-existing data?

Creating and collecting data
- Will you be using new types of data?

Processing data
- Data storage systems and file naming conventions

Giving access to data
- Can all of your data become completely open immediately?
Effectiveness: Version history

3 Will you be using new types of data?

- Horizon 2020 DMP
- Named versions only

<table>
<thead>
<tr>
<th>March 2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>21. 3.</td>
</tr>
</tbody>
</table>

15:51

Set phase to Before Submitting the DMP

Marek Suchánek

- a. No, all of my data will fit in common formats

- Clear answer

Answered 4 minutes ago by Marek Suchánek.

4 How will you be collecting and Keeping your metadata?

- Horizon 2020 DMP
- Science Europe DMP

For the re-usability of your data by yourself or others at a later stage, a lot of information about the data, how it was collected and how it can be used should be stored with the data. Such data about the data is called metadata, and this set of questions are about this metadata.

- Set phase to Before Submitting the DMP

Marek Suchánek
Effectiveness: DMP Export in Various Templates

Data Management Plan

Science Europe Example DMP

Contact person: Jana Freeman (jana.freeman@ds-wizard.org)
Czech Technical University in Prague
Based on: Common DSW Knowledge Model, 2.3.0 (doi:10.20964/2.3.0)
Created by: Jana Freeman (jana.freeman@ctu.cz)
CVUT
Generated on: 18 Mar 2021

Data Management Plan created in Data Stewardship Wizard - ds-wizard.org
The Questionnaire is fully customisable

- Chapters
- Questions (different types)
- Answers (choice of option)
- References and Experts
- Tags
- Integration for answer suggesting
The questionnaire is fully customisable

Built-in Knowledge Model Editor

Create „from the scratch“

Edit existing ones:

- Delete non-relevant questions
- Add relevant ones
- Change existing questions
- Structure change
Embedding in the FAIR Ecosystem
Integrations through REST API

2.a.7.a.1 Funder

Univerzita Karlova v Praze

http://dx.doi.org/10.13039/100007397

Specify the name of the funder that you ask for funding for your project. If the funder is not present in the suggested list, please specify a complete URL to the funder website.
Machine-Actionable DMPs
Machine-Actionable DMPs

New document

Name
My first maDMP

Answered (current phase): 9/45
Answered: 9/58

Template
maDMP (RDA DMP Common Standard) 1.4.0
Machine-actionable DMP according to RDA Common Standard

Format
- ◼️ JSON
- ◼️ Turtle
- ◼️ N3
- ◼️ RDF/XML
- ◼️ JSON-LD
- ◼️ N-Triples
- ◼️ TriG

Create

```xml
@prefix dct: <http://purl.org/dc/terms/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dco: <https://w3id.org/dcso/ns/core#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dcs: <https://w3id.org/dcso/1.0/> .

# DMP
https://demo.ds.wizard.org/questionnaires/b3d89775-64aa-4342-916e-c010629036f7

a dcs:DCS0;  
dct:title "Marek" ;  
dct:description 'This maDMP has been created using Data Stewardship Wizard (DSW, ds-wizard.org) and is based on knowledge model Common DSW Knowledge Model (dsw:root:2.3.0). The questionnaire used for this DMP is identified by UUID "b3d89775-64aa-4342-916e-c010629036f7" within https://demo.ds.wizard.org DSW instance." ;  
dct:language dcs:lang:en ;  
dcs:ethicalIssuesExist 'unknown' ;  
dcs:hasDCS0Id <https://demo.ds.wizard.org/questionnaires/b3d89775-64aa-4342-916e-c010629036f7> Identifier ;  
dct:created "2021-03-22T07:11:40Z"^^xsd:dateTime ;  
dct:modified "2021-03-22T07:11:40Z"^^xsd:dateTime .

# DMP identifier
https://demo.ds.wizard.org/questionnaires/b3d89775-64aa-4342-916e-c010629036f7/identifier

a dcs:DCS0ID ;  
dcs:identifier_type "url" ;  

# DMP contact

This file has been exported from Data Stewardship Wizard (ds-wizard.org)  
It is using DCSO 3.0.2 (see https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard)
```
• **FAIR Implementation Profile** represents the implementation strategies of a FAIR community

• Such implementation strategies entail decisions about which FAIR-enabling resources are used, planned-to-be-used, or not-used

"Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence"

[https://link.springer.com/chapter/10.1007/978-3-030-65847-2_13](https://link.springer.com/chapter/10.1007/978-3-030-65847-2_13)
CRF Wizard

[Diagram]

- WHO CRF Metadata
  - enter metadata only once
  - provenance license rights

- Semantic Model
  - COVID-19 Case
    - patient ID
    - lab results
    - vital signs
  - DSW
    - on admission
    - daily
    - on discharge

- FAIR Data Point
  - submit
  - update
  - uses

- Triple Store
  - store

- SPARQL queries

- DSW
  - submit

- WHO CRF Data
  - fill eCRF form

[Link]
We will use guided choices through the DSW to help people find the ELIXIR Tools and Toolkits most appropriate for their case.

The Wizard is supposed to ask the same questions that an expert would ask if a researcher comes to ask for help.
How to Start
How to Start

https://ds-wizard.org/get-started.html

Demo
For exploring the DSW features
- Easy to sign up and use
- A shared instance with other users
- Not for serious usage
- Questionnaires & DMPs
- Knowledge Models Management

Try the demo

Researchers
For individual researchers
- Easy to sign up and use
- Ready to use Knowledge Models
- Privacy and stability
- Questionnaires & DMPs

Start planning

Self-Managed
For organizations
- All the DSW features available
- Your own instance
- You need to host and run the instance by yourself
- Questionnaires & DMPs
- Knowledge Models Management
- User Management & Organization Settings
- Features configuration & visual style customizations

Contact us

DSW Cloud
For organizations
- We offer managing the DS Wizard instance for interesting projects that want to use it seriously but don't want to run it by themselves.
- Questionnaires & DMPs
- Knowledge Models Management
- User Management & Organization Settings
- Features configuration & visual style customizations

Contact us

Free for research institutions (ELIXIR funded)
Going professional

• Codevence professional services
  • Professional cloud instances with custom branding & styling
  • Deployment and configuration support
  • Consulting
  • Guaranteed user support
  • Tailored training plans
  • Custom document templates and exports
  • Integration into company infrastructure
  • Custom features development
Adoption

- ~1500 registered users in the ELIXIR DEMO/Researchers cloud instances
- 14 institutions having their own cloud version
- ~15 known on-premises installations
Governance

• Strongly user-driven:
  • User feedback & bug report – from the UI, goes into GitHub
  • User ideas https://ideas.ds-wizard.org

• Advisory Board
Learn More

- Features overview: [https://ds-wizard.org/features.html](https://ds-wizard.org/features.html)
- Resources (webinars, workshops, presentations, publications, posters): [https://ds-wizard.org/resources.html](https://ds-wizard.org/resources.html)
- Comparison with other tools: [https://ds-wizard.org/comparison.html](https://ds-wizard.org/comparison.html)
DSW Core Team

Robert Pergl  
Project Coordinator

Rob Hooft  
DS Knowledge Expert

Vojtěch Knaisl  
Backend Developer

Jan Slifka  
Frontend Developer

Marek Suchánek  
Leonardo da Vinci

Jana Freeman  
Community Manager

Tereza Machačová  
Intern

Jiří Vondrášek  
Business and Strategic Advisory
Acknowledgement

The development and operation of DSW is supported by ELIXIR CZ research infrastructure (MŠMT Grant No.: LM2018131).
Thank you!

info@ds-wizard.org