DSW
DATA STEWARDSHIP WIZARD

Scientific Data Management Training

5. 12. 2024

GAČR
CZECH SCIENCE FOUNDATION

ostrails

elixir
CZECH REPUBLIC

elixir
Recommended Interoperability Resources

FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

Jana Martínková
Jana.martinkova@cvut.cz
0000-0001-8575-6533

About me

Ing. Jana Martínková

- FIT ČVUT
- ELIXIR CZ
- OSTrails

- DSW Team, DMP Methodology Specialist

jana.martinkova@cvut.cz

Outline

- Motivation
- Research Data Management
- FAIR
- Data Management Plan
- Data Stewardship Wizard
  - Demonstration

Motivation

- **Data management plans are necessary**
- *DMP is a formal requirement*

  → but they also gather valuable information about your research
- *Writing a DMP is a tedious work*

  → answer questions instead
- *Researchers are not always aware of the best practices*

  → provide guidance

Why?

- Required by organizations or funding agencies
- Planning resources and equipment
- Defines roles and responsibilities
- Identifies risks and its solutions
- Facilitates data sharing, reusability, and preservation
- Prevention vs. firefighting

DSW

DATA STEWARDSHIP WIZARD

Research Data Management

Data Life Cycle

https://rdmkit.elixir-europe.org

Data Life Cycle: Plan

- Planning on how we will work with data during (and after) the project
- Output should be a **Data Management Plan**

Data Life Cycle: Collect

- Collecting new data (methods differ according to the research area)
- Using existing data (e.g., from previous projects)
- Recording the origin of the data (samples, researchers, tools, etc.)
- Emphasis on the quality of recorded data

Data Life Cycle: Process

- Conversion of data from the recorded format to a format suitable for analysis
- Exclusion of bad data or data of low quality
- Pseudonymization/anonymization of sensitive data

Data Life Cycle: Analyse

- Exploration of collected data
- Main part of research – gaining new knowledge
- The workflow used in the analysis should be reproducible
- Big data analysis may require significant computational power

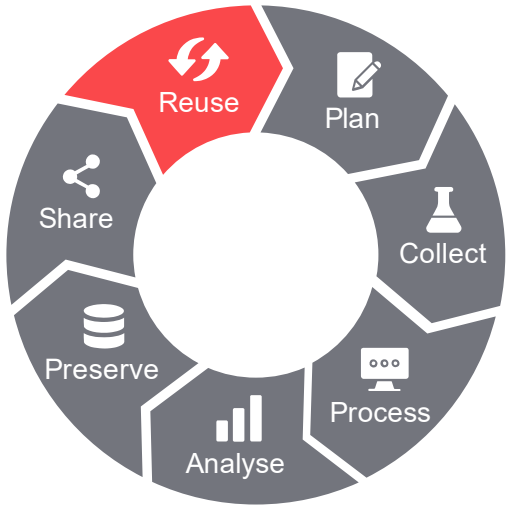
Data Life Cycle: Preserve

- Ensuring long-term preservation of data after the project ends
- The possibility to verify the project's results even after several years
- Using the data in the future for other purposes (teaching, other research)

Data Life Cycle: Share

- Sharing data with others (e.g., in another research project)
- Sharing does not mean that the data must be publicly available; they can be shared with limited access
- Consideration of all ethical, legal, licensing, and other restrictions

Data Life Cycle: Reuse



- Using data for a purpose other than for which they were collected, for example:
  - As reference data for other research
  - Verification of the original research results
  - Combining results from multiple studies into meta-studies

RDM Roles

- **Researcher** = work with data during the research project, create DMP for their project
  - PhD, grant applicants, project managers…
- **Data Steward** = takes care of data on several levels
  - **Policy** = cooperate with management and grant agencies; oversee processes, ethic and legal issues
  - **Research** = cooperation with the scientists, help with DMP creation
  - **Infrastructure** = cooperation with IT; IT solutions for RDM, infrastructure…

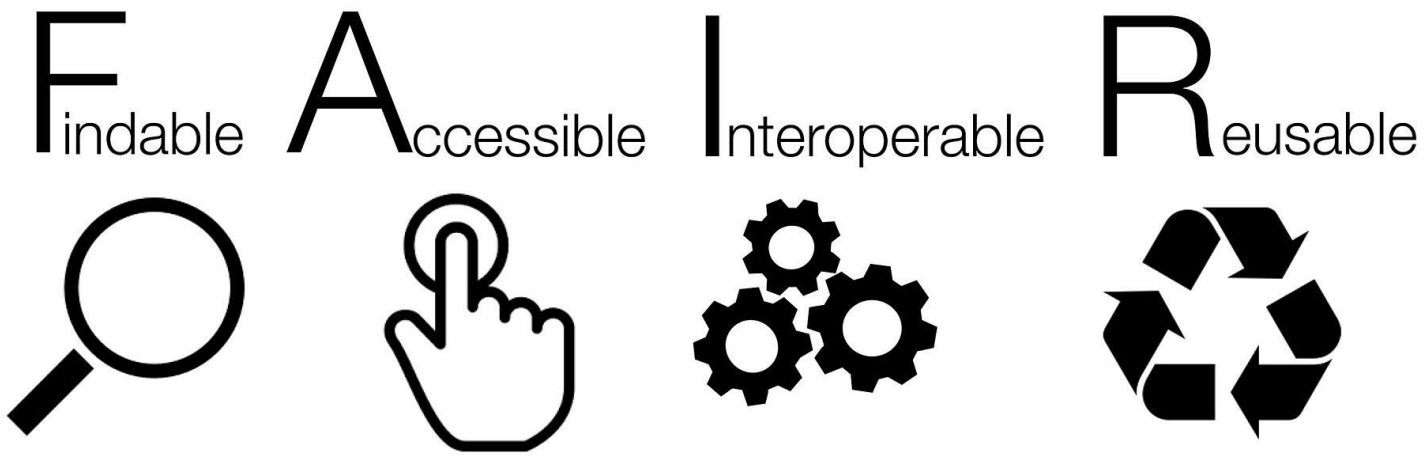
DSW

DATA STEWARDSHIP WIZARD

FAIR

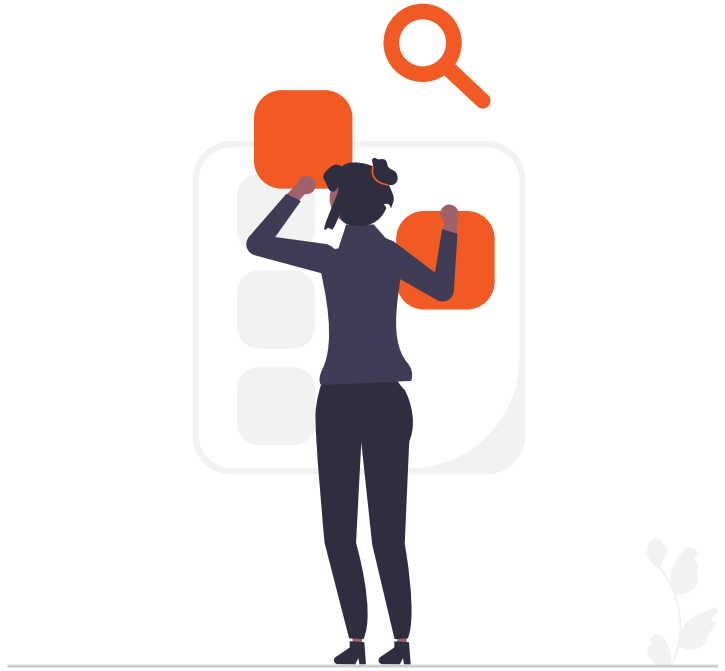
FAIR Data

- What is FAIR?
- Principles applied on data to make them:



Findability

- The first step towards reusability is finding the data
  - Data are described with rich metadata
  - (Meta)data have a globally unique persistent identifier
  - (Meta)data are registered in a searchable resource

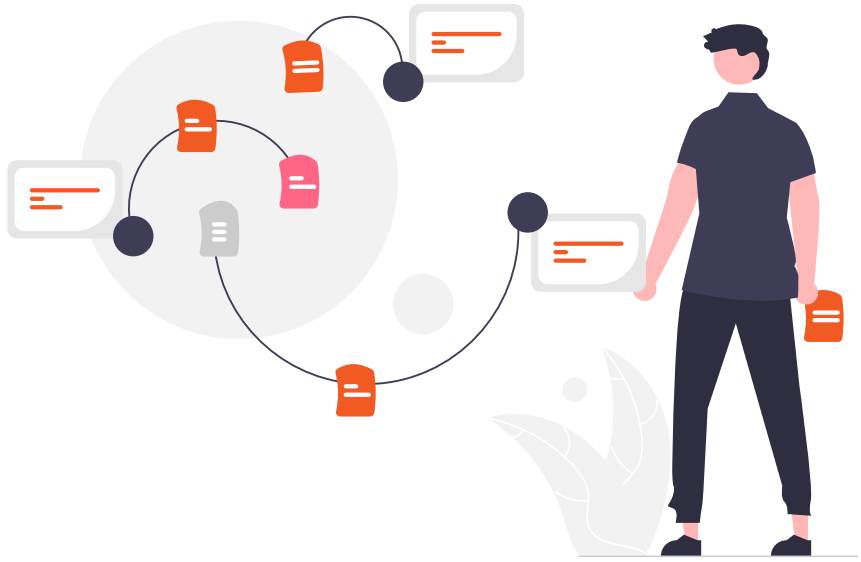


Accessibility

- After finding the data, it must be clear how they can be accessed
  - **Accessible ≠ Open**
  - Access to (meta)data via a standard communication protocol
  - Use of authentication and authorization if needed

Interoperability

- Ability to integrate data with other data
- Ability to process in various applications and workflows
- Use of standard formats, vocabularies, and ontologies (RDF, JSON-LD, OWL)



Reusability

- The main goal of FAIR is reusability
- Quality description of (meta)data
  - Licenses
  - Origin
  - Community standards in the given domain

Why should we apply FAIR Principles?

- Who will reuse the data?
  - **You**
  - Colleagues in the team
  - Colleagues within the institution
  - Someone else

DSW

DATA STEWARDSHIP WIZARD

Data Management Plan

Data Management Plan (DMP)

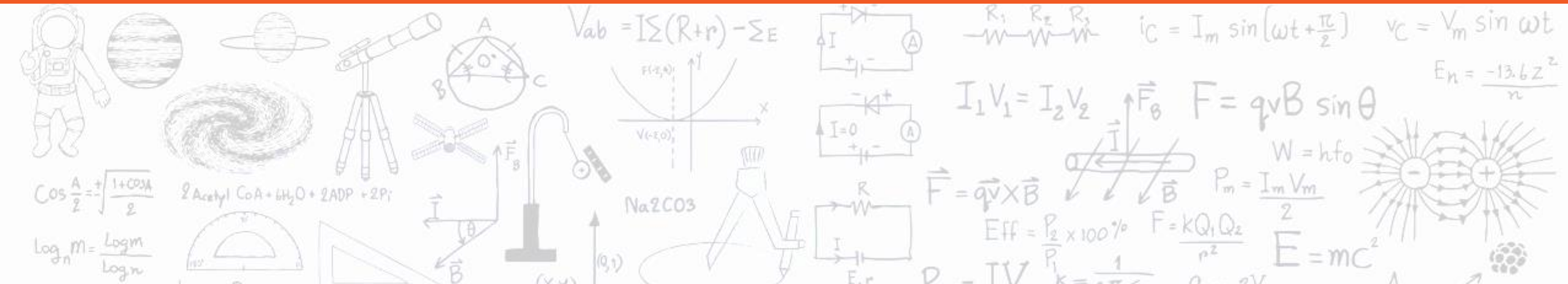
- General project information
- Description of data
- Metadata and ontologies, documentation
- Data storage, security, data preservation
- Data sharing
- Costs and human resources
- Ethical and legal issues, licenses

DMP as pre-flight checklist

DSW

DATA STEWARDSHIP WIZARD

Data Stewardship Wizard



What is Data Stewardship Wizard?

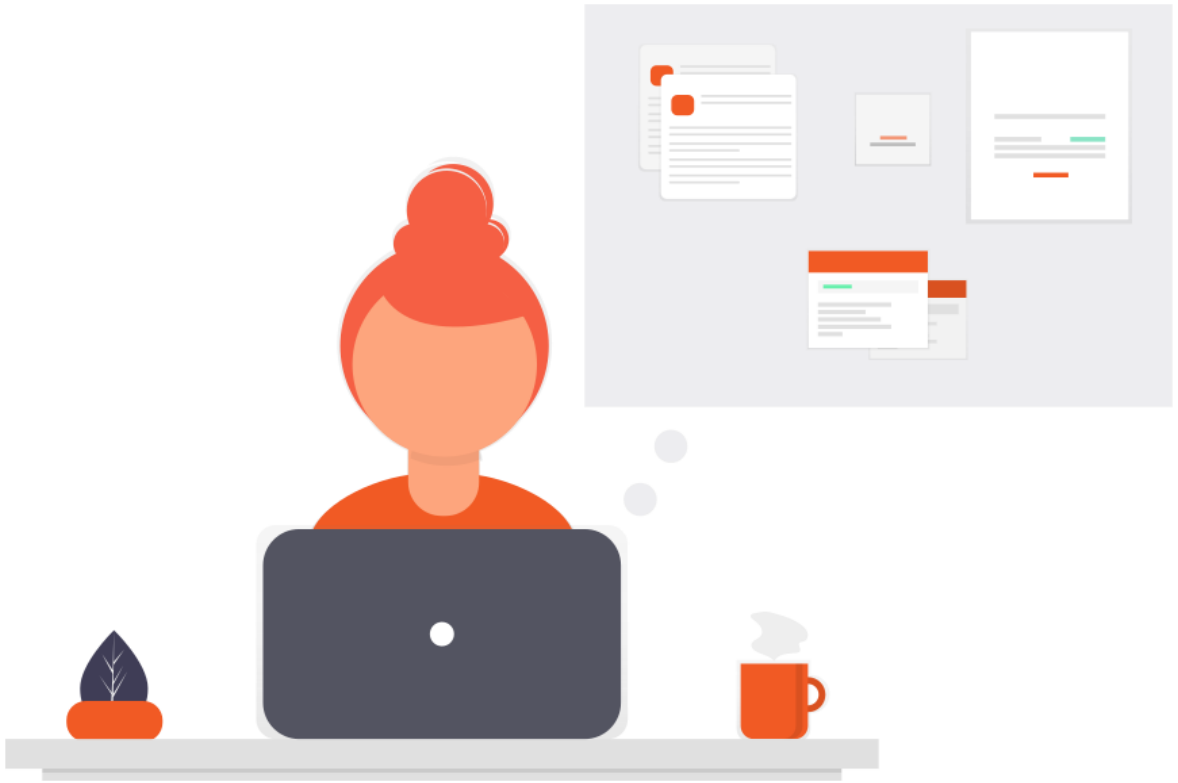
- Expert system for data management planning and creating DMPs
  - *"From burden to benefit"*
  - *"Plans are worthless, but planning is everything"*
- Open-source tool
- Tool suitable for everyone (from beginners to data stewards)
- Serves as a check-list before starting the project
- Supports data management planning with respect to (current) best practices
- A recommended tool in the **Horizon Europe Program Guide, ELIXIR Recommended Interoperability Resources**

DSW Main Ideas

- Minimum of writing = DMP is not an essay, as little writing as possible
- Guidance = DSW guides users through the smart Questionnaire
- Flexibility = easy to edit the content and integrate with other services
- Openness = anybody can use it and create own content
- User-oriented = DSW development is strongly user feedback driven

Demonstration

DSW

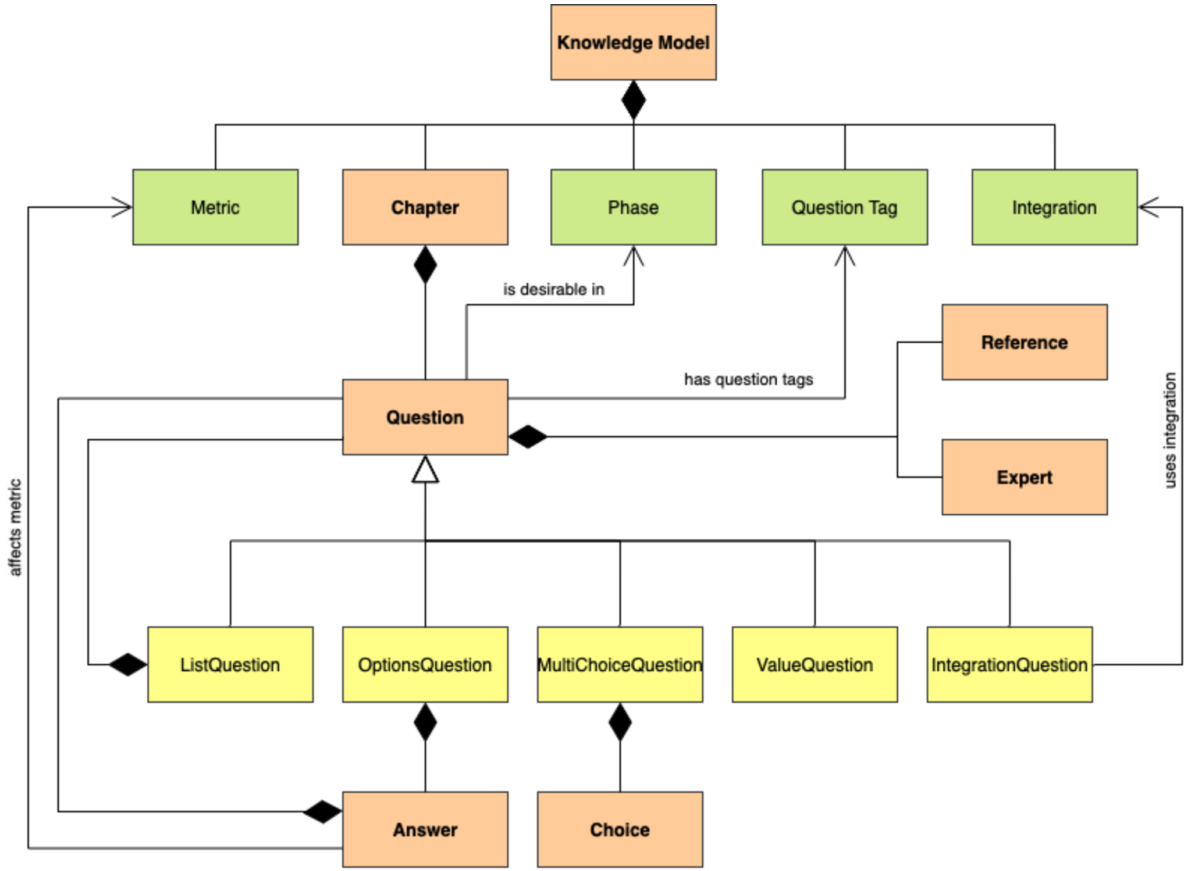


Data Stewardship Wizard

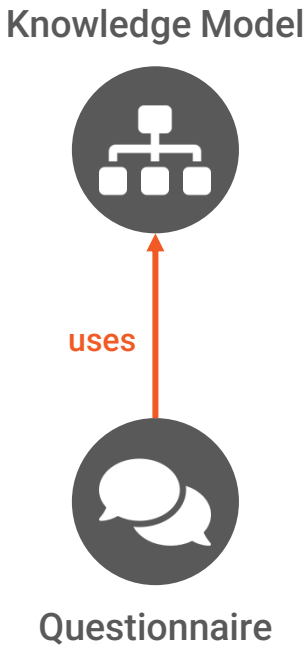
Knowledge Model

Knowledge Model (KM)

- Contains **knowledge** about what to ask for and how
- **Template** for structured questionnaire
- Tree structure build of **chapters, questions, answers**, follow-up questions and other resources

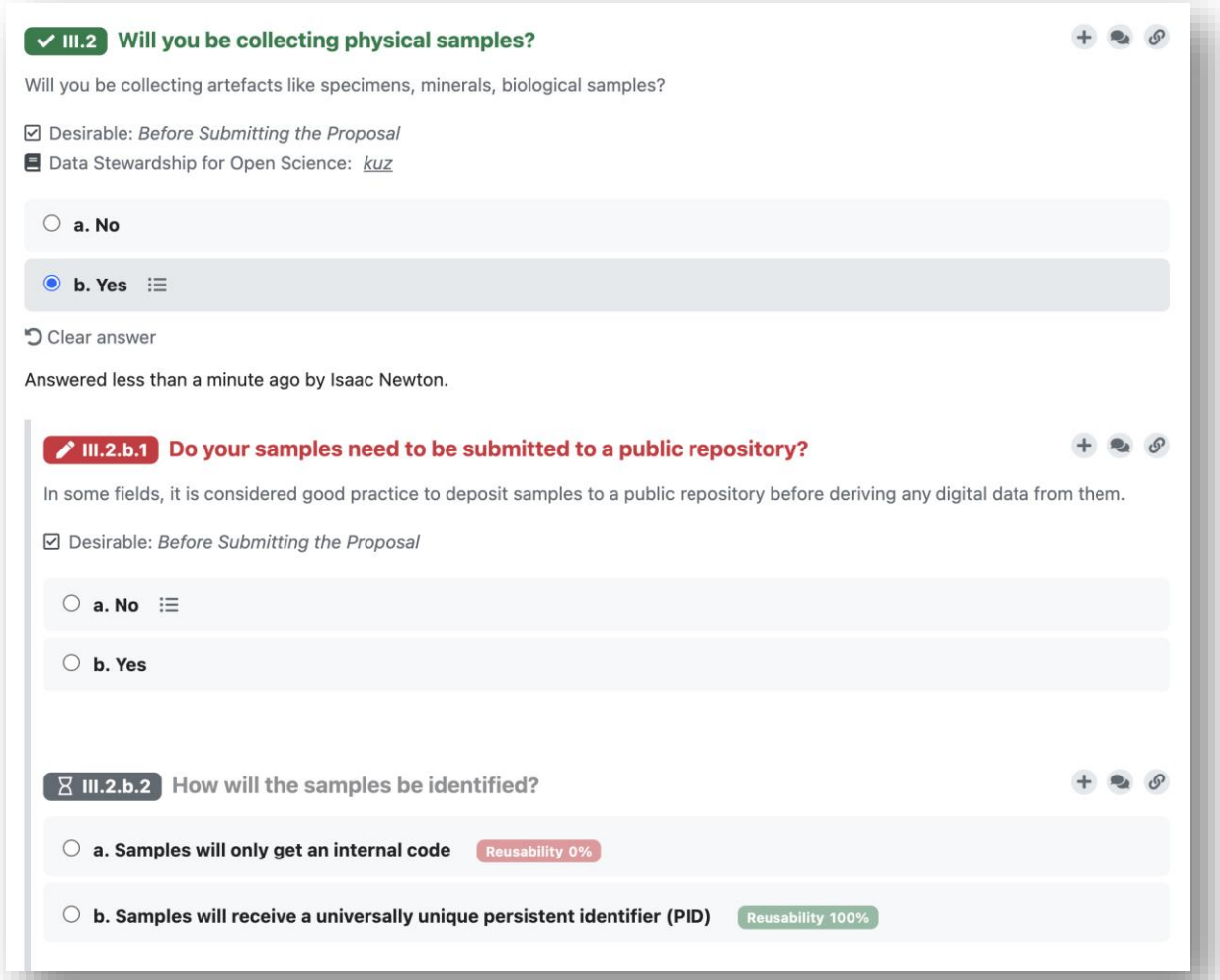


Data Stewardship Wizard

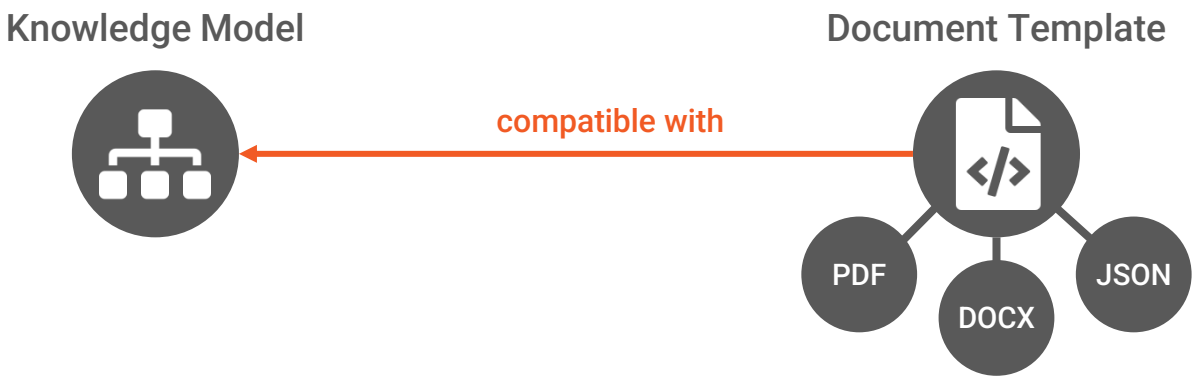


Questionnaire

- Can be filled in any order
- Only relevant questions are asked based on the previous answers
- Links to additional resources
- Metrics indications

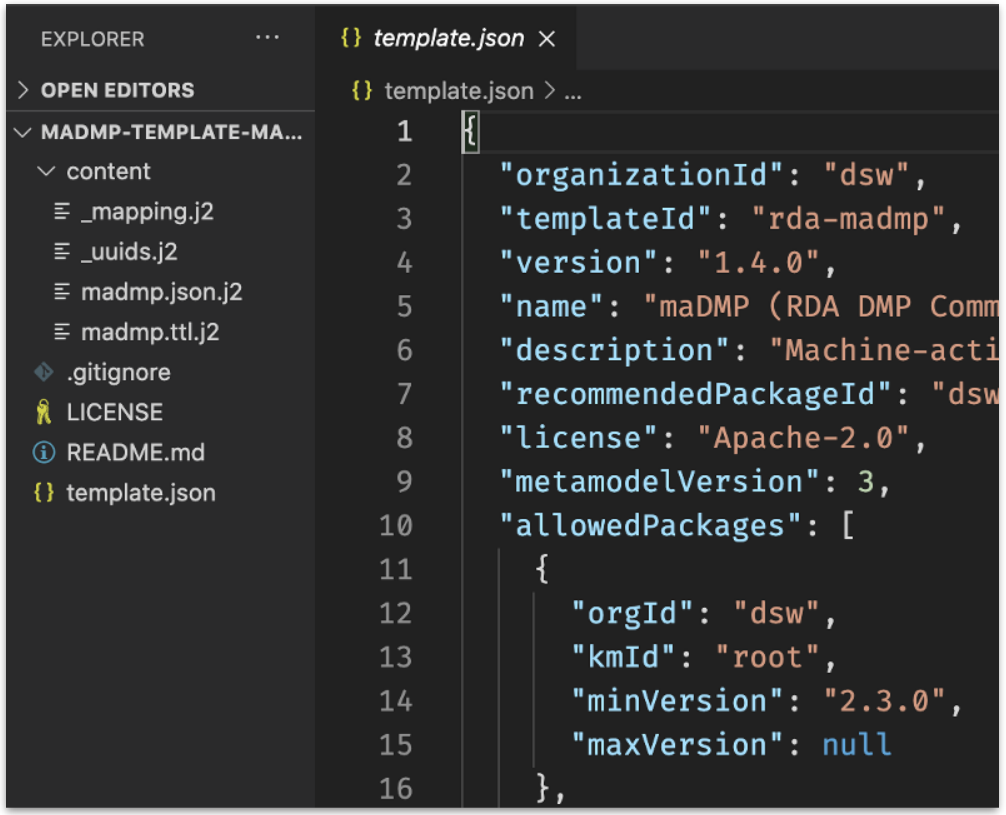


Data Stewardship Wizard

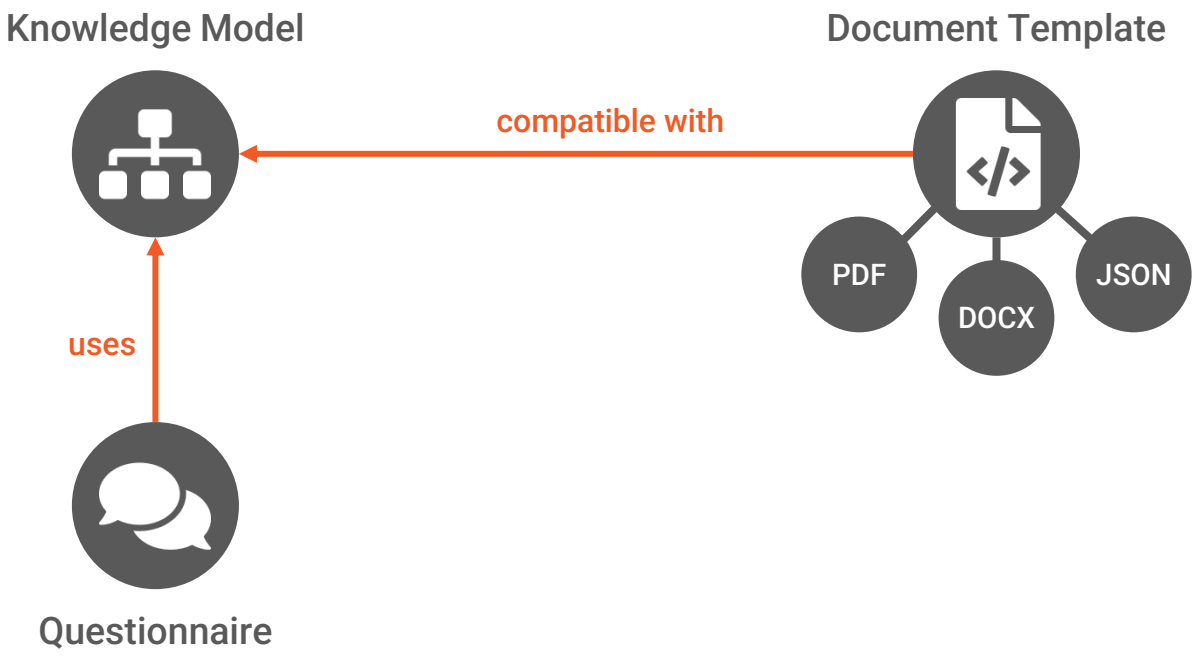


Document Template

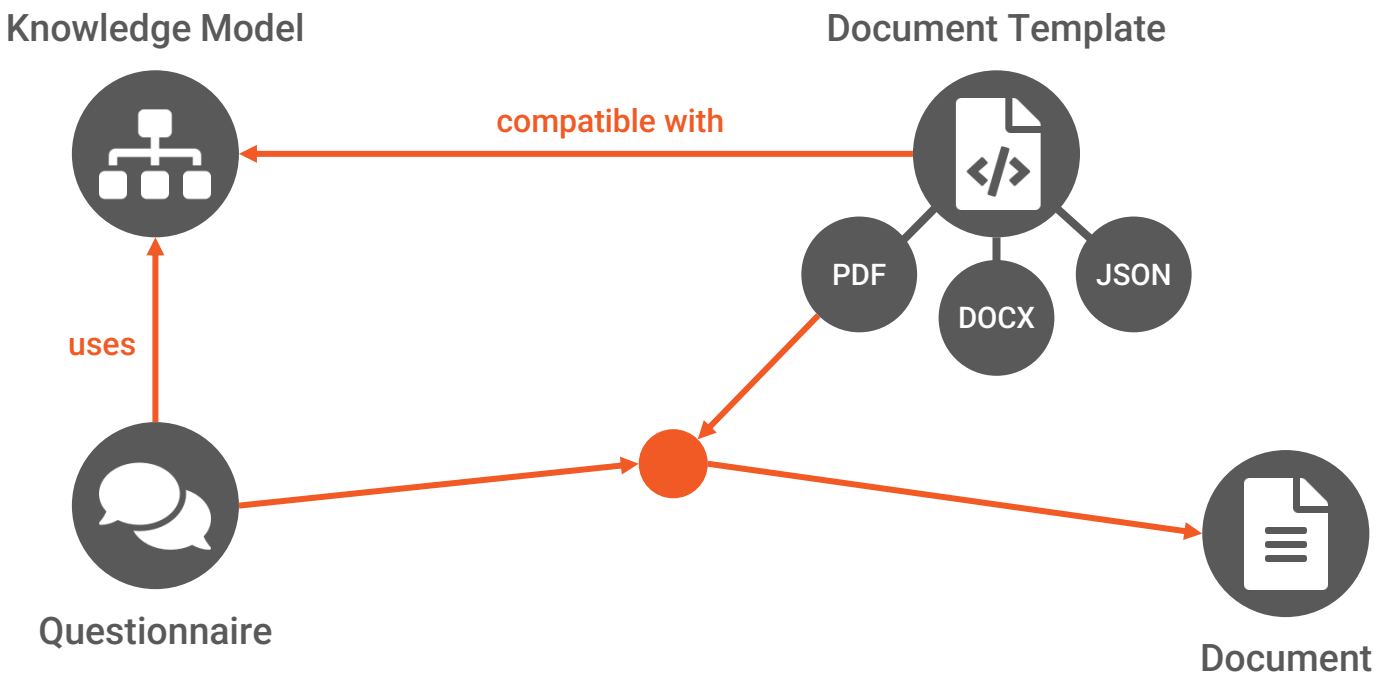
- Template development requires some **technical knowledge**
- Templates are composed of **JSON** metadata, **Jinja2** templates and other files
- **DSW Template Development Kit** (TDK)



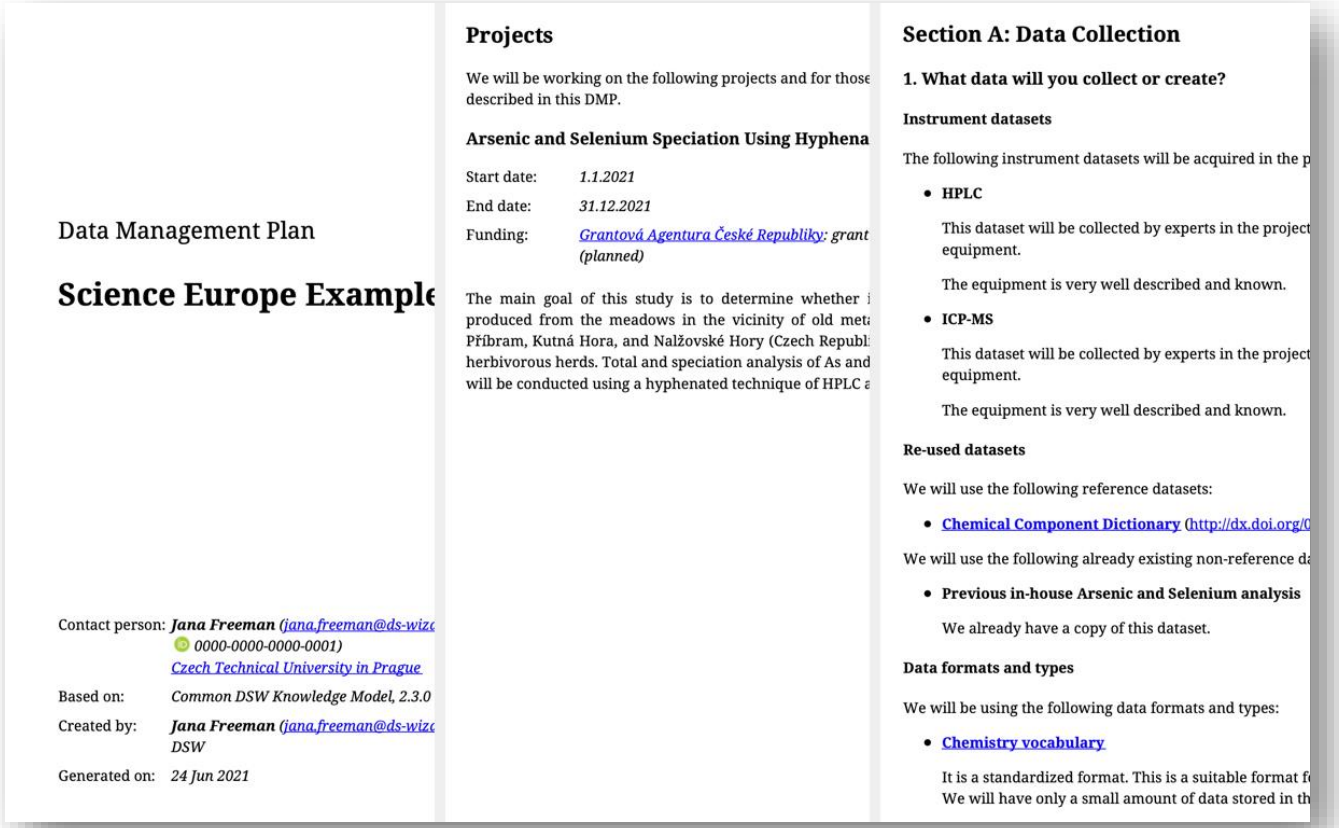
Data Stewardship Wizard



Data Stewardship Wizard



Generated Documents



Data Management Plan

Science Europe Example

Contact person: ***Jana Freeman*** *(jana.freeman@ds-wiz*
0000-0000-0000-0001)
Czech Technical University in Prague

Based on: *Common DSW Knowledge Model, 2.3.0*

Created by: ***Jana Freeman*** *(jana.freeman@ds-wiz*
DSW

Generated on: *24 Jun 2021*

Projects

We will be working on the following projects and for those described in this DMP.

Arsenic and Selenium Speciation Using Hyphena

Start date: *1.1.2021*

End date: *31.12.2021*

Funding: *Grantová Agentura České Republiky: grant (planned)*

The main goal of this study is to determine whether i produced from the meadows in the vicinity of old met Příbram, Kutná Hora, and Nalžovské Hory (Czech Republ herbivorous herds. Total and speciation analysis of As and will be conducted using a hyphenated technique of HPLC a

Section A: Data Collection

1. What data will you collect or create?

Instrument datasets

The following instrument datasets will be acquired in the p

- **HPLC**

  This dataset will be collected by experts in the project equipment.

  The equipment is very well described and known.

- **ICP-MS**

  This dataset will be collected by experts in the project equipment.

  The equipment is very well described and known.

Re-used datasets

We will use the following reference datasets:

- **Chemical Component Dictionary** (http://dx.doi.org/0

We will use the following already existing non-reference da

- **Previous in-house Arsenic and Selenium analysis**

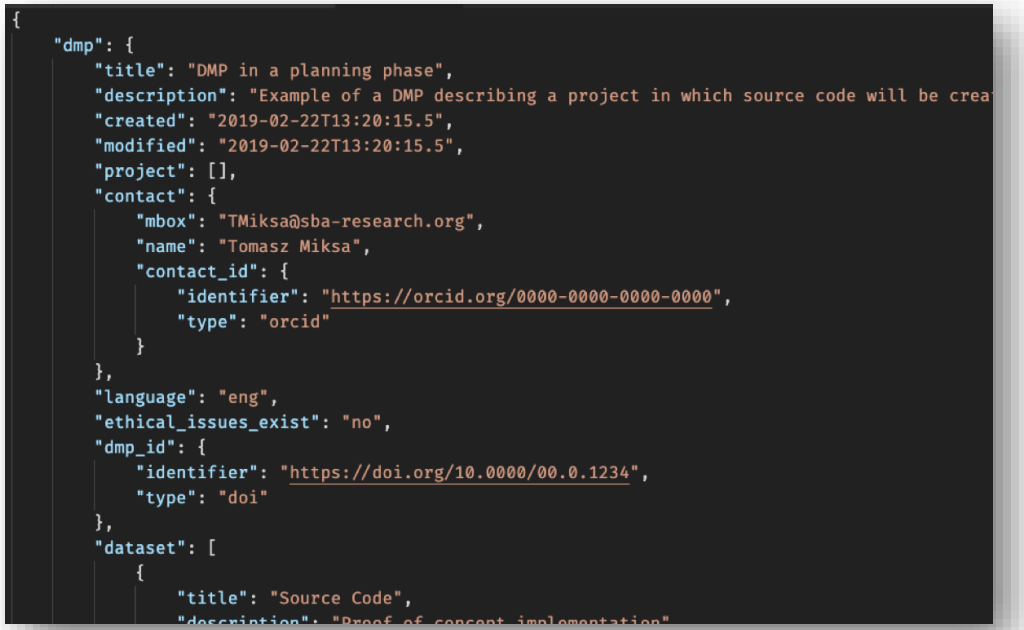
  We already have a copy of this dataset.

Data formats and types

We will be using the following data formats and types:

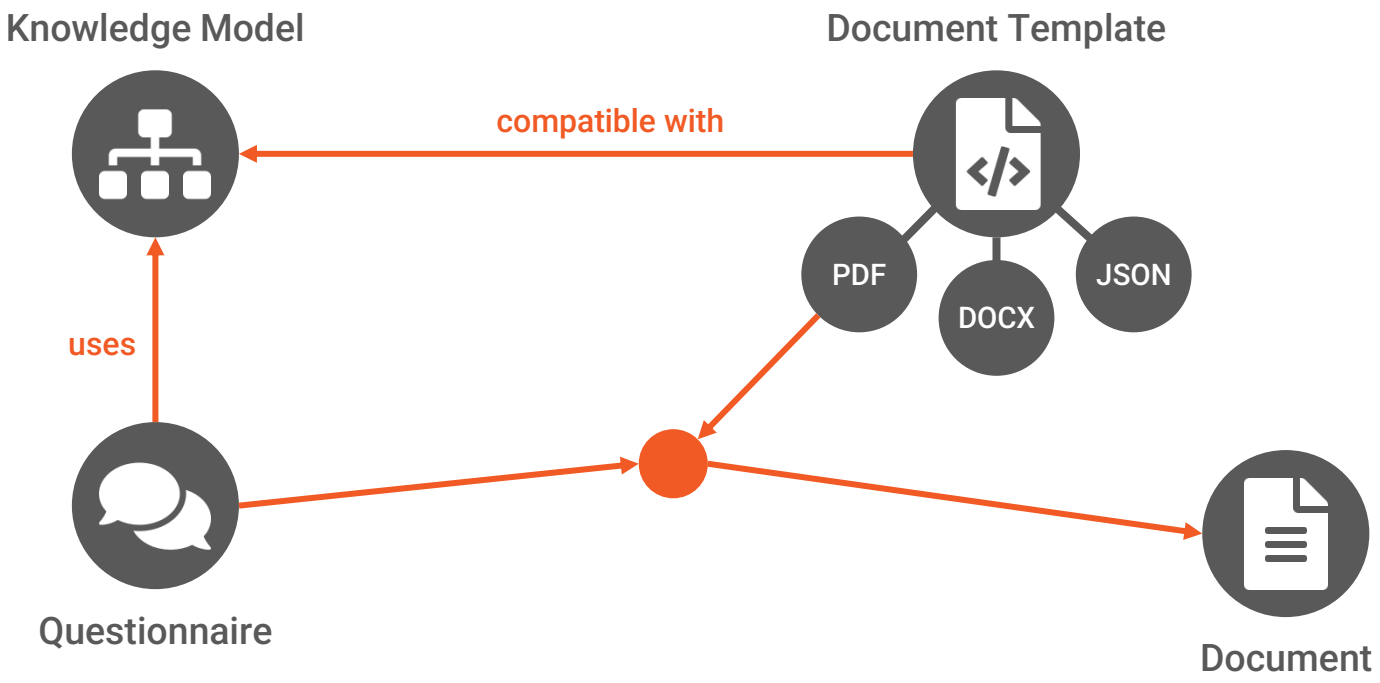
- **Chemistry vocabulary**

  It is a standardized format. This is a suitable format f
  We will have only a small amount of data stored in th



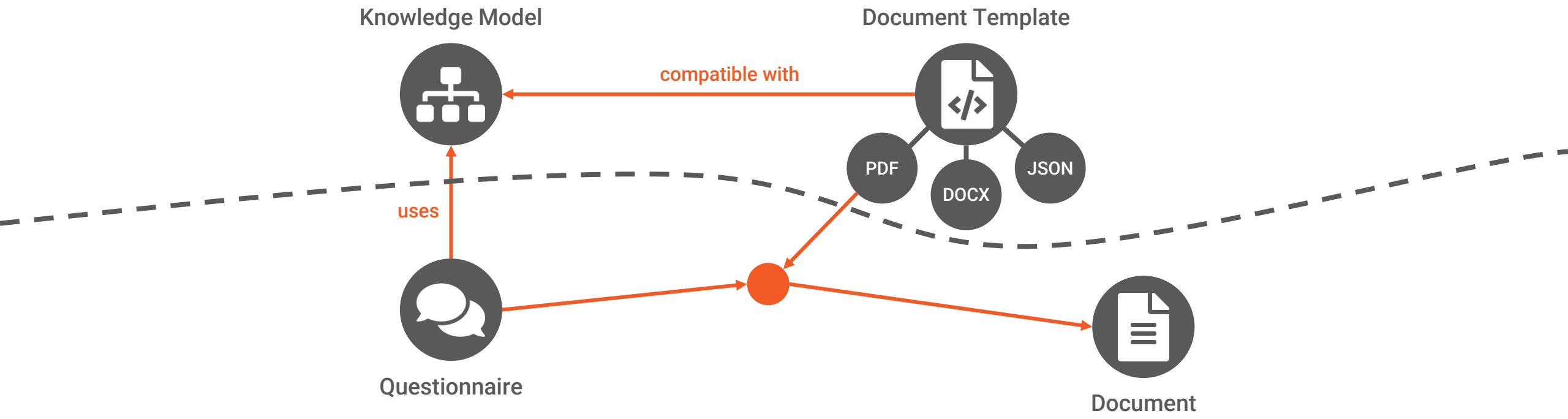

```
{
    "dmp": {
        "title": "DMP in a planning phase",
        "description": "Example of a DMP describing a project in which source code will be crea
        "created": "2019-02-22T13:20:15.5",
        "modified": "2019-02-22T13:20:15.5",
        "project": [],
        "contact": {
            "mbox": "TMiksa@sba-research.org",
            "name": "Tomasz Miksa",
            "contact_id": {
                "identifier": "https://orcid.org/0000-0000-0000-0000",
                "type": "orcid"
            }
        },
        "language": "eng",
        "ethical_issues_exist": "no",
        "dmp_id": {
            "identifier": "https://doi.org/10.0000/00.0.1234",
            "type": "doi"
        },
        "dataset": [
            {
                "title": "Source Code",
                "description": "Proof of concept implementation"
```

Data Stewardship Wizard



Data Stewardship Wizard

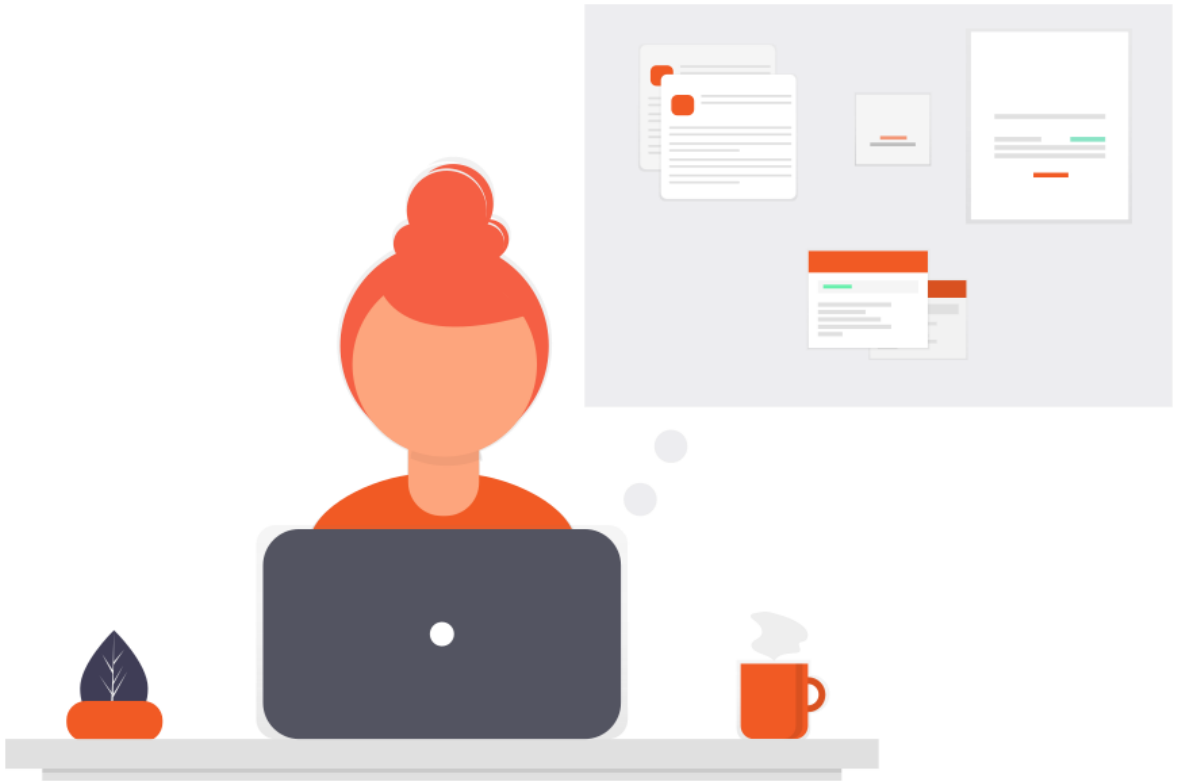
Data Steward



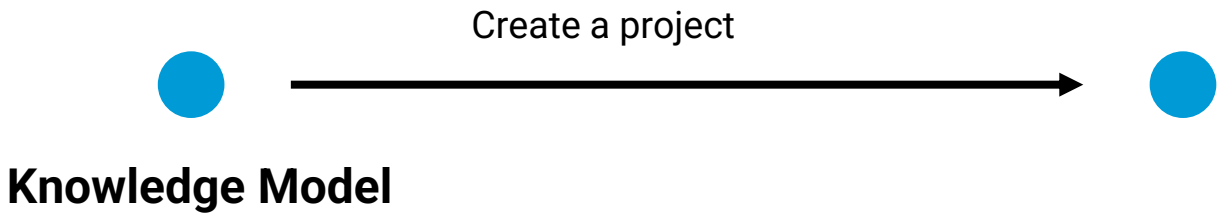
Researcher

Researchers

DSW

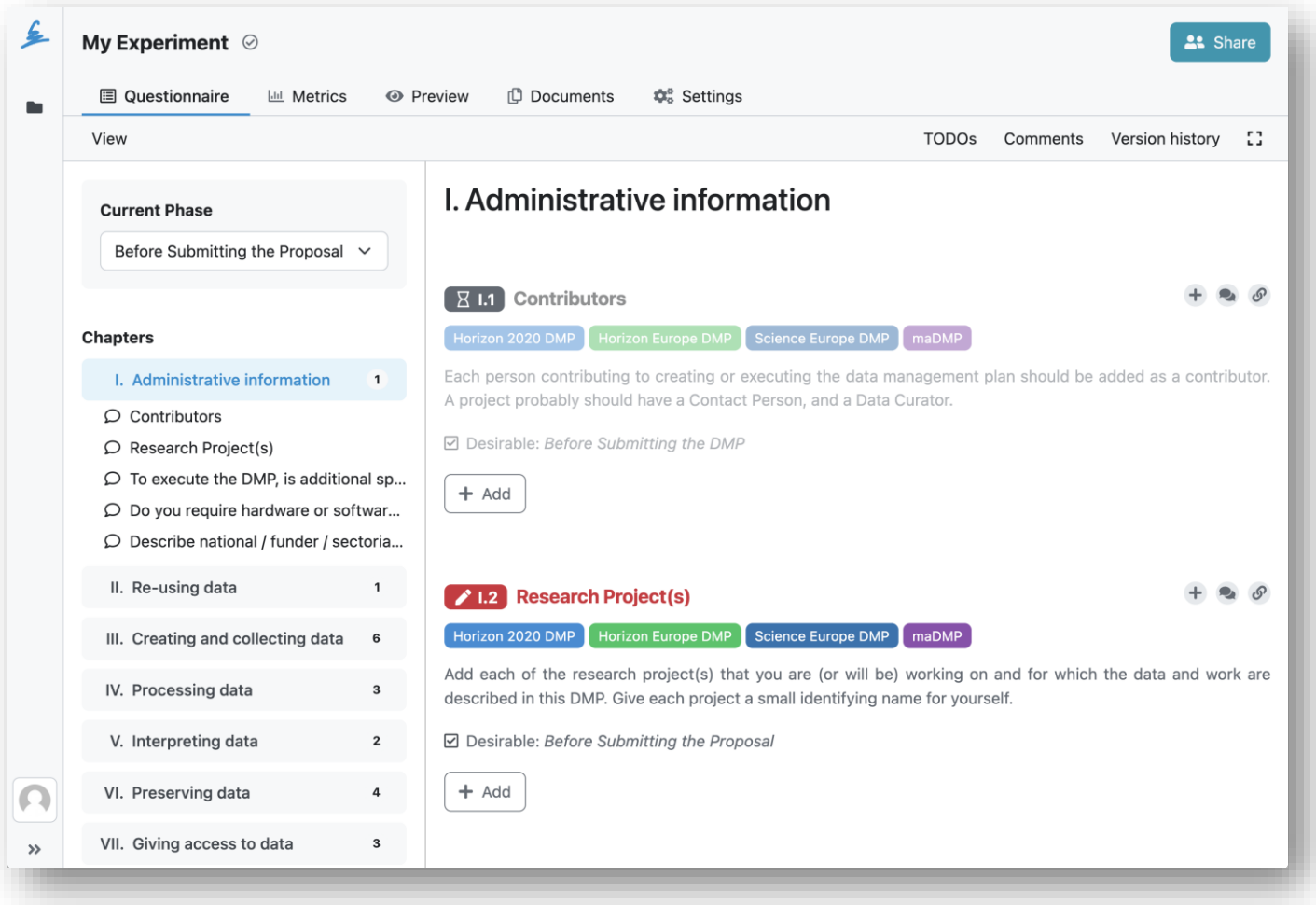


Researchers Workflow

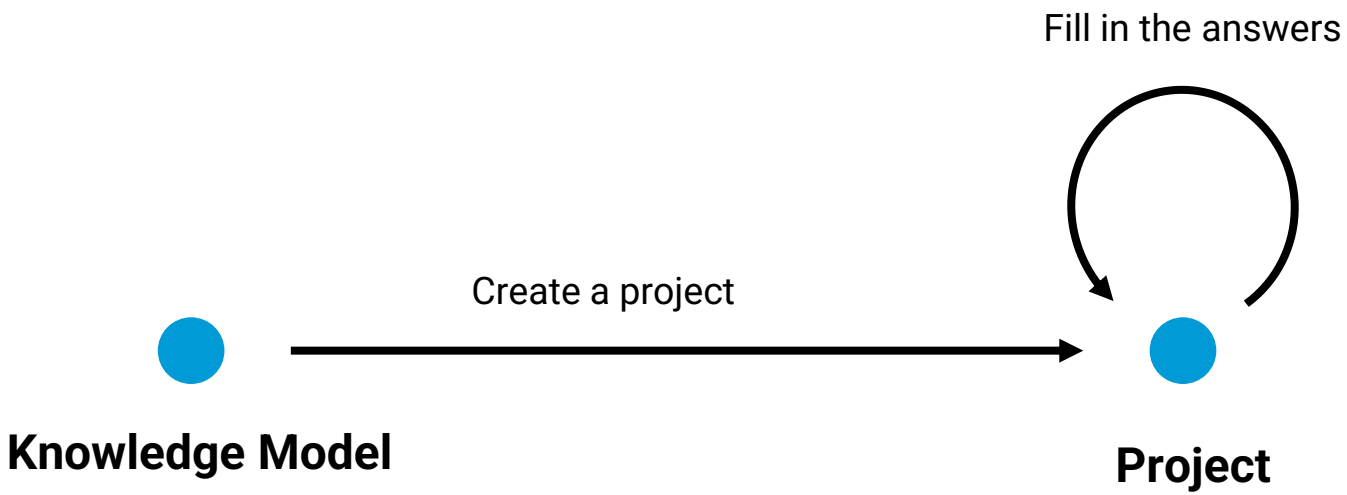


Project

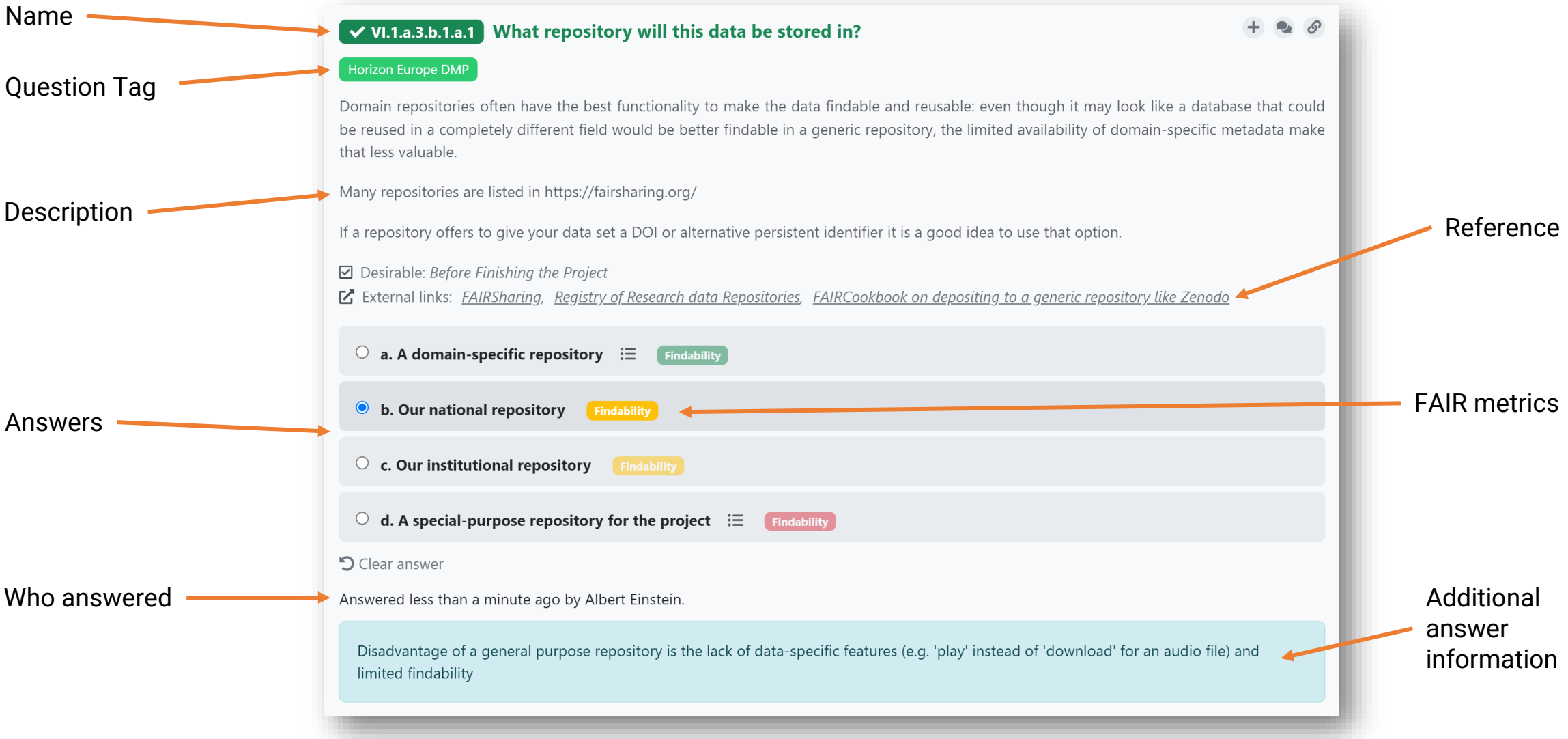
- Questionnaire
- Metrics
- Preview
- Documents
- Settings



Researchers Workflow



Question


Name
Question Tag
Description
Answers
Who answered
Reference
FAIR metrics
Additional answer information
VI.1.a.3.b.1.a.1
What repository will this data be stored in?
Horizon Europe DMP
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/
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, Registry of Research data Repositories, FAIRCookbook on depositing to a generic repository like Zenodo
a. A domain-specific repository
Findability
b. Our national repository
Findability
c. Our institutional repository
Findability
d. A special-purpose repository for the project
Findability
Clear answer
Answered less than a minute ago by Albert Einstein.
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

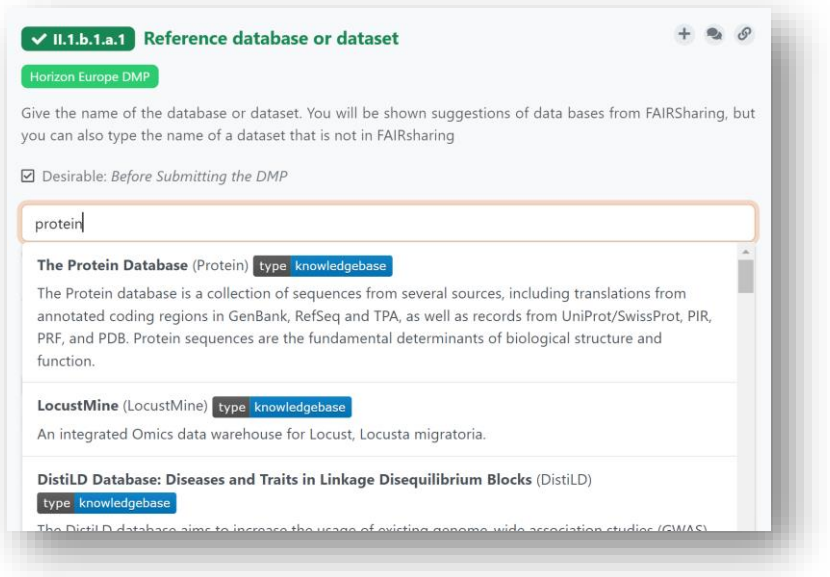
Question types

- Value
- Options
- Multi-Choice
- List of items
- Integration
- Item Select
- File

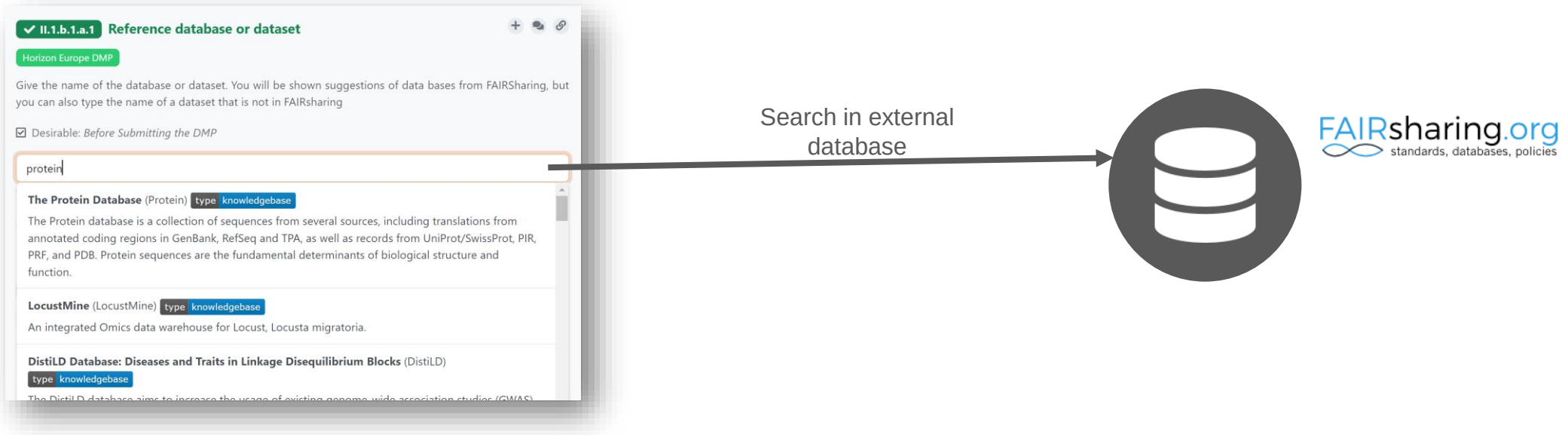
Integration question

- Connection to **external database** or service on the Knowledge Model level
- Answer is not just a text but also a **link to selected item**

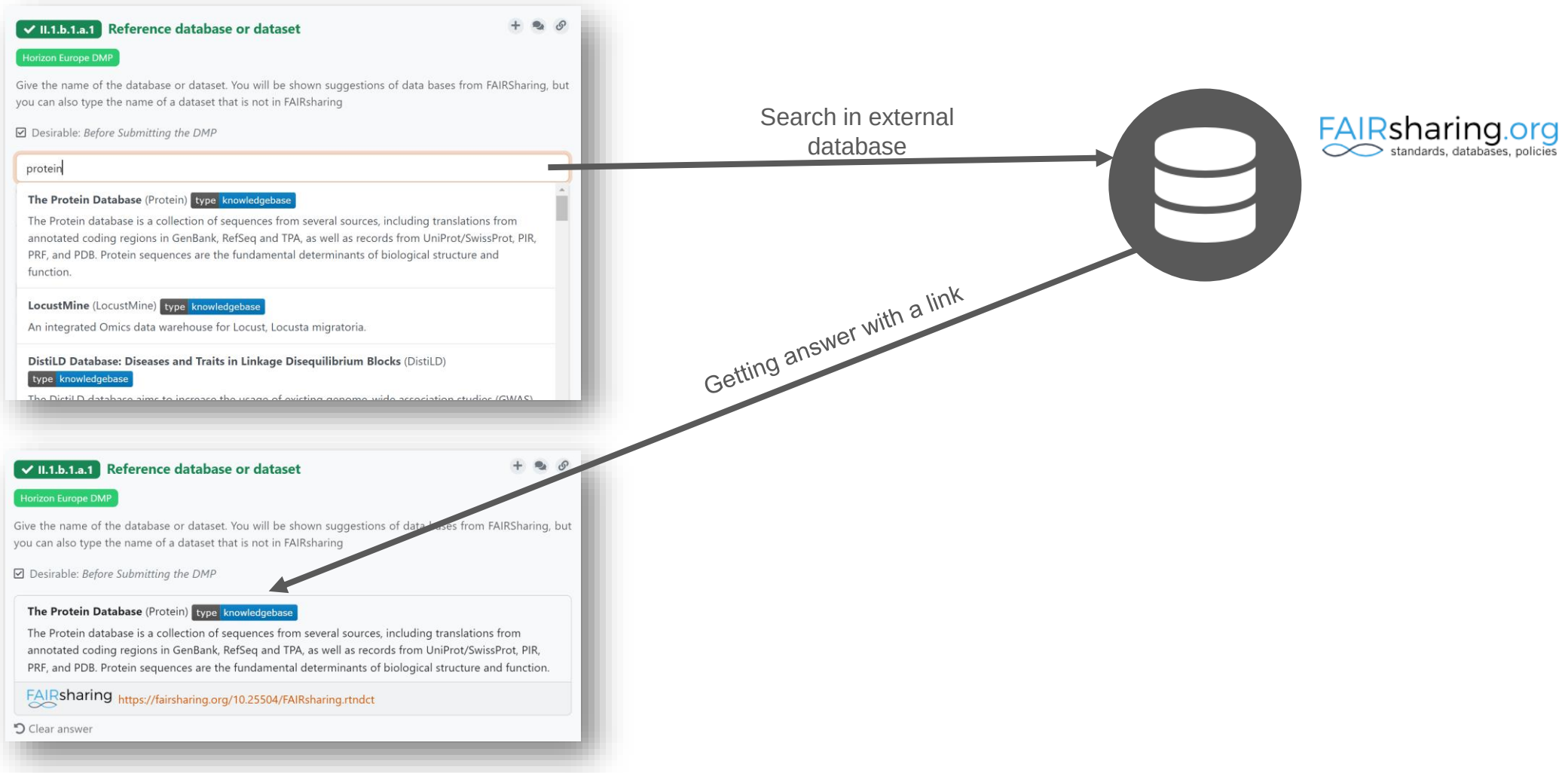
Integration question



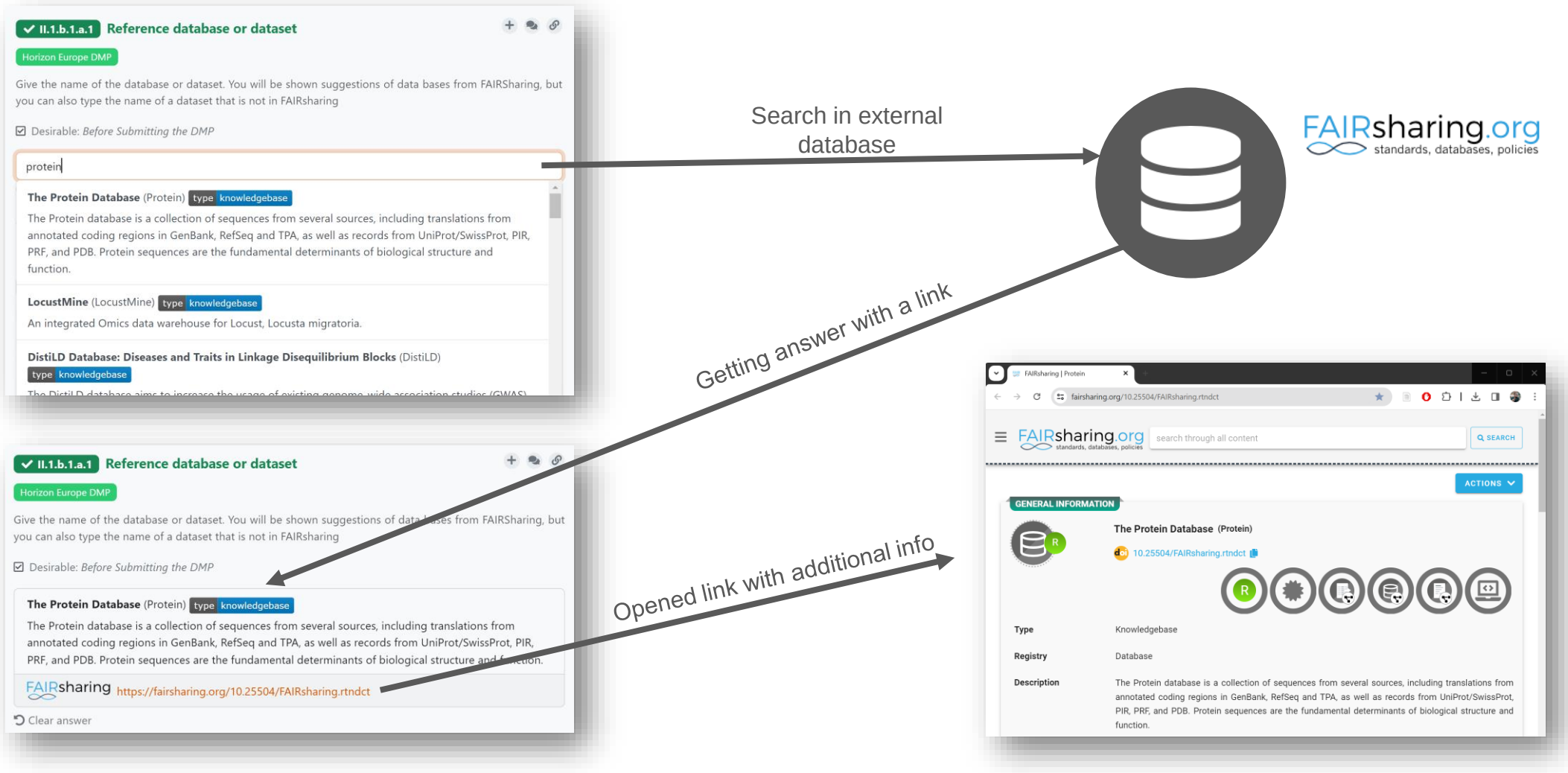
Integration question



Integration question

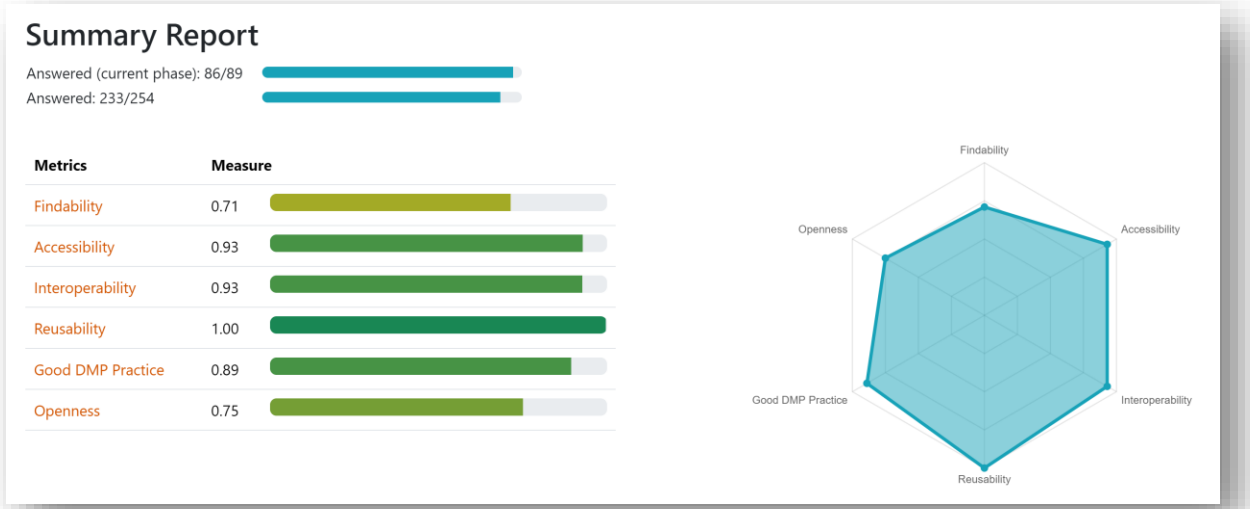


Integration question

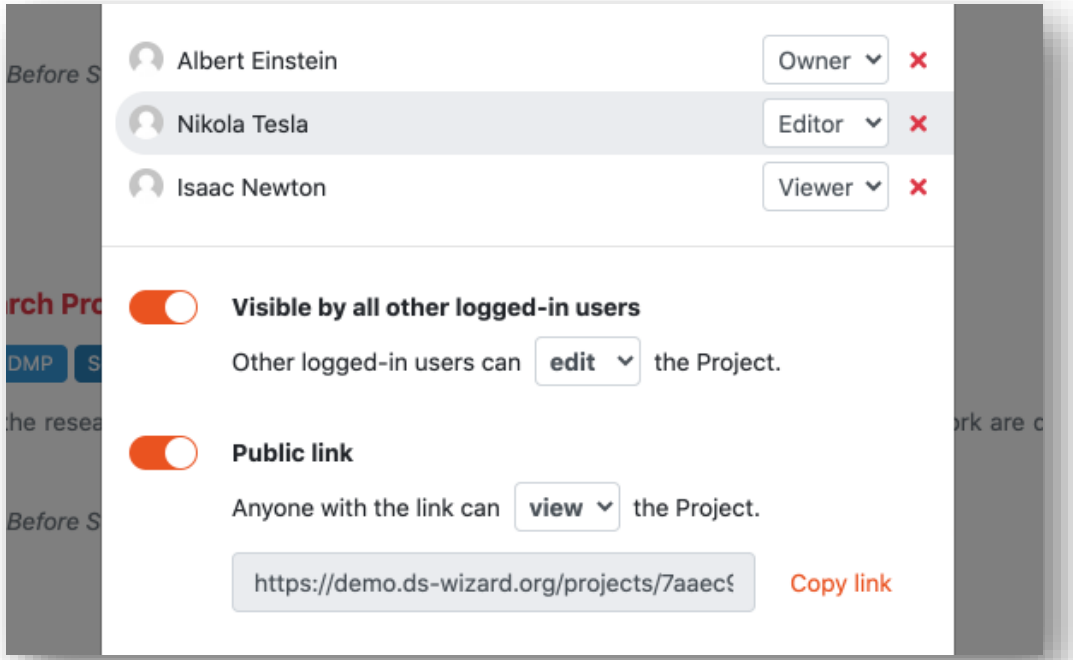


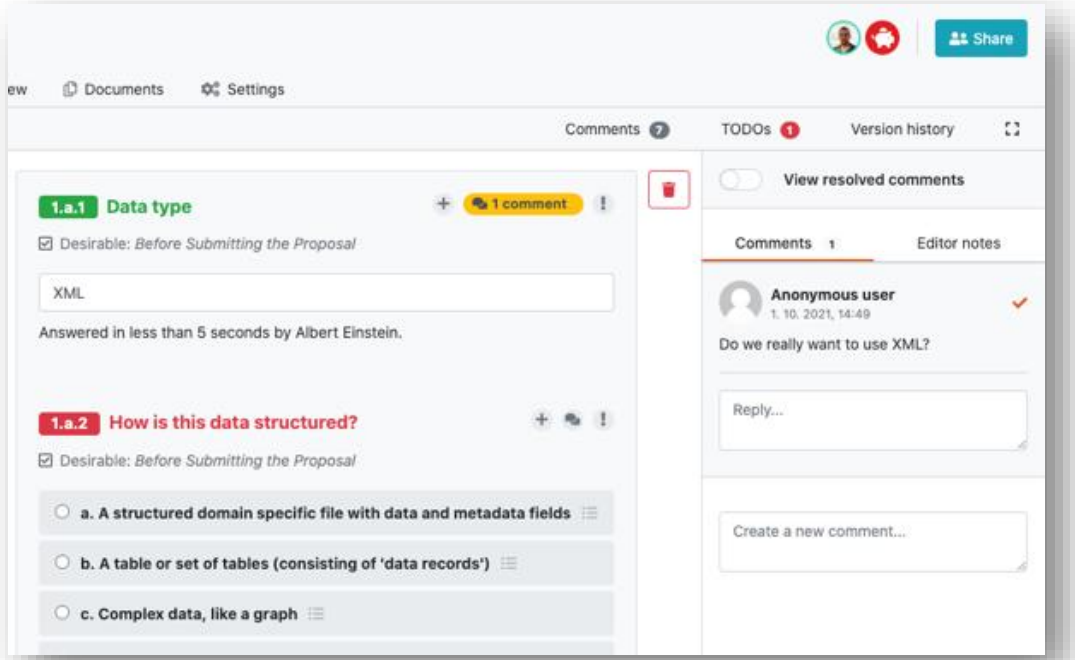
FAIR metrics

- Each answer can affect **metrics**
- The resulting value is calculated as a weighted average of all responses that affect the given metric



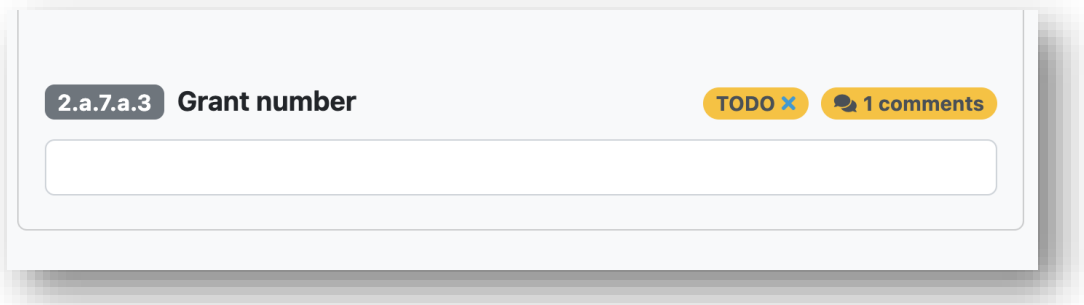
Online cooperation

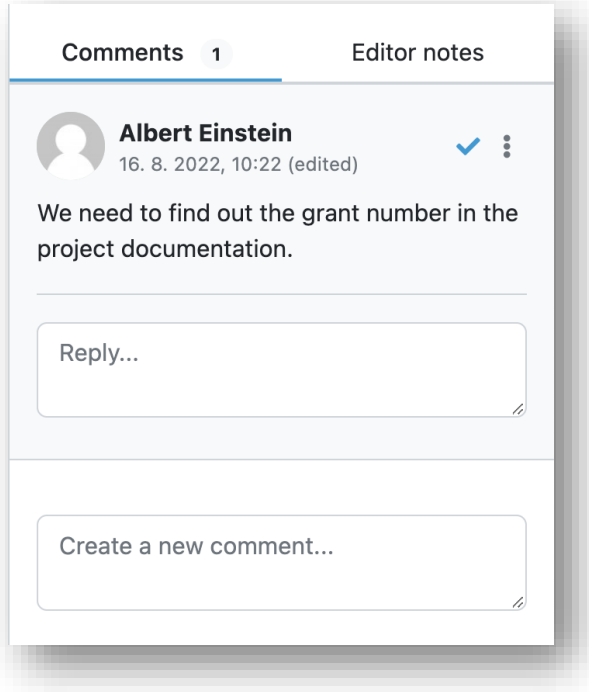




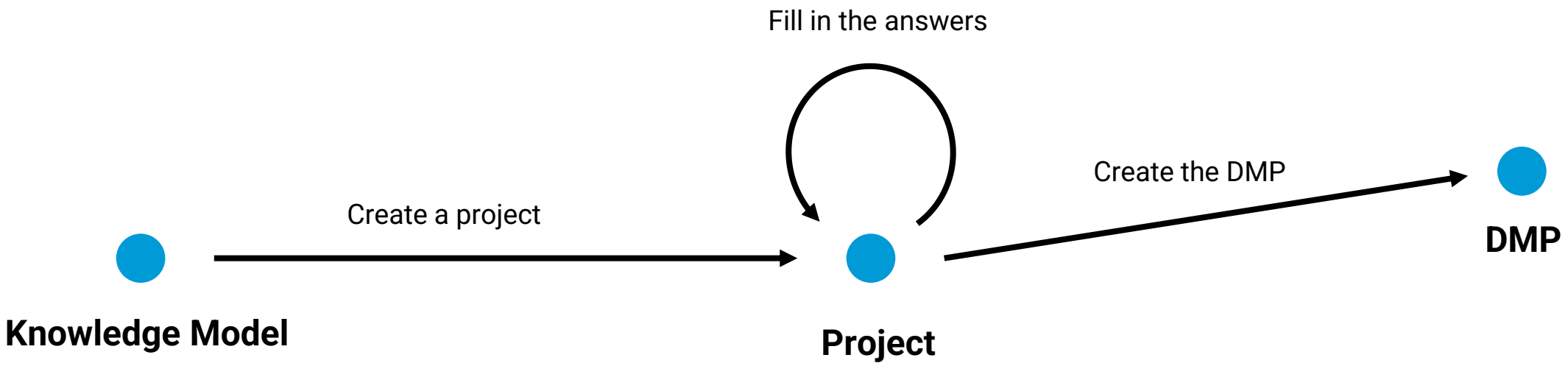
TODOs & Comments

- **TODOs** – when you're unsure how to answer a question, you can add TODO
- **Comments** – for discussion about the questions
  - Comments vs Editor notes



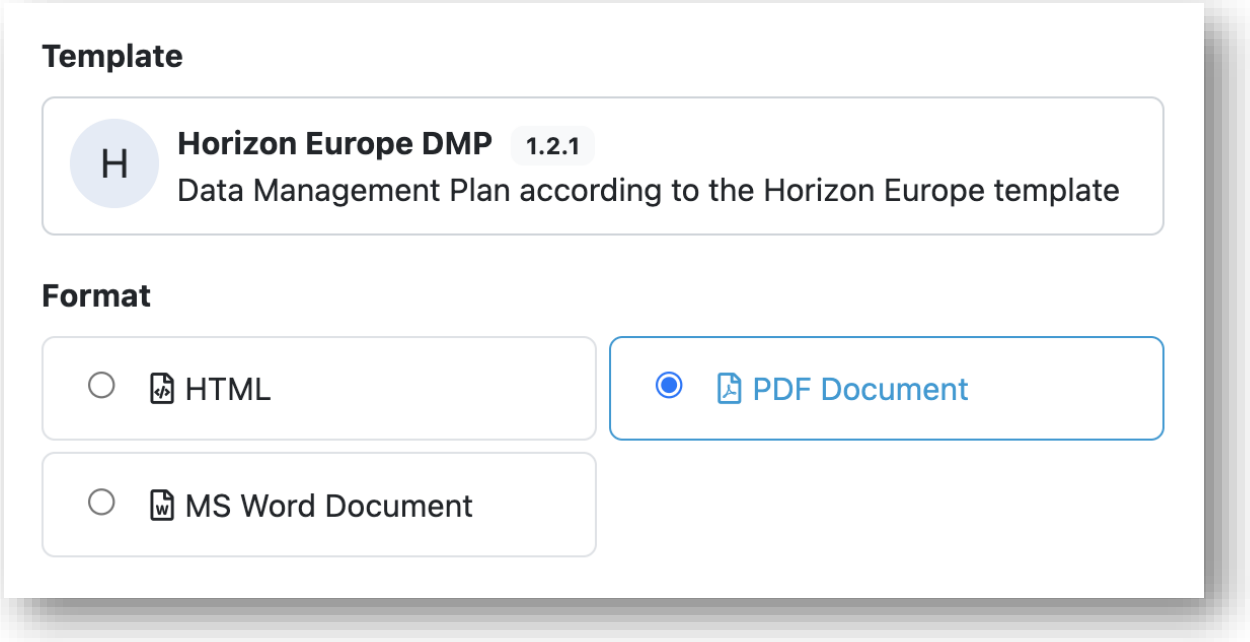


Researchers Workflow

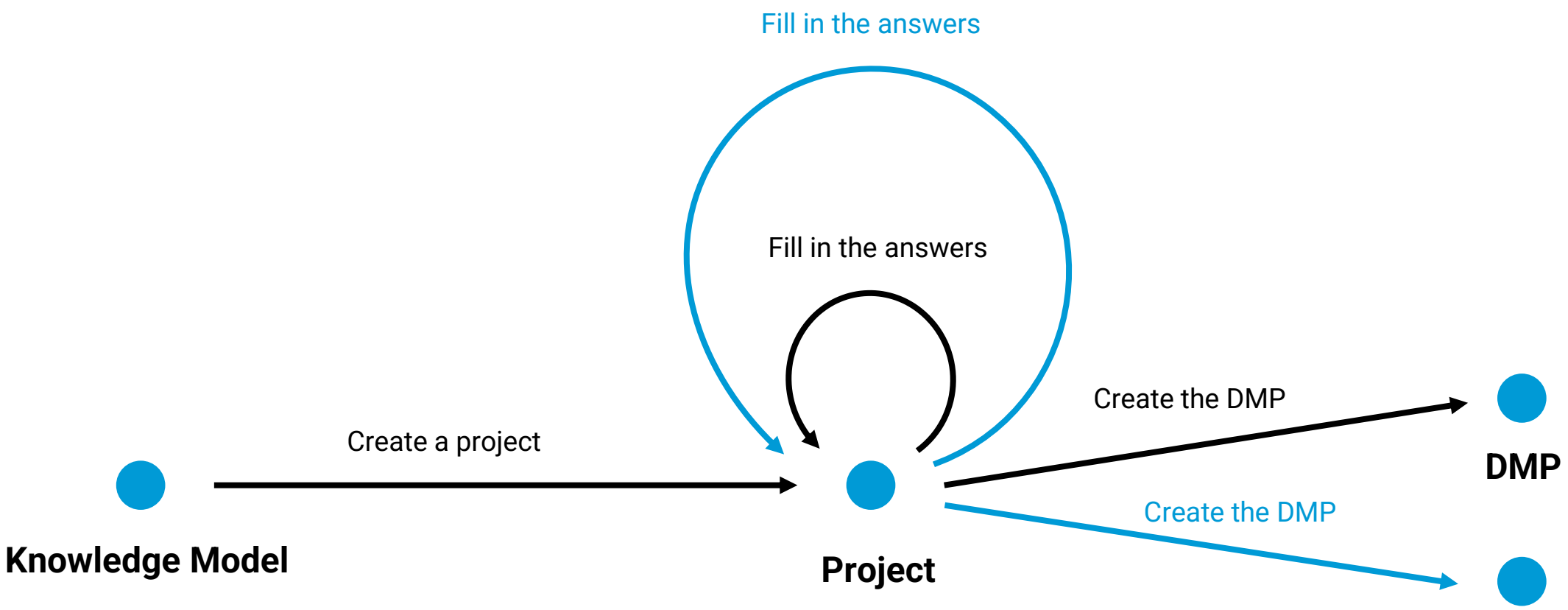


Creating Data Management Plan

- Created from a filled questionnaire using a document template
- Document formats based on template
- Saved in the DSW, can be downloaded

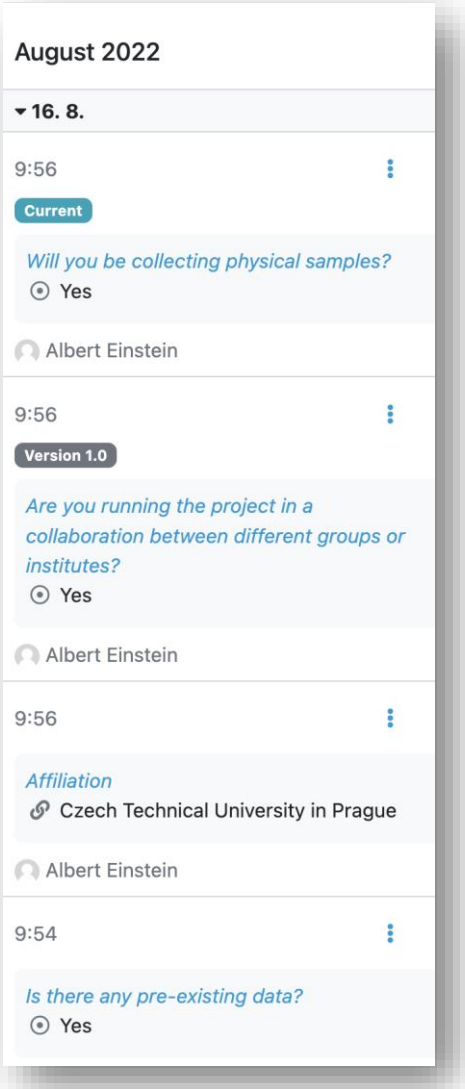


Researchers Workflow



Version History

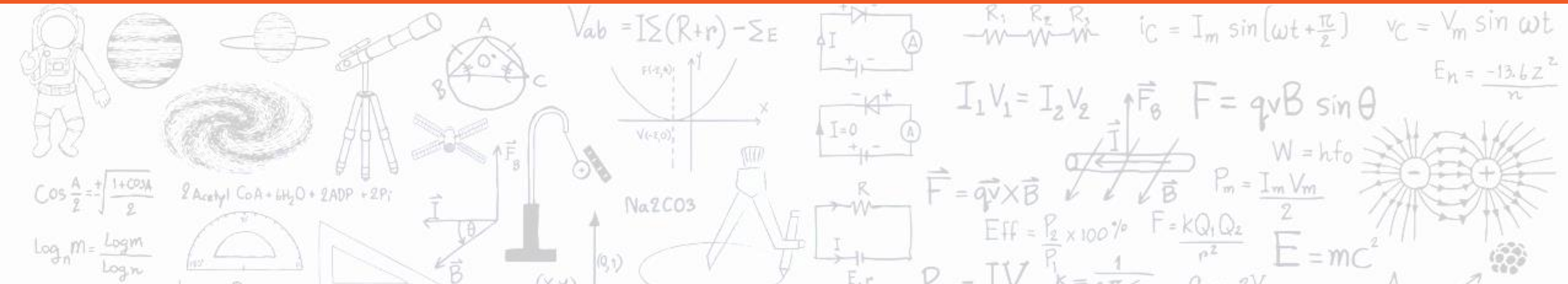
- List of all changes made while filling in the questionnaire
- Named versions
- Revert to version
- Create document from an older version



DSW

DATA STEWARDSHIP WIZARD

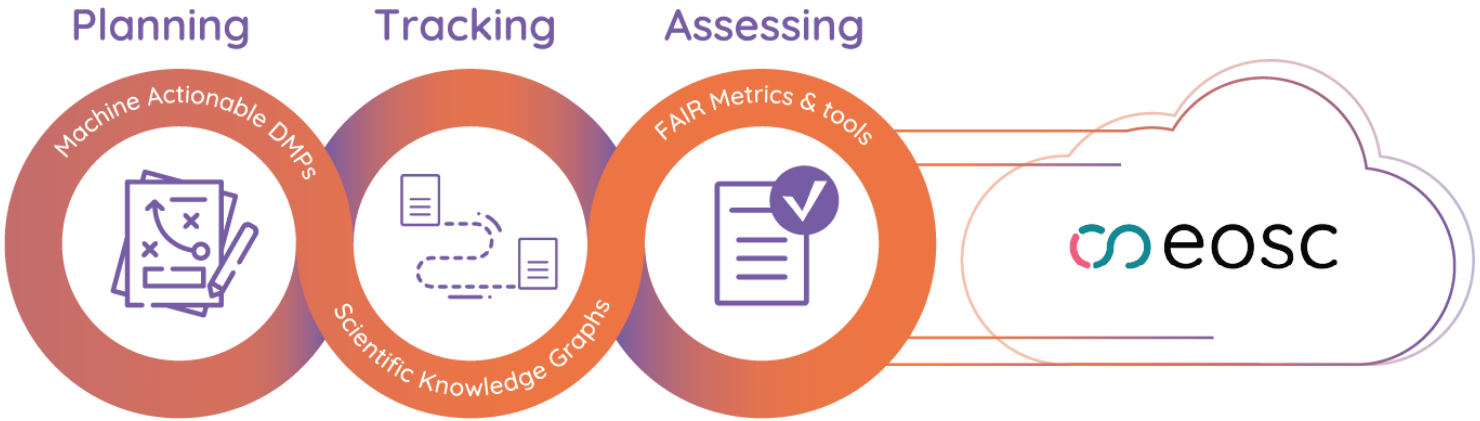
Possibilities for Institutions



Customizations and use of DSW

- ELIXIR infrastructure (provided by CESNET, managed by us):
  - ELIXIR is an intergovernmental organisation that brings together life science resources from across Europe.
  - researchers.dsw.elixir-europe.org - can be simply used by researchers, not customized
  - Institutional instance – can be customized by institution (admins), they manage it
- Self-hosted (on-premise / own cloud solution)
  - DSW is open-source, documentation and Docker images available
  - Institution can host their own instance, manage it, update it, keep the data in-house
- Commercial cloud (https://fair-wizard.com)
  - Enterprise-ready with additional services for institutions and commercial subjects

Acknowledgements



This presentation was supported by the EU-funded OSTrails project (Grant Agreement No 101130187)

More information on Data Stewardship Wizard support and funding is available at https://ds-wizard.org/#acknowledgements

DSW

DATA STEWARDSHIP WIZARD

Discussion and question