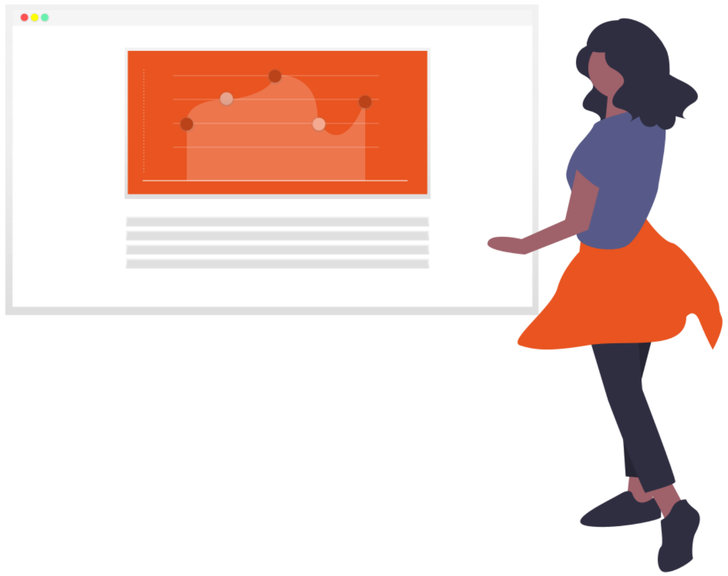
Data Stewardship Wizard Workshop

Jan Slifka <jan.slifka@ds-wizard.org>

24. 2. 2020

Outline

- Introduction
- DSW for Researchers + Demo
- DSW for Data Stewards + Demo
- How to Get Started
- Questions & Discussion

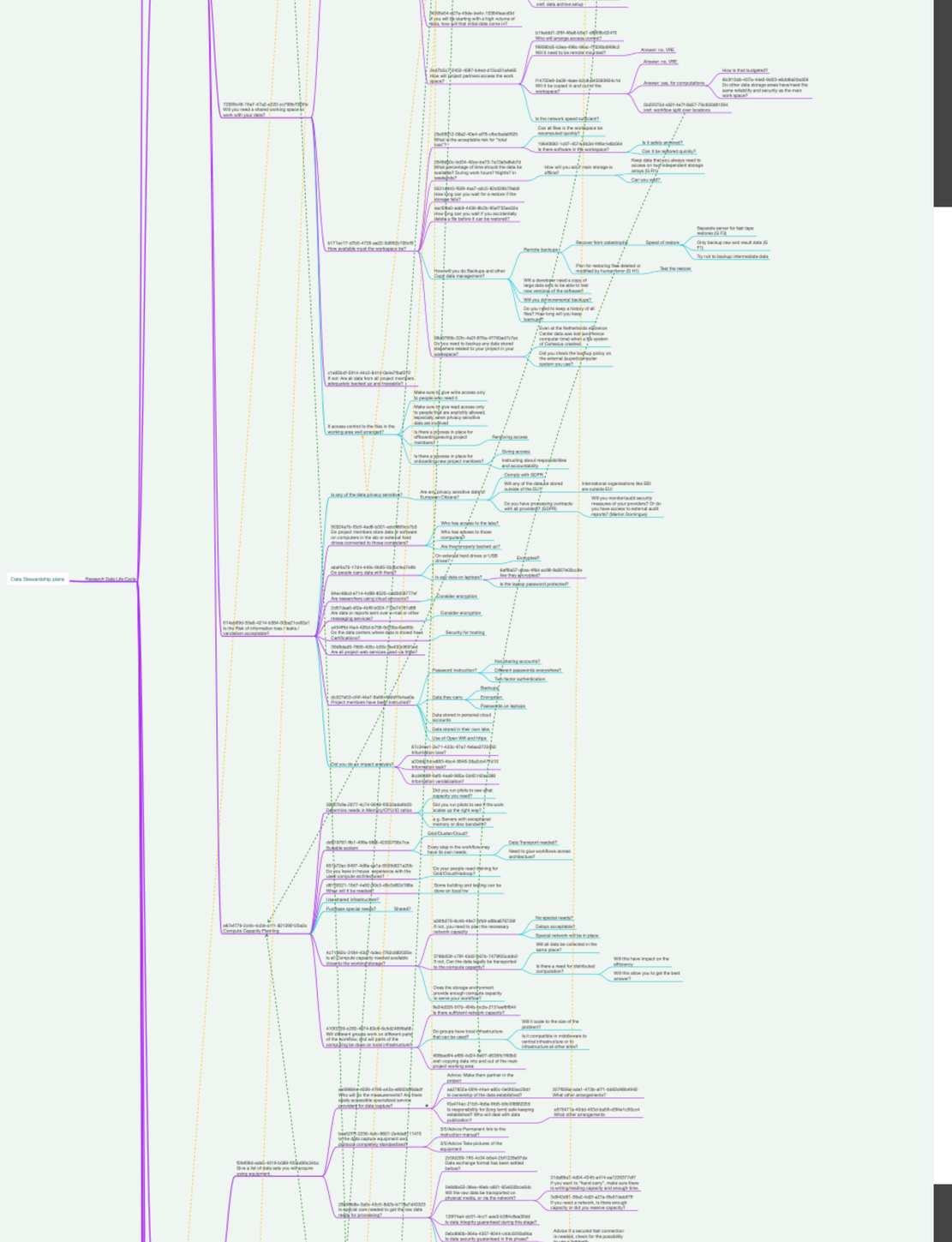


Introduction

Data Stewardship Mindmap

- A mind map of data stewardship topics
- Created by Rob Hooft

- https://doi.org/10.5281/zenodo.2614820



Data Stewardship Wizard

- A tool for generating data management plans
- Using smart questionnaires to guide researchers
- Customizable questionnaire templates

- Can be used in other domains

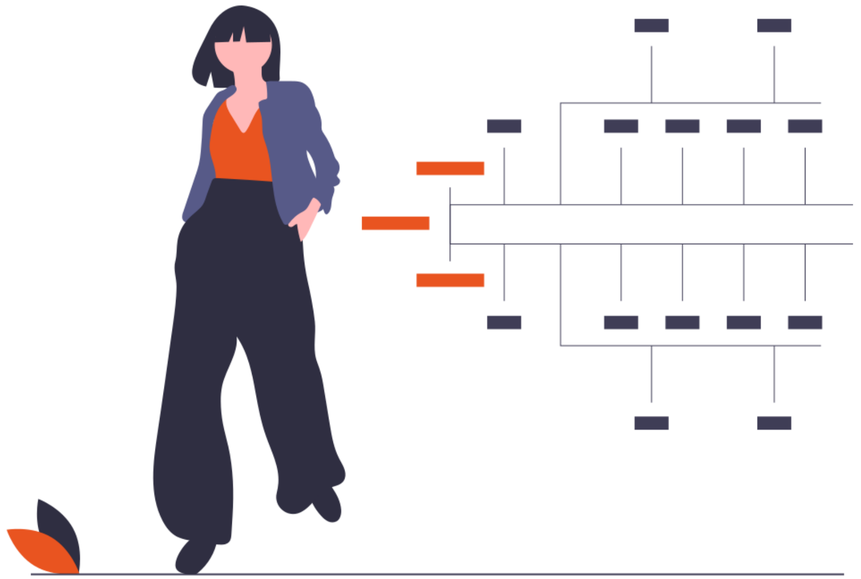
DSW for Researchers

Researchers Workflow

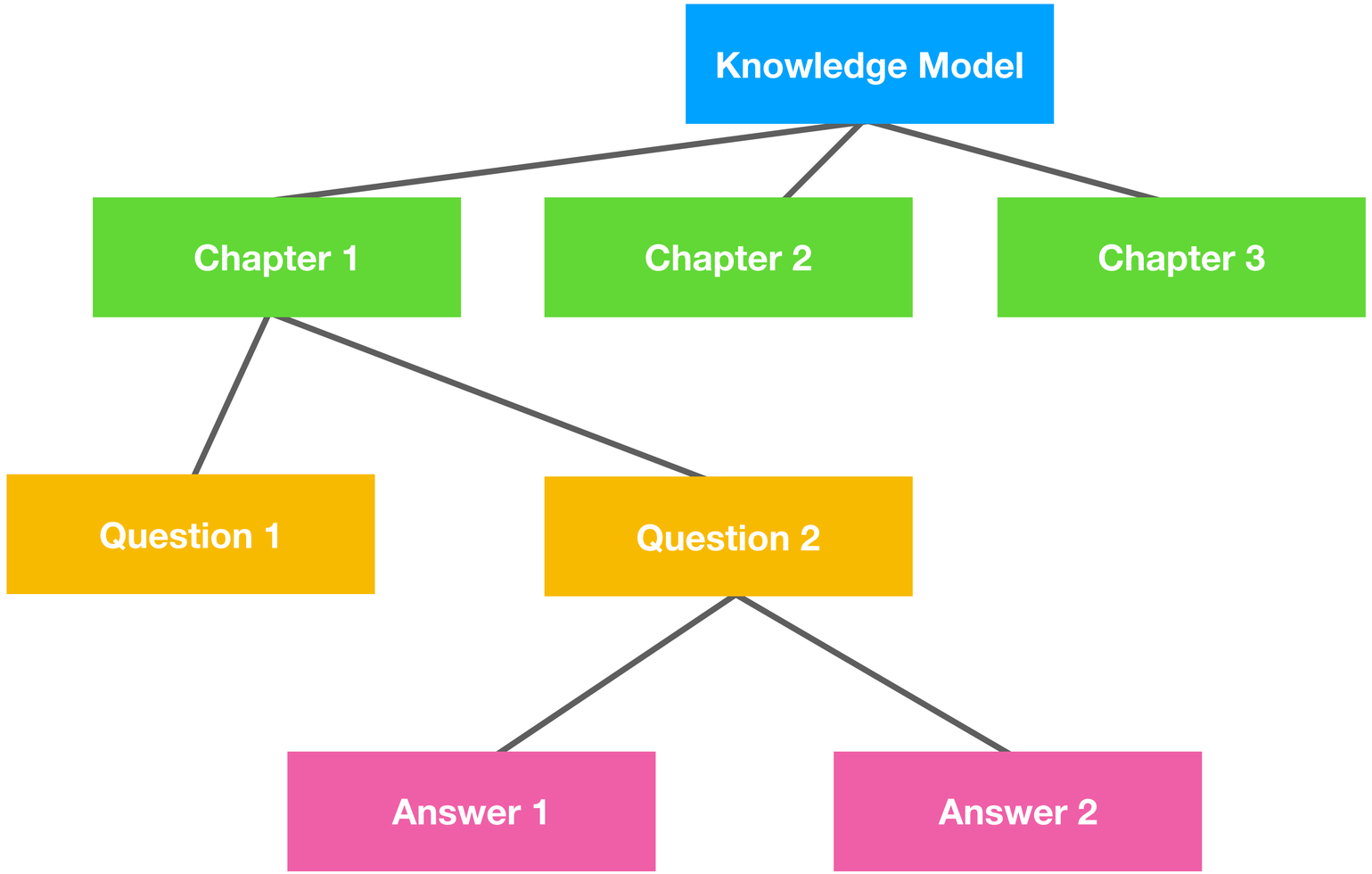
Knowledge Model

Knowledge Model

- Contains the **knowledge** about what should be asked and how
- **Tree-like structure** of Chapters, Questions, Answer and References or Experts
- “**Template**” for the Questionnaire
- Provided by DSW or institution data stewards



Knowledge Model

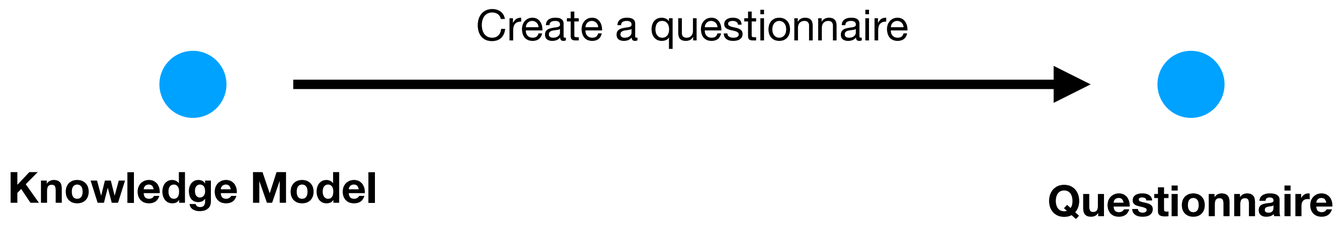


- Core DS Knowledge Model
  - Design of experiment
    - Is there any pre-existing data?
      - ☑ No
      - ☑ Yes
      - atq
    - Will reference data be created?
    - Will you be storing samples?
    - Will you be collecting experimental data?
  - Data design and planning
    - What data formats/types will you be using?
      - Data format/type:
      - Is this a standard data format used by others too?
      - Does this data format enable sharing and long term archiving?
        - ☑ No
        - ☑ Yes
      - njy
    - 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 stora
    - Will you need a shared working space to work with your data?
    - Is the risk of information loss, leaks and vandalism acceptably low?
    - Do you need to do compute capacity planning?
  - Data Capture/Measurement
  - Data processing and curation

Researchers Workflow

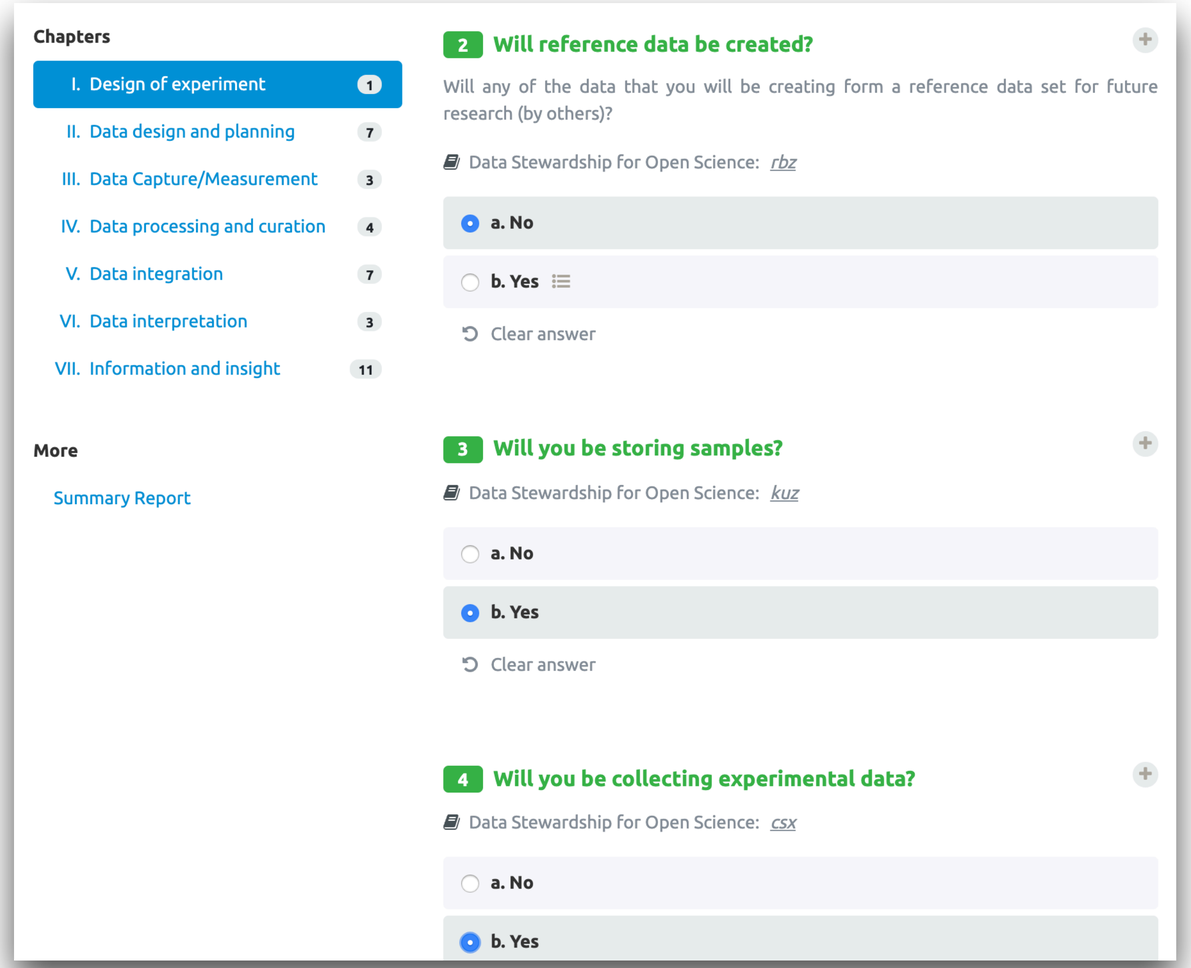
Knowledge Model

Researchers Workflow



Questionnaire

- Interactive form how to get the answers from users based on the Knowledge Model
- Contains report on the FAIR metrics
- Answers can be exported to a document (e.g., PDF)



Questionnaire Accessibility

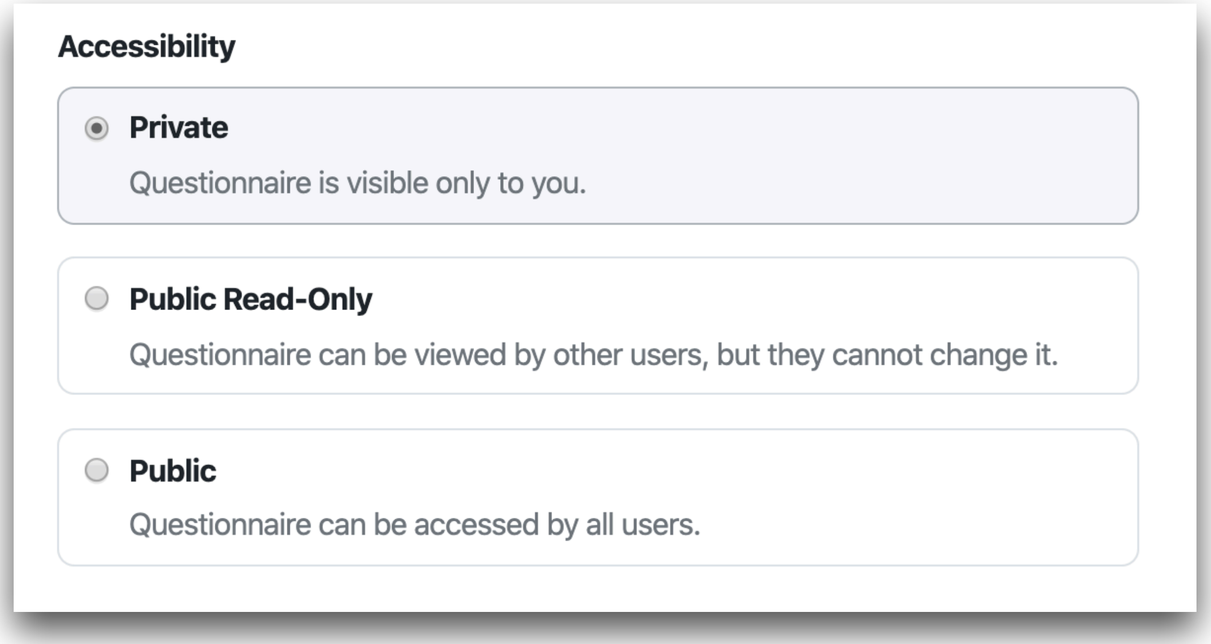
- Define who and how can access the questionnaire

Questionnaire Accessibility: Private

- Use when you don’t want to share it with anyone

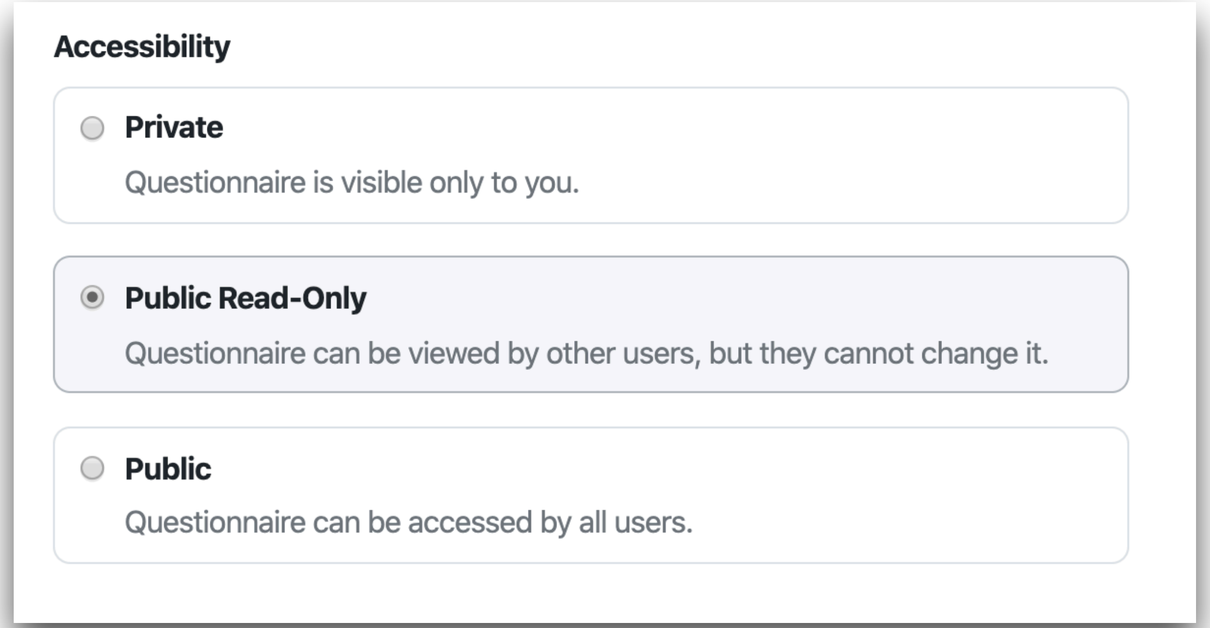
- You: View and Edit

- Other users: Nothing



Questionnaire Accessibility: Public Read-Only

- Use when you want to share your answers with other users
- Can be used for template questionnaires



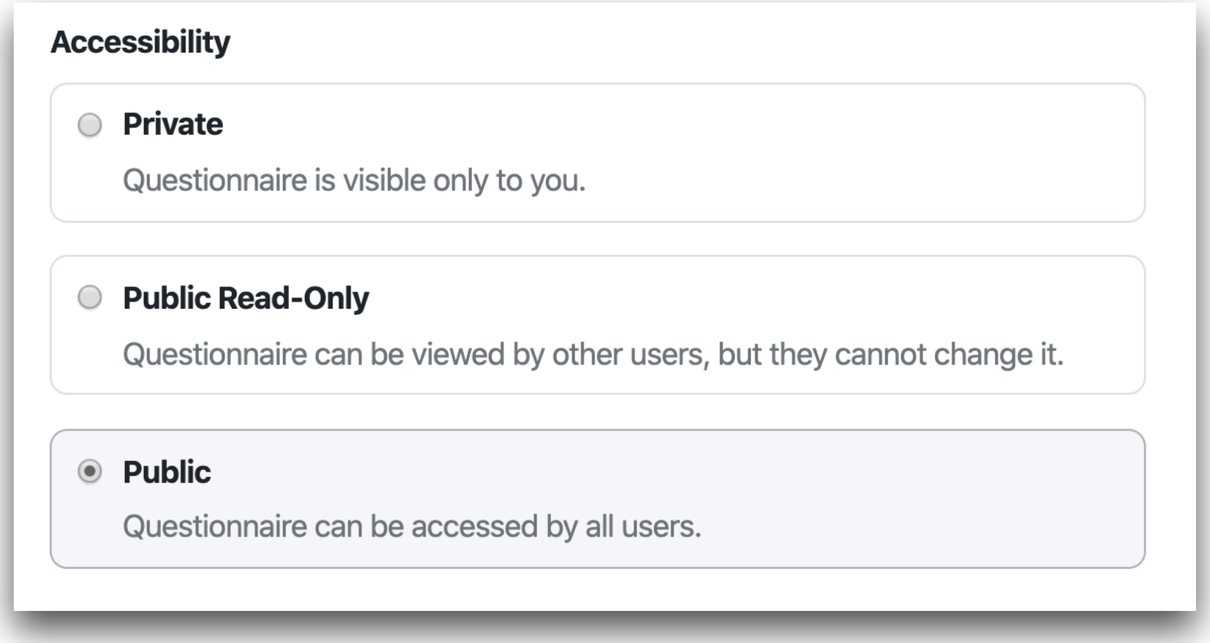
- You: View and Edit
- Other users: View

Questionnaire Accessibility: Public

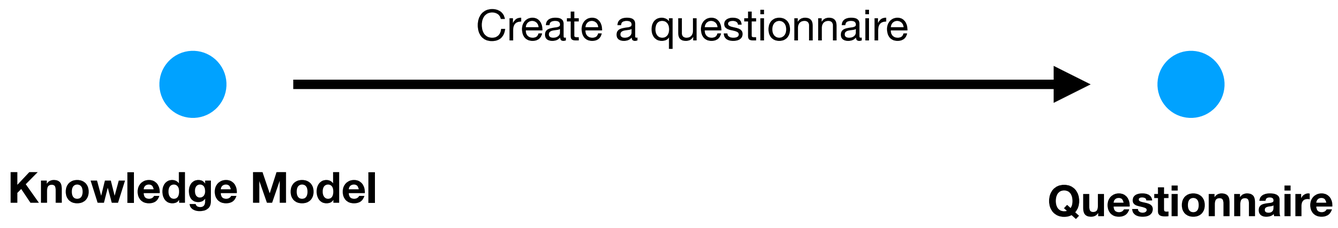
- Use when you want to collaborate with other users on the questionnaire

- You: View and Edit

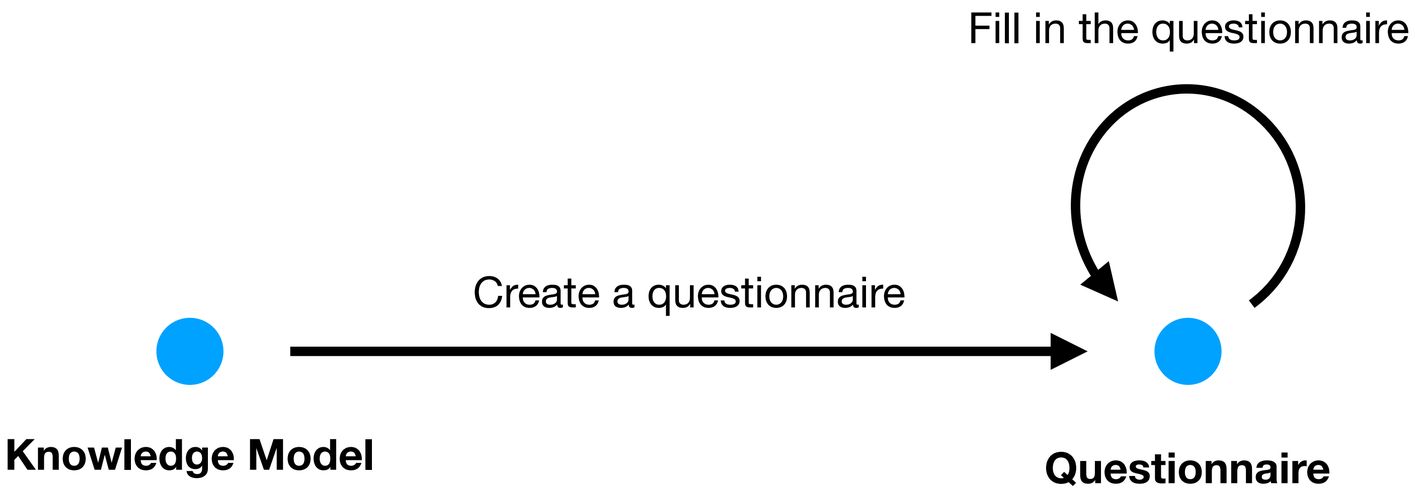
- Other users: View and Edit



Researchers Workflow

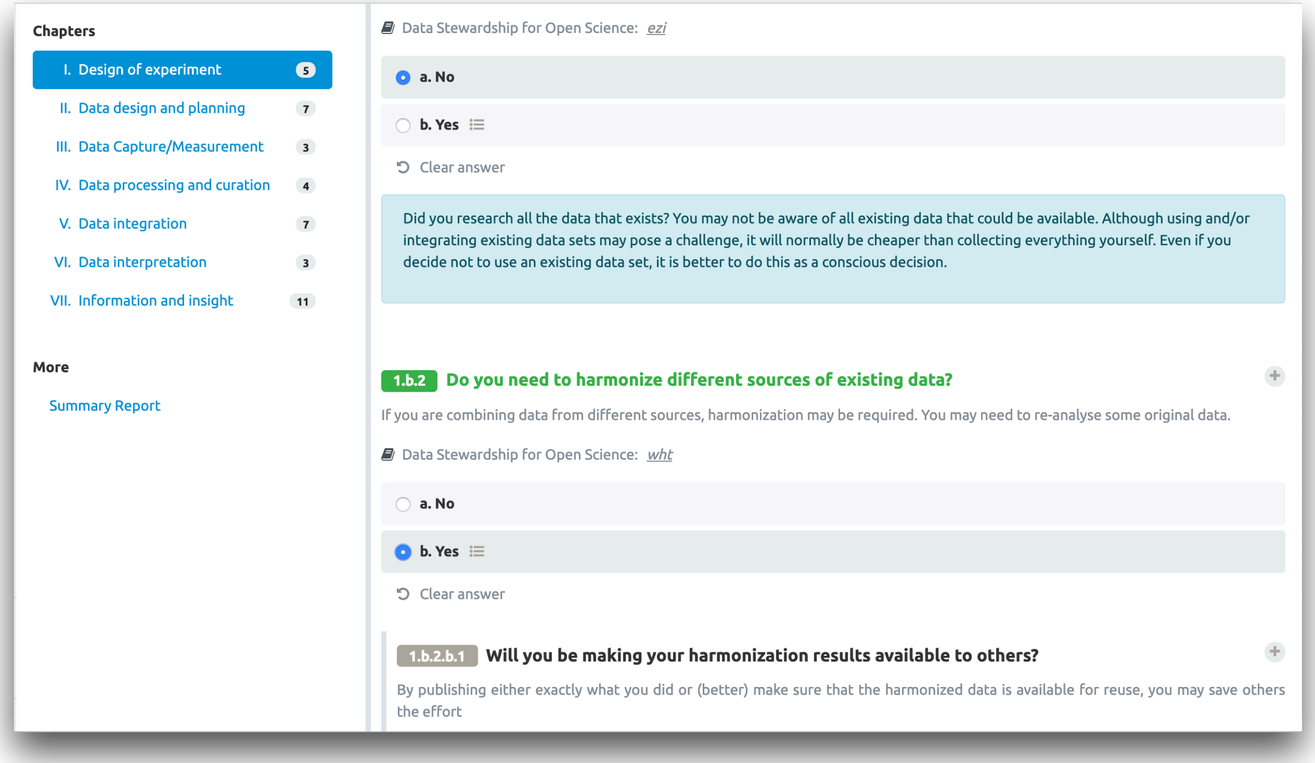


Researchers Workflow



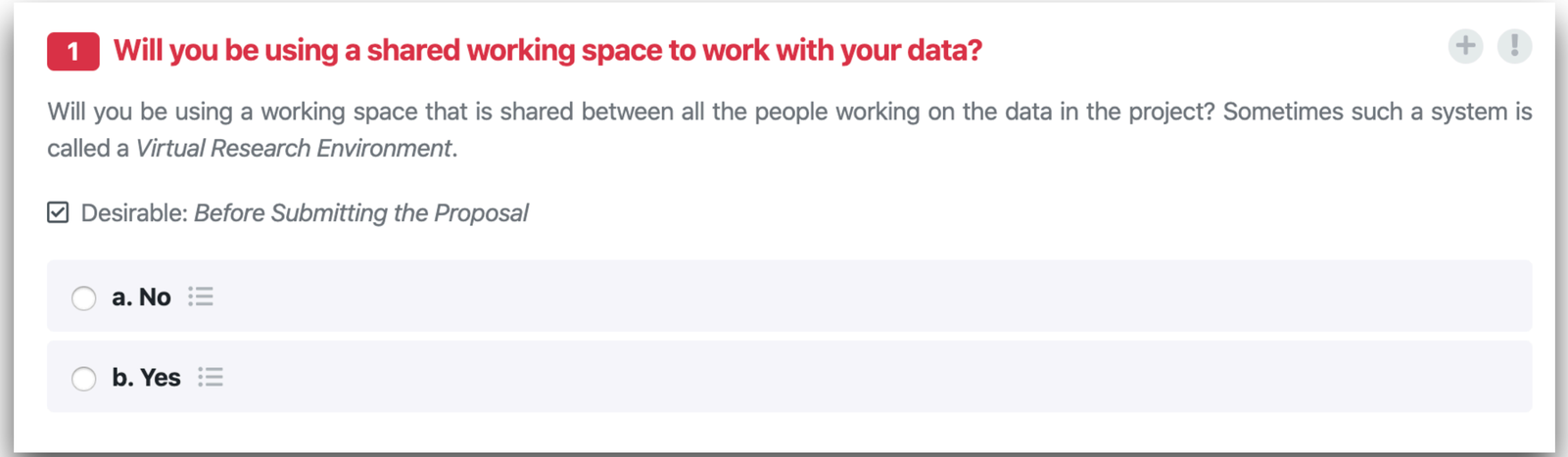
Filling the Questionnaire

- Free navigation through the Questionnaire (no given order of answering)
- Can be saved anytime

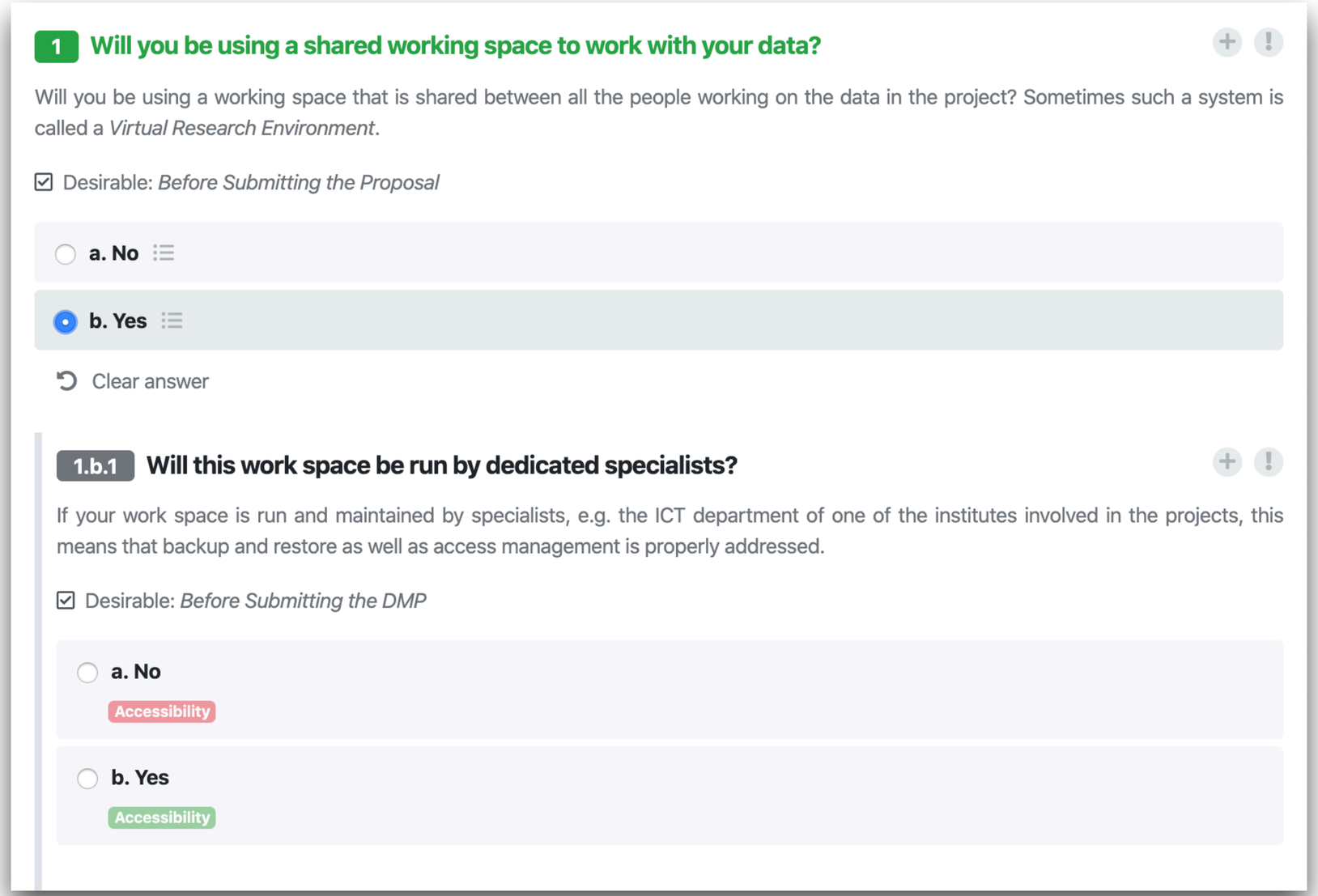


Filling the Questionnaire

- Only relevant questions are asked based on previous answers



Filling the Questionnaire



Data Stewardship for Open Science

- By Barend Mons

- References from questions in the Common DSW Knowledge Model

Data Stewardship for Open Science: Chapter 1.1

With kind permission of CRC Press Taylor & Francis Group

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 'reuse' data that had been created earlier, in all kinds of surveys and increasingly of course from sources such 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 format, 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.

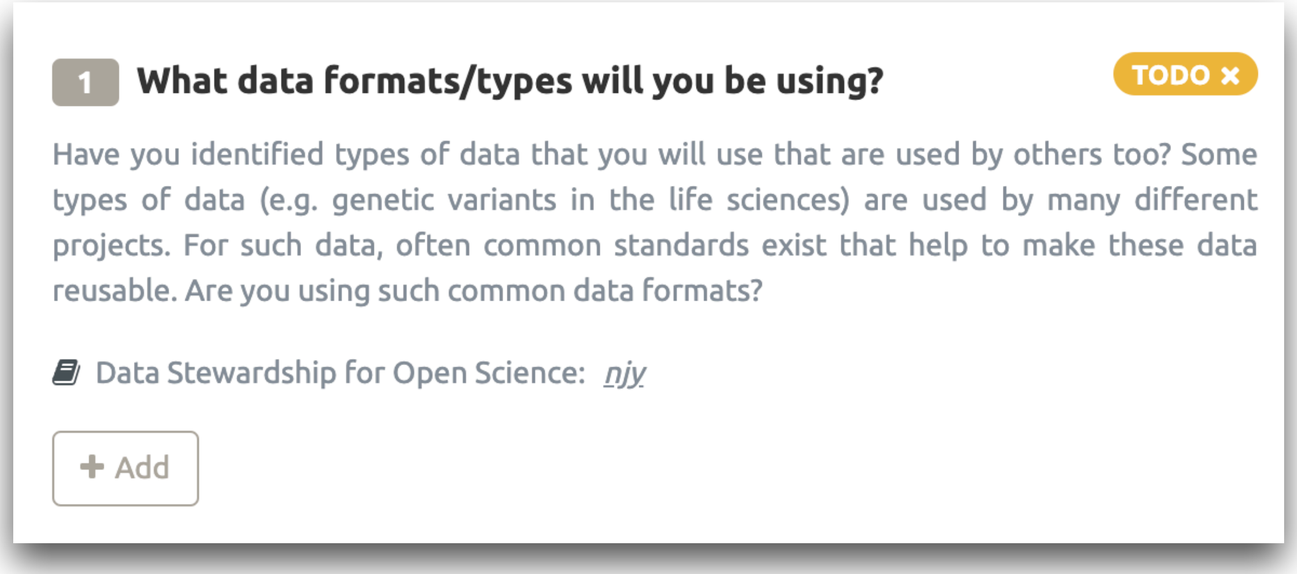
Do

- 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.

Questionnaire TODOs

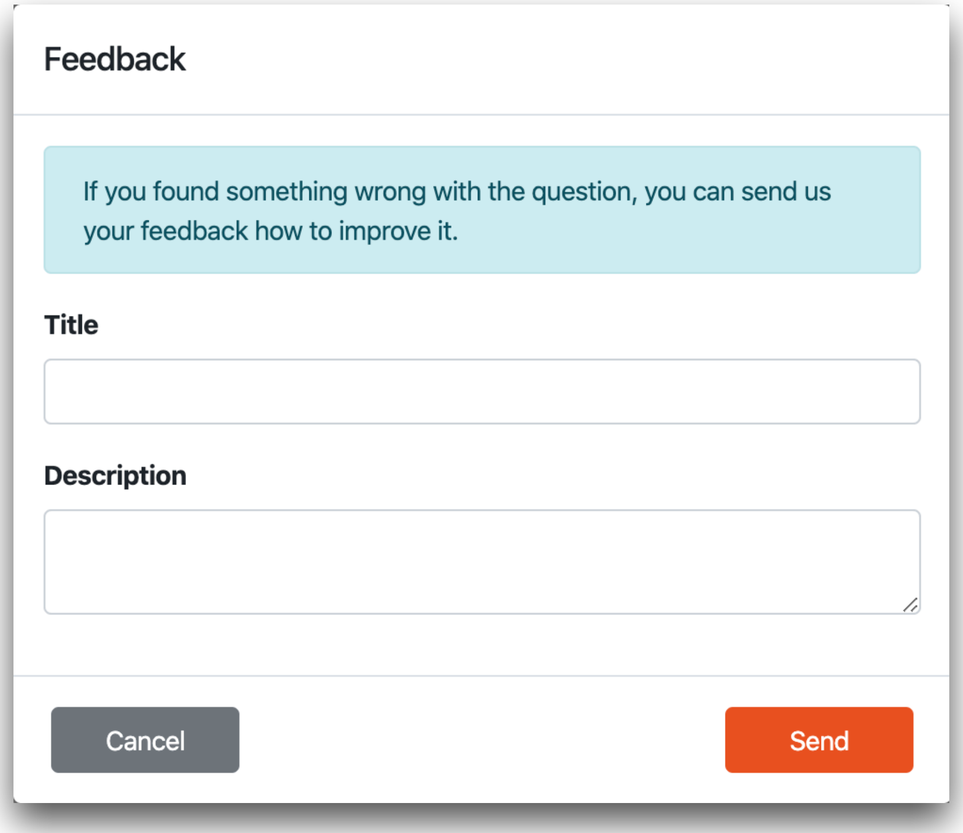
- When you are unsure how to answer a question, you can add TODO and come back later

- List of TODOs



Feedback

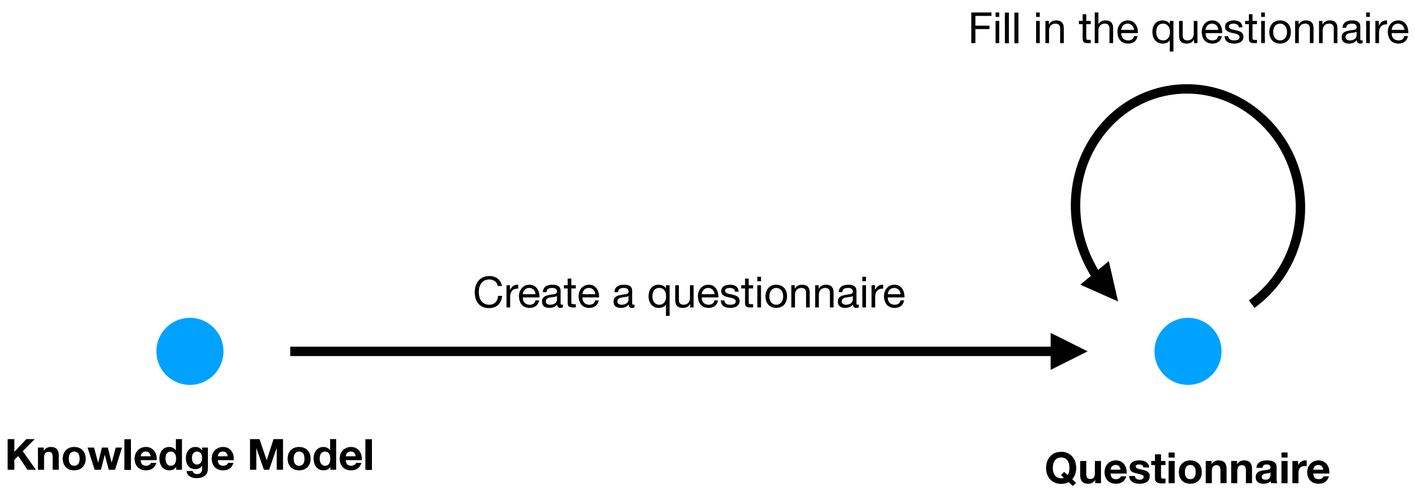
- When a researcher find something unclear or wrong with a question
- Integrated to GitHub issues



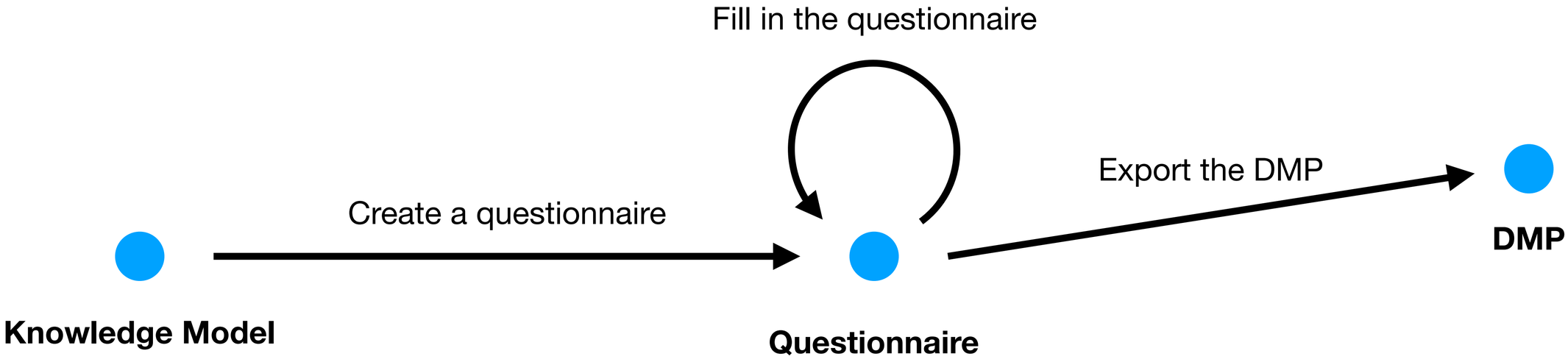
Summary Report

- Information about FAIR metrics for each chapter
- Calculated as a weighted average of answers

Researchers Workflow

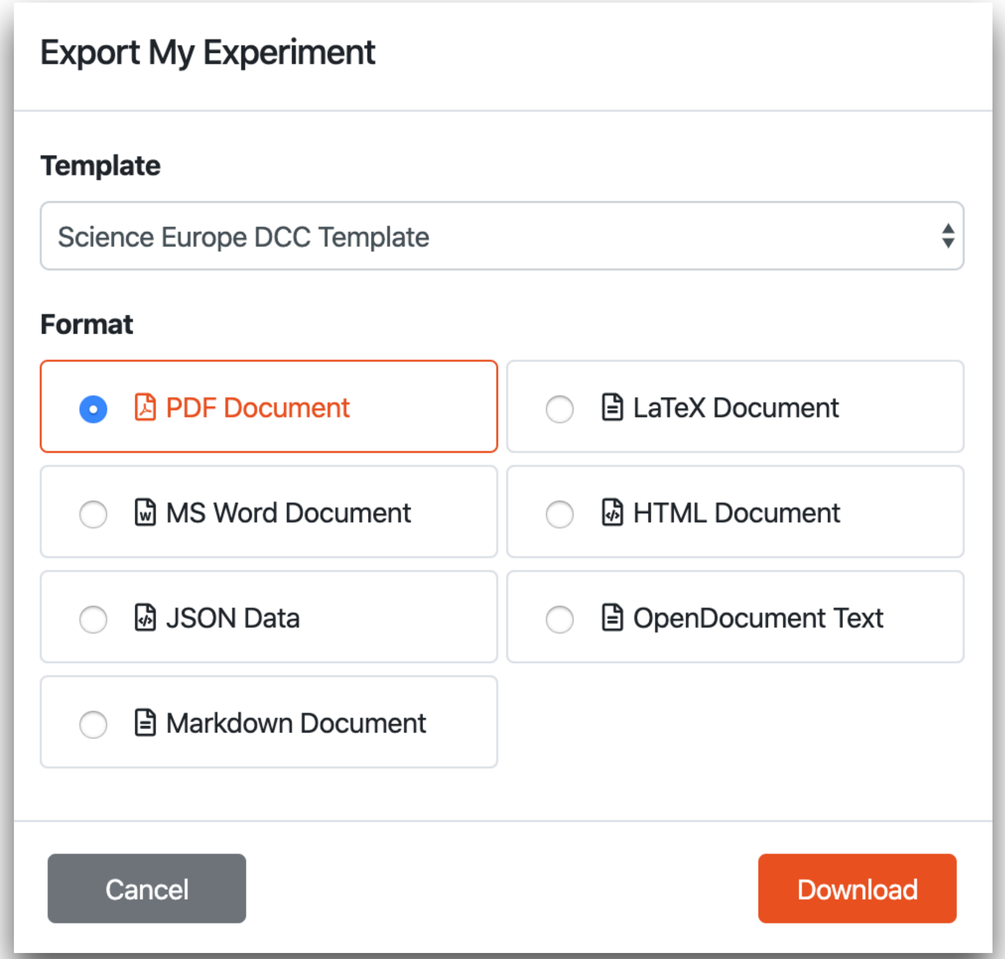


Researchers Workflow



Exporting the DMP

- Documents can be generated from a filled questionnaire
- Customizable document templates
  DSW provides **Science Europe template**
- Various document formats



Exporting the DMP

Data Management Plan

HVSC: Hypothetical Vascular Study by a Chemist

Contact person: ***Bob A*** *(bob.a@example.com, 9999-5559-9999-9999)*
Example Corporation

Based on: *Life Sciences DSW Knowledge Model, 2.0.0 (dsw:lifesciences:2.0.0)*

Generated on: *24. 02. 2020*

Data Management Plan created in Data Stewardship Wizard <https://ds-wizard.org>

Abstract

We will be deriving models of the "virtual" age of arterial walls by combining new data with existing biobanks and reference data, in order to create early warning systems for aging patients.

Section A: Data Collection

1. What data will you collect or create?

Re-used datasets

We will use the following reference datasets:

- **Human Protein Atlas**

  We will use version "19.1" of this dataset. If a new version becomes available during the project, new analyses will be done with the new version.

- **UniProt Knowledgebase** (https://www.uniprot.org/uniprot/)

  We will use version "2019_09" of this dataset.

2. How will the data be collected or created?

There will be no instrument dataset in this project.

Section B: Documentation and Meta-data

3. What documentation and meta-data will accompany the data?

List of data to be published is given in Section E, Question 9. This also includes information about catalogs where the data can be found. Information about data types used is given in Section A, Question 1.

Section C: Ethics and Legal Compliance

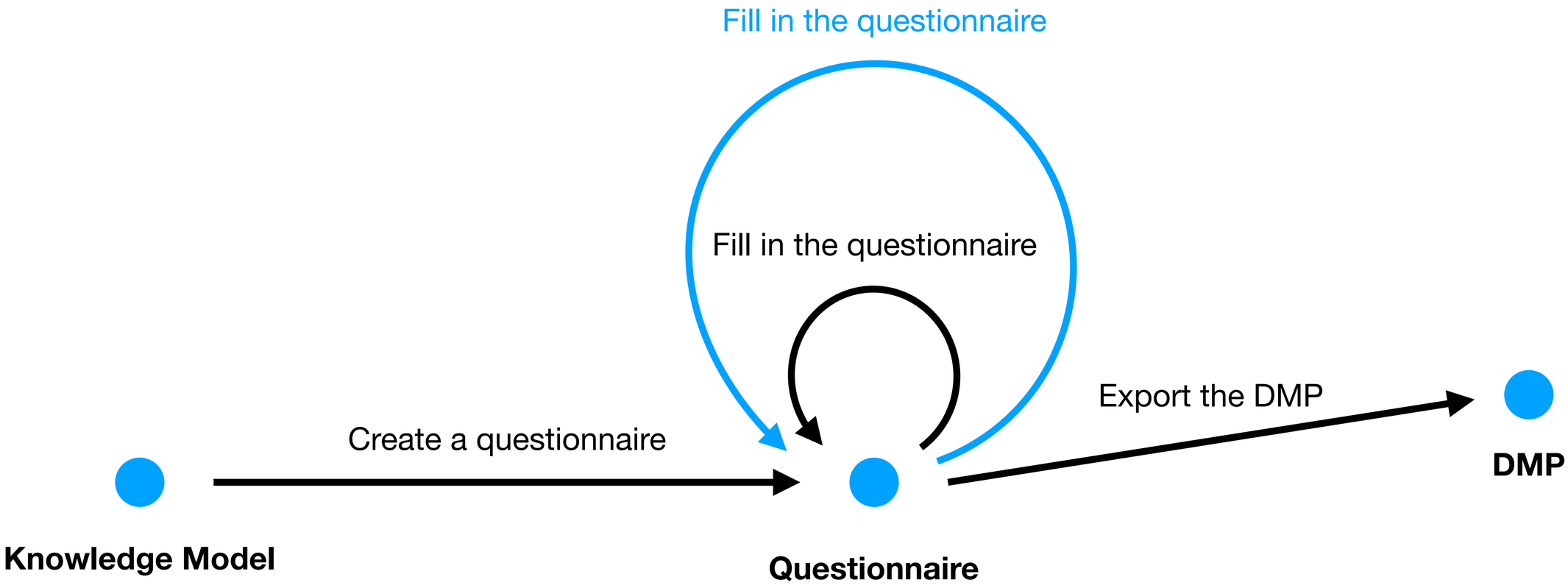
4. How will you manage any ethical issues?

5. How will you manage copyright and Intellectual Property Rights (IPR) issues?

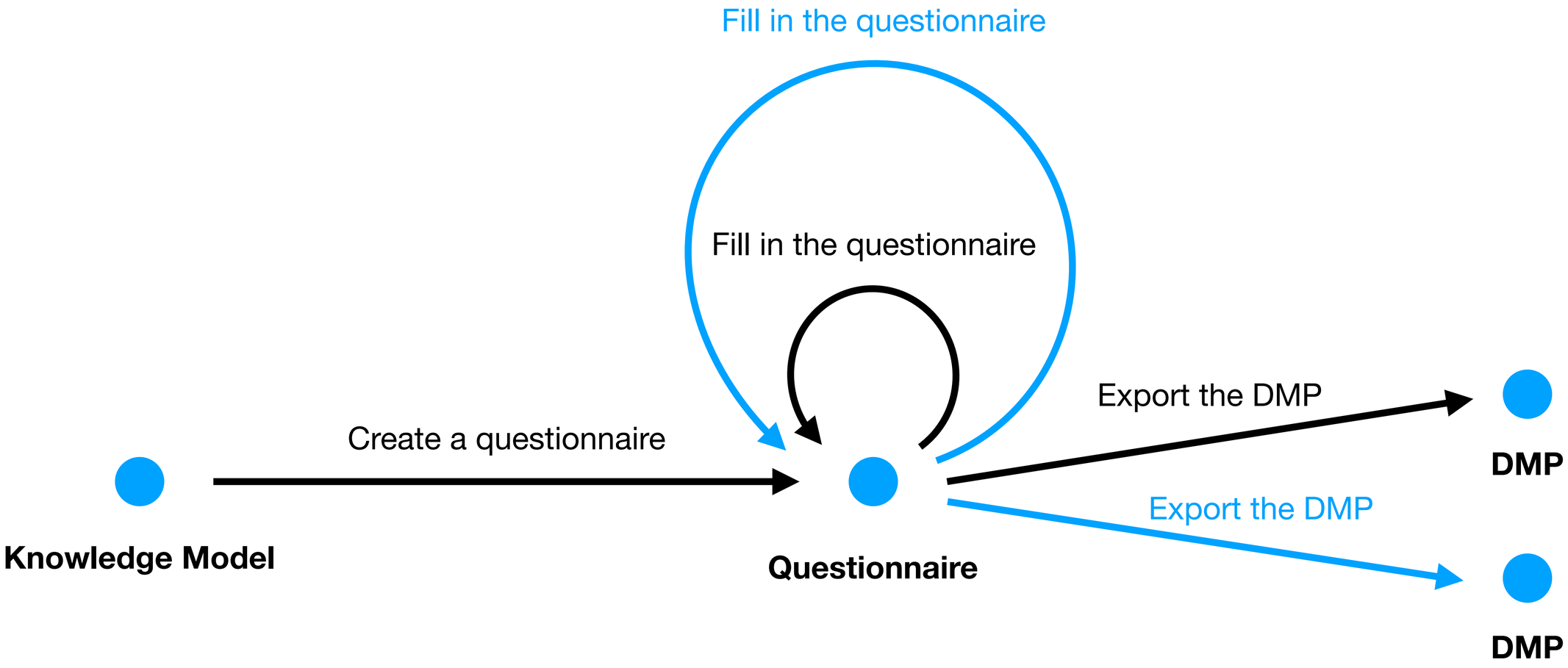
For the reference and non-reference data sets that we reuse, conditions are as follows:

- **Human Protein Atlas** – available under specific restrictions, which we will

Researchers Workflow



Researchers Workflow



Additional Researcher Workflows

- Questionnaire cloning
- Questionnaire migration

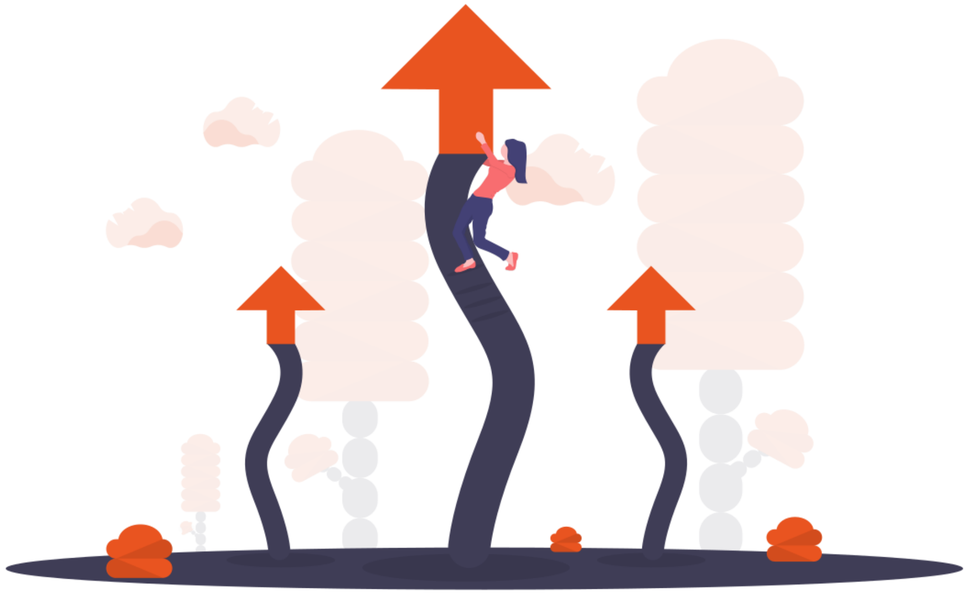
Questionnaire Cloning

- Create a **copy of existing questionnaire**

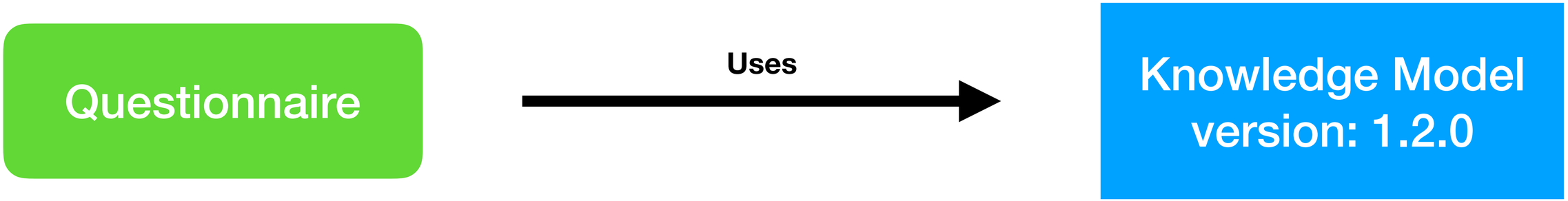
- Can be used to create partially filled questionnaire templates
  - Public Read-Only questionnaire as a template
  - Researcher can clone it
  - And fill in the specific details

Questionnaire Migration

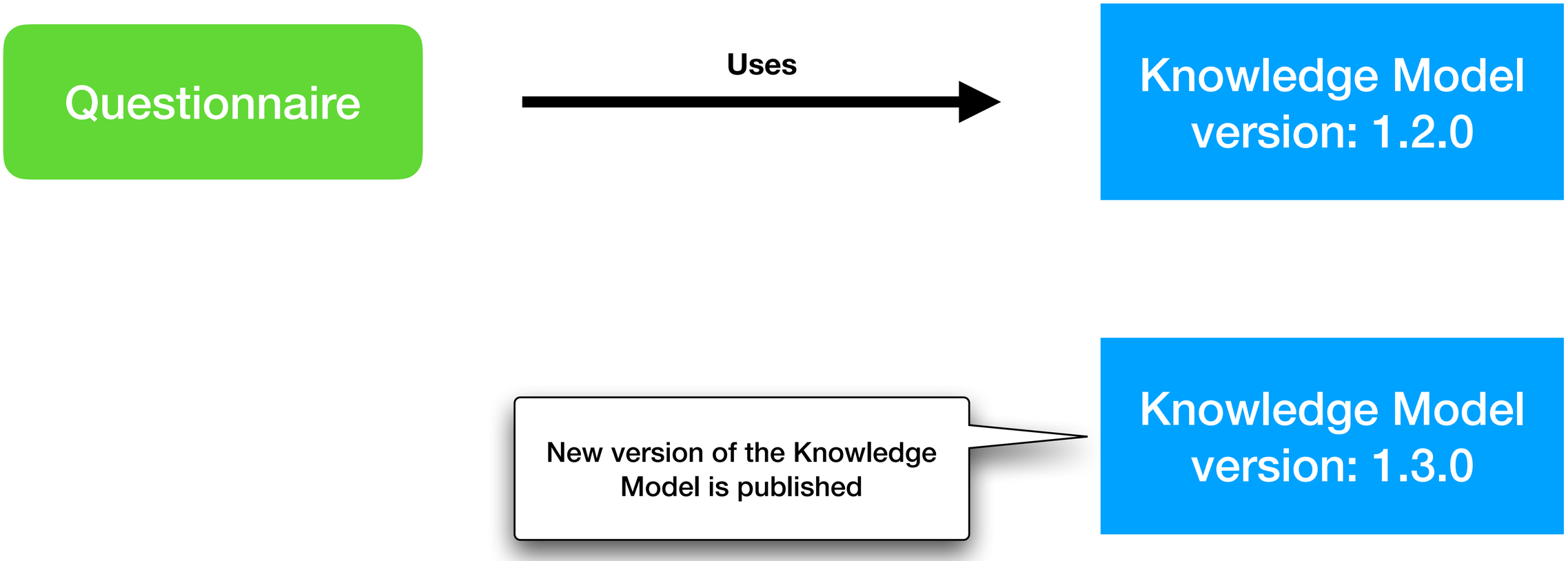
- Process of upgrading existing Questionnaires with their answers to new versions of Knowledge Models



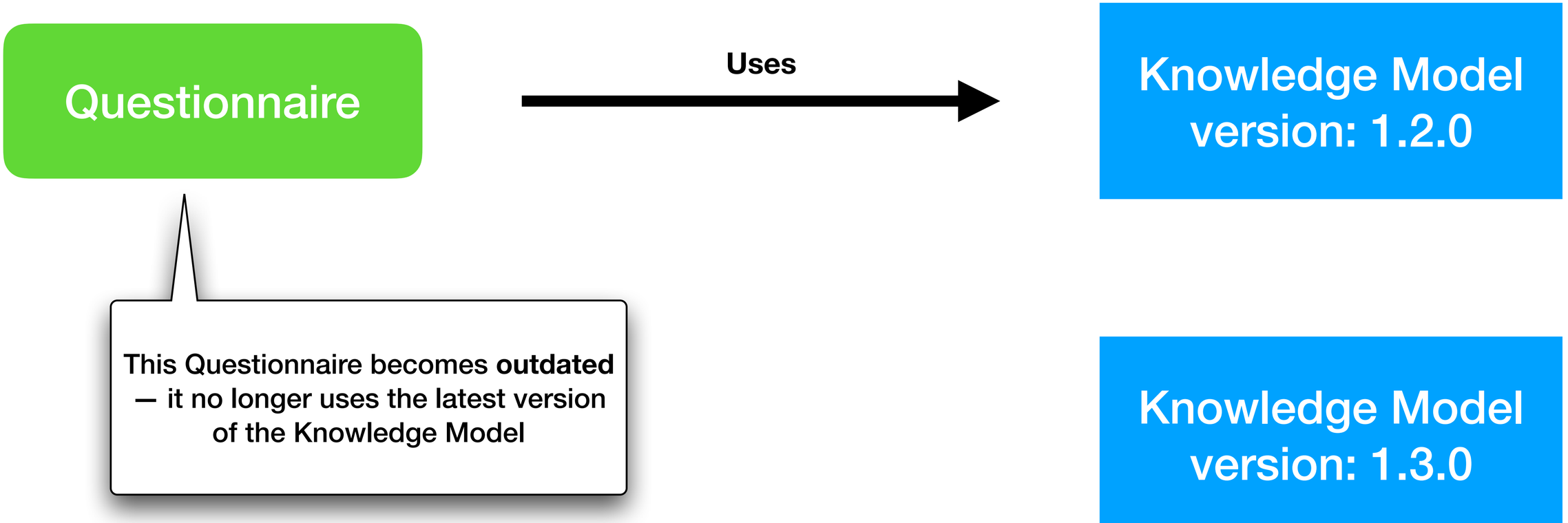
Questionnaire Migration



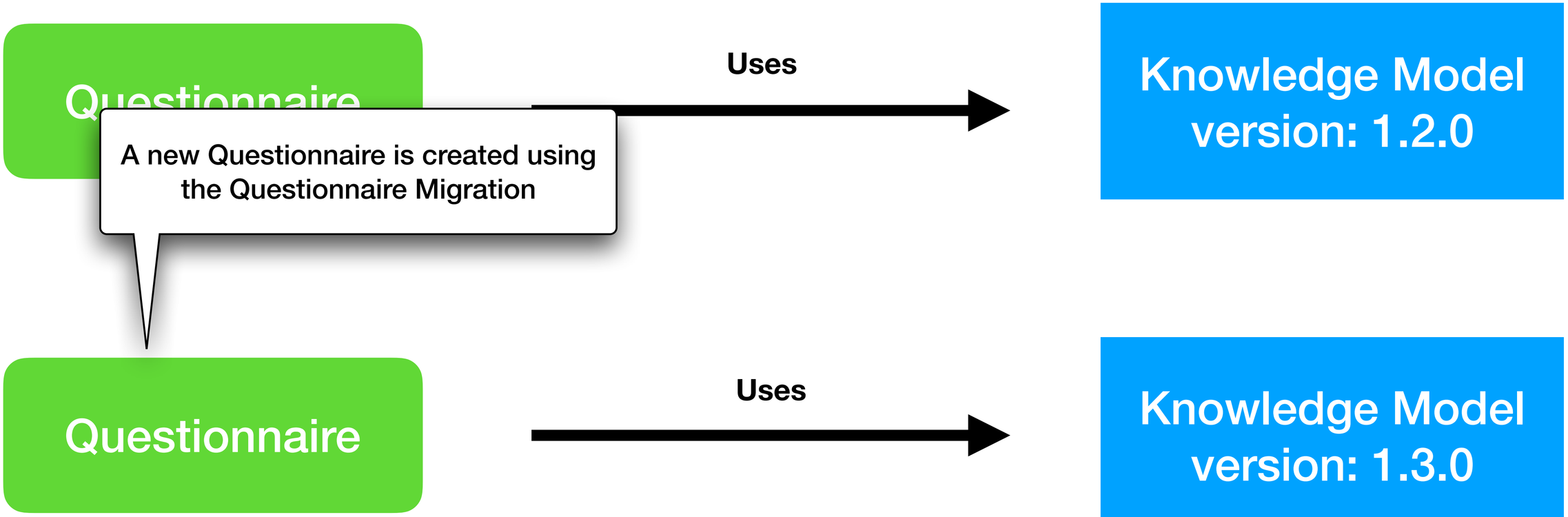
Questionnaire Migration



Questionnaire Migration



Questionnaire Migration



Researchers Workflow Demo

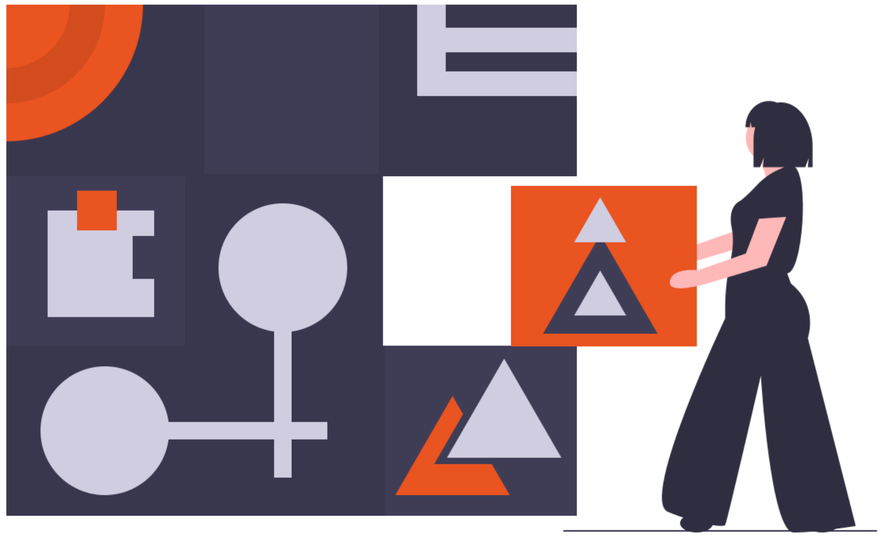
DSW for Data Stewards

Data Steward Responsibilities in DSW

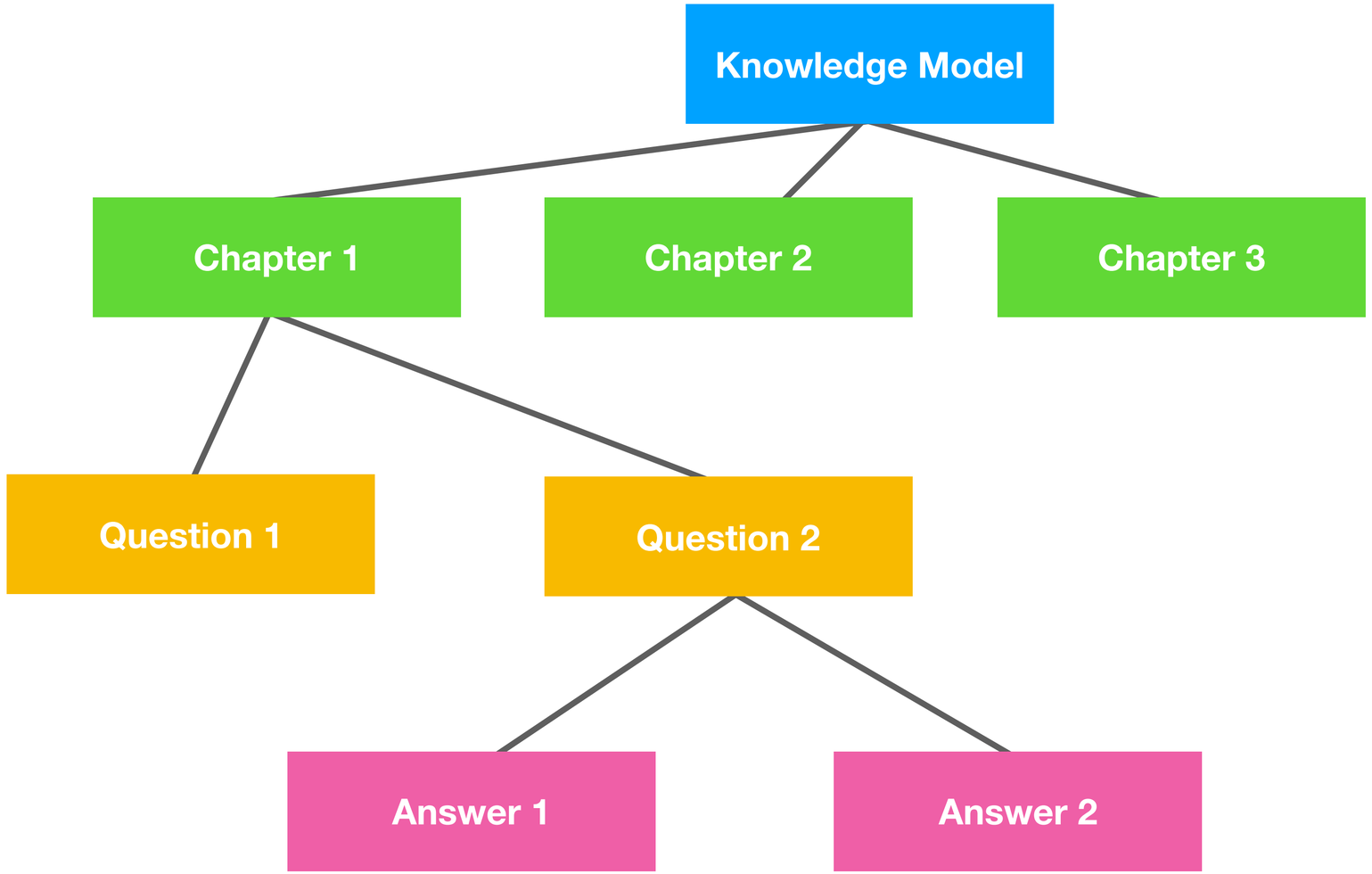
- Manage existing Knowledge Models
- Build & publish new Knowledge Models
- Import Knowledge Models from other DSW instances or Registry

Building a Knowledge Model

- From scratch
- Extending existing Knowledge Model



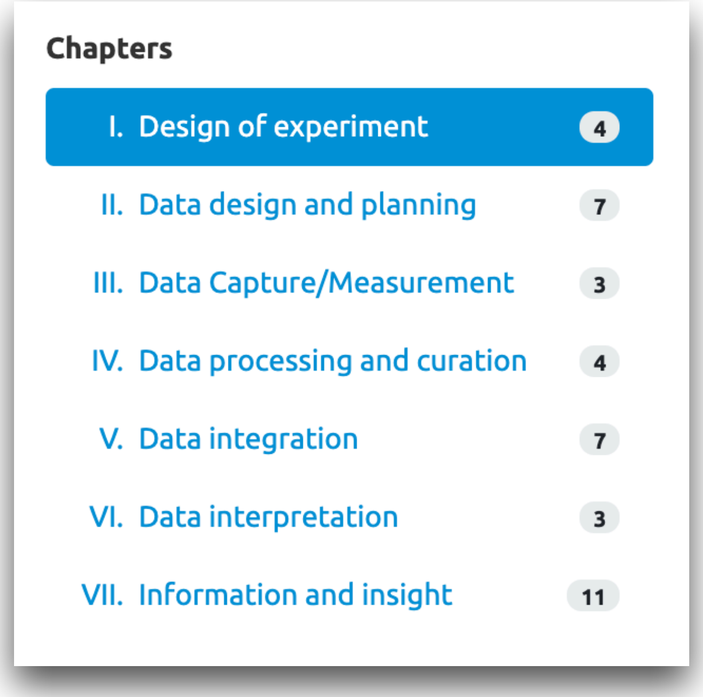
Knowledge Model



- Core DS Knowledge Model
  - Design of experiment
    - Is there any pre-existing data?
      - ☑ No
      - ☑ Yes
      - atq
    - 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?
      - Data format/type:
      - Is this a standard data format used by others too?
      - Does this data format enable sharing and long term archiving?
        - ☑ No
        - ☑ Yes
      - njy
    - 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 stora
    - Will you need a shared working space to work with your data?
    - Is the risk of information loss, leaks and vandalism acceptably low?
    - Do you need to do compute capacity planning?
  - Data Capture/Measurement
  - Data processing and curation

Chapter

- At the top level of each Knowledge Model
- Contains **short introduction** and **questions**

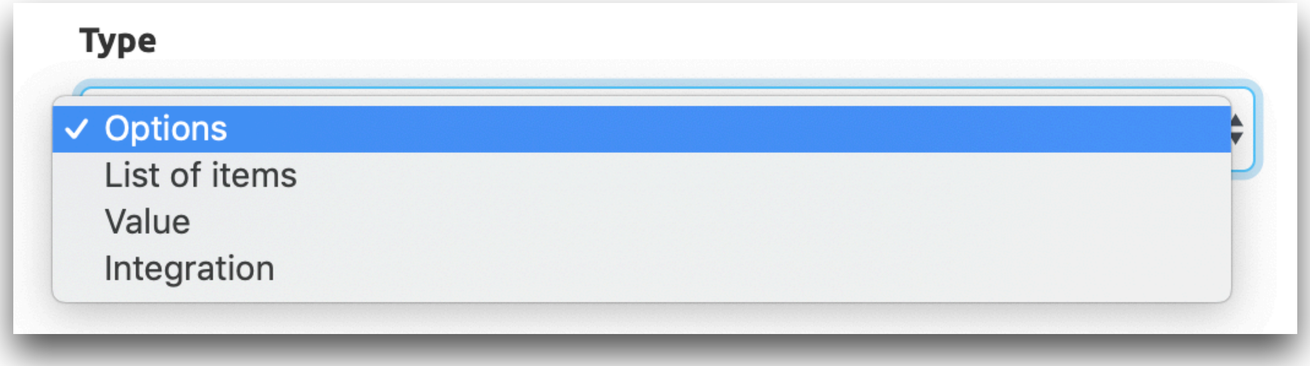


Question

- Core entity of the Knowledge Model
- Each question has a **Title** and a description **Text**
- Can be of different types
- Can have **References** and **Experts** assigned

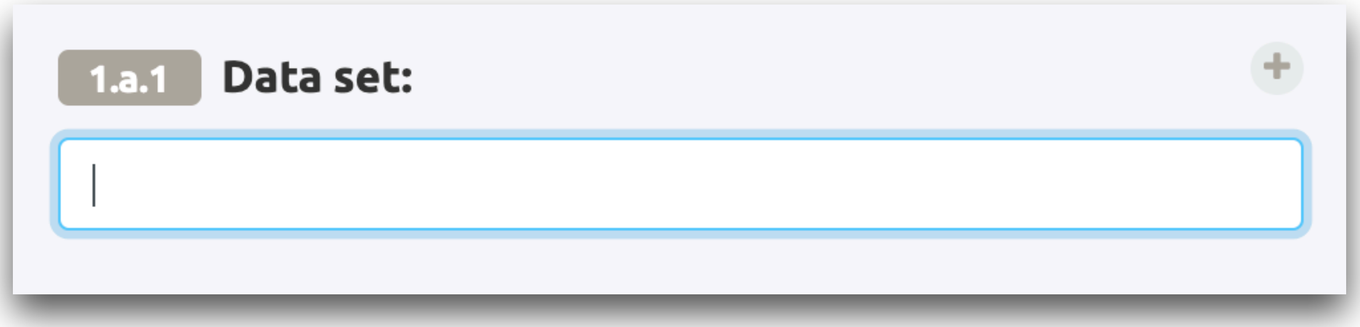
Question Type

- Value
- Integration
- Options
- List of items



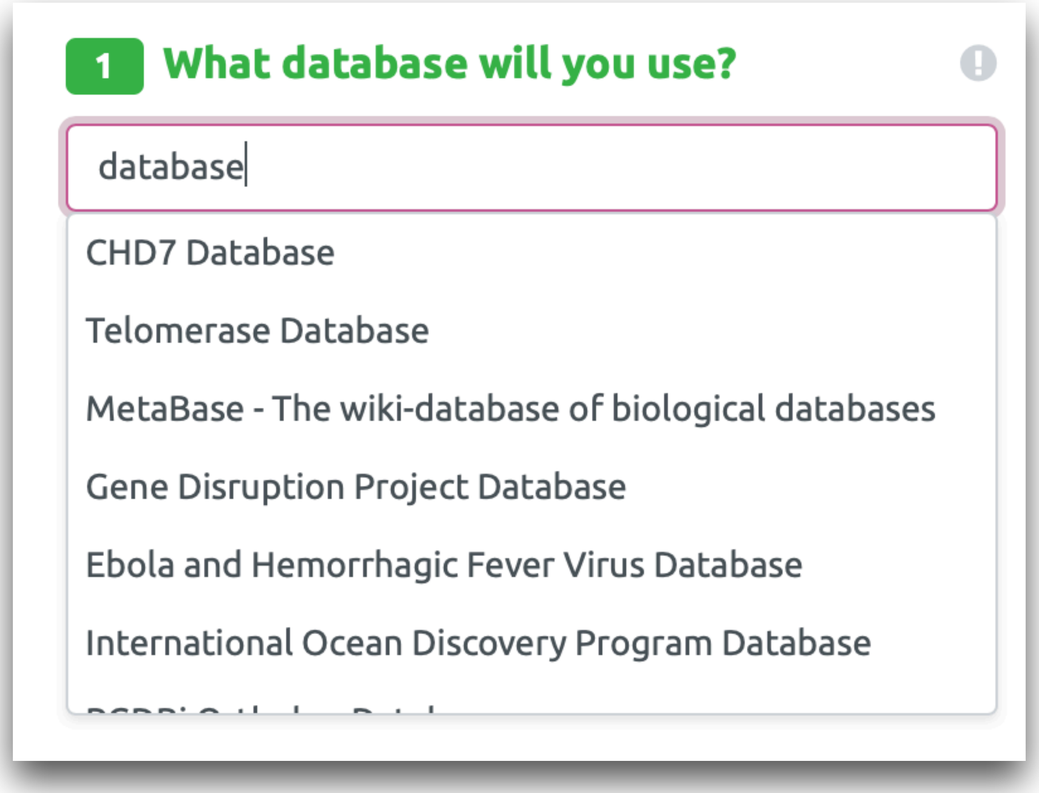
Question Type: Value

- Simple value that users type
- Value types
  - Number
  - Date
  - String
  - Text



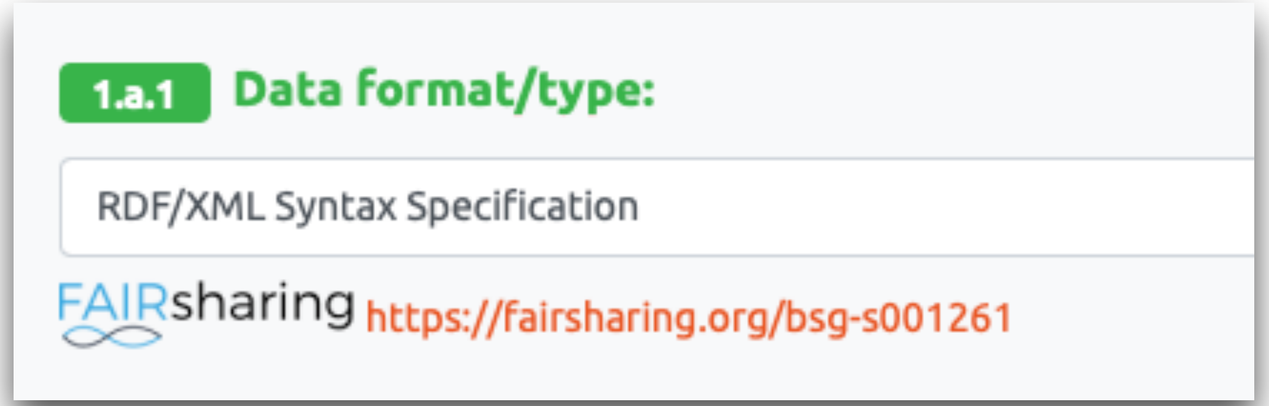
Question Type: Integration

- Similar to Value type
- The answers can be taken **from external resource**
- More complex to set up



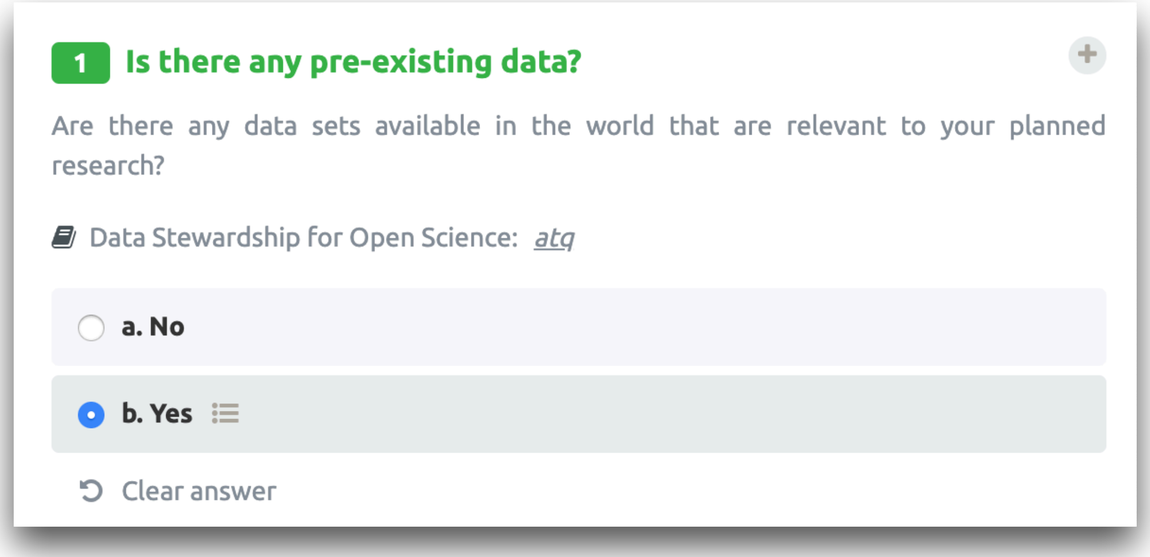
Question Type: Integration

- Similar to Value type
- The answers can be taken **from external resource**
- More complex to set up

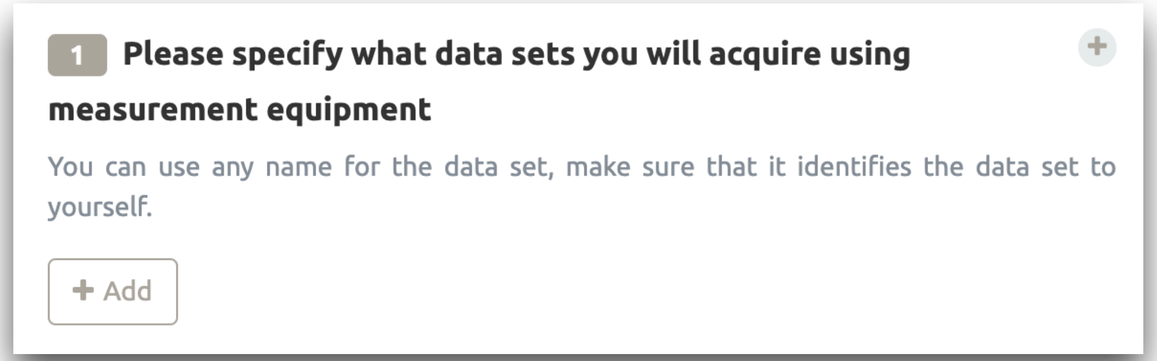


Question Type: Options

- Closed list of **answers**
- Users can pick one
- Answers can have **follow-up questions**



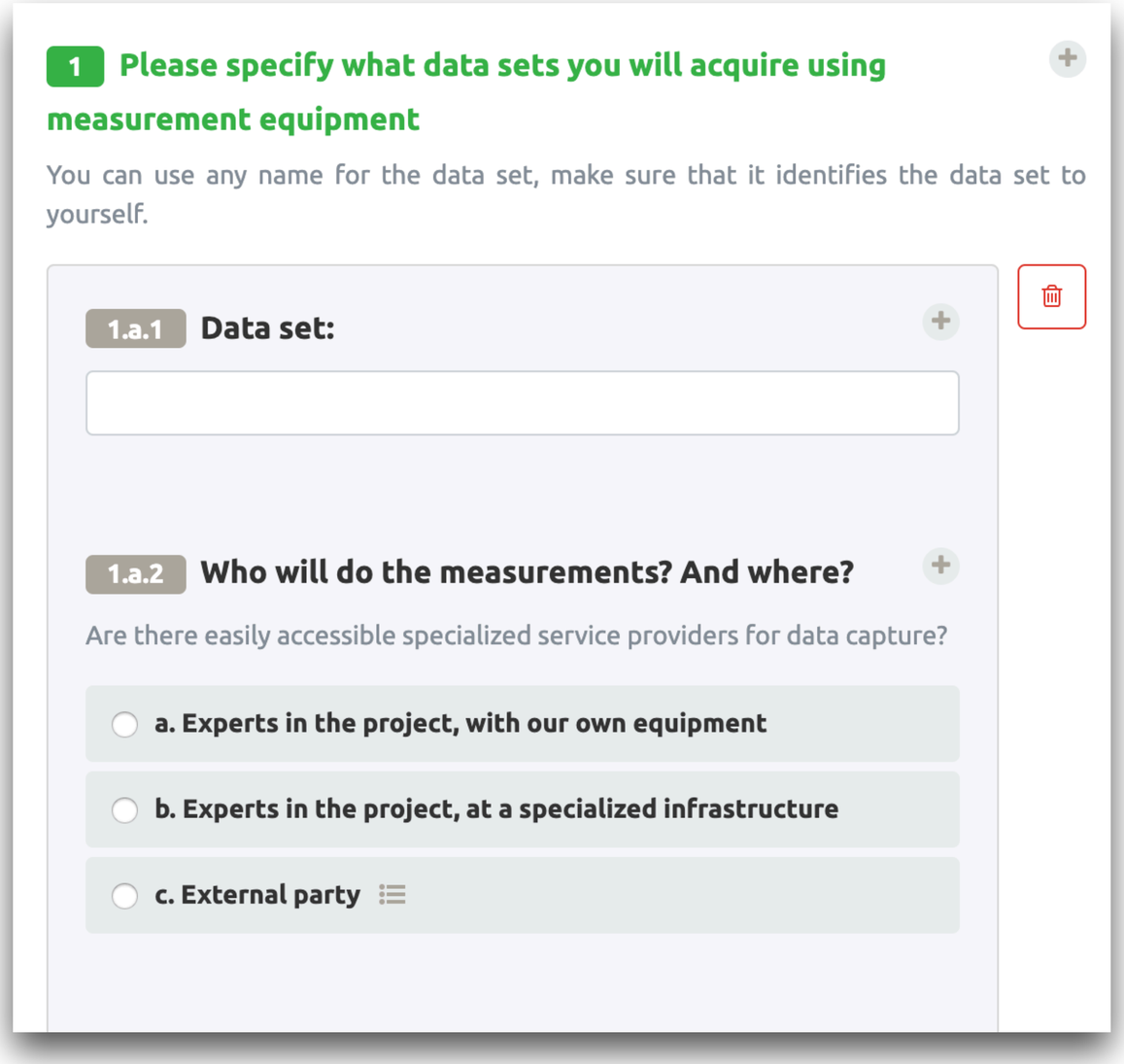
Question Type: List of Items



- Multiple **items** of the same type
- Each item has the same set of questions

Question Type: List of Items

- Multiple **items** of the same type
- Each item has the same set of questions

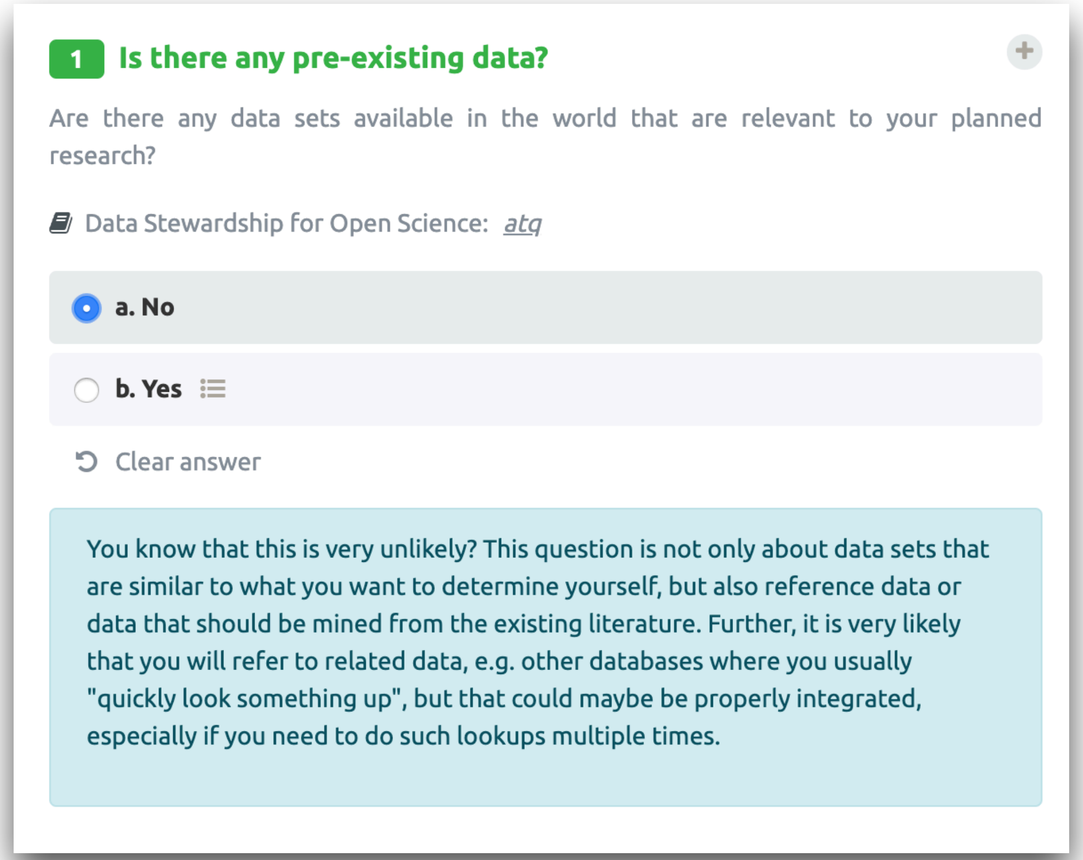


Answer

- Used in **Options Question**
- Can have **advice**, **follow-up questions** and **FAIR metrics** assessment

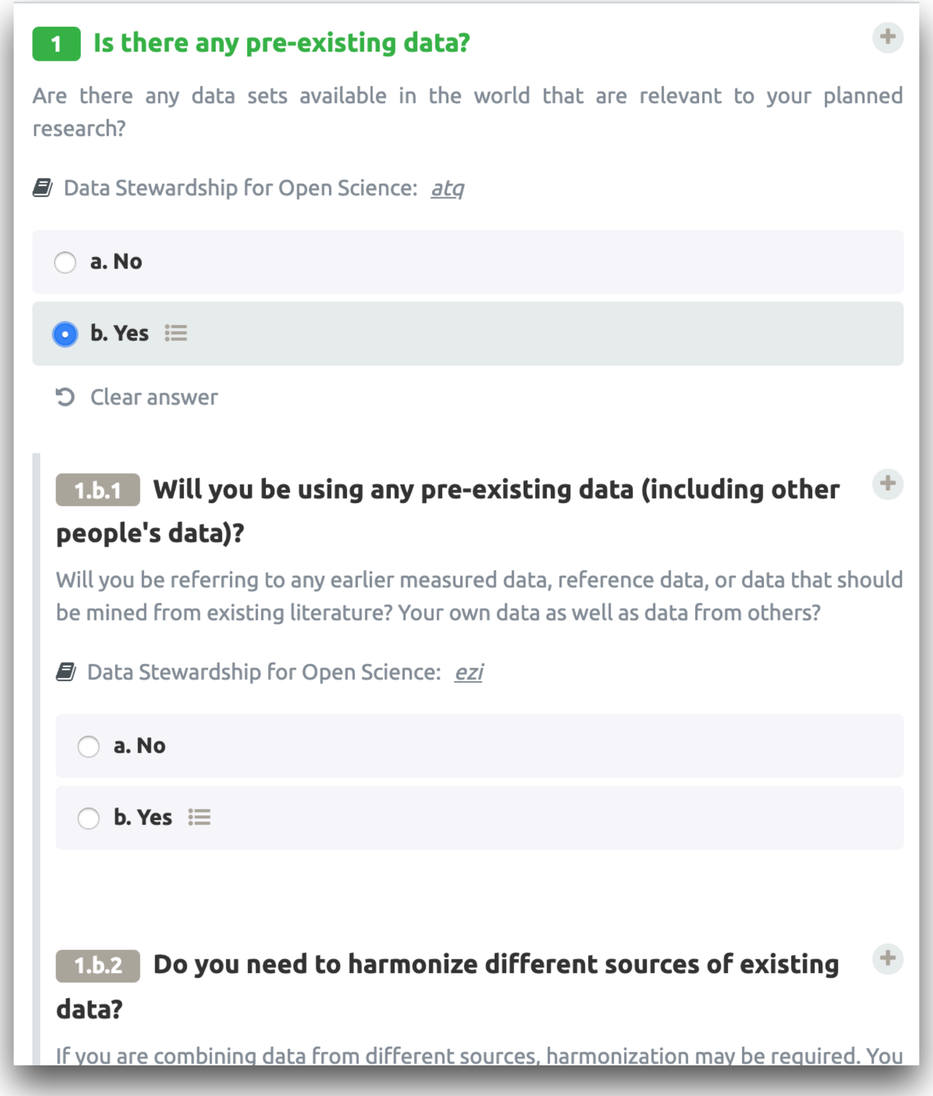
Answer Advice

- When an answer is selected, advice can be provided
- Guide users to achieve better results



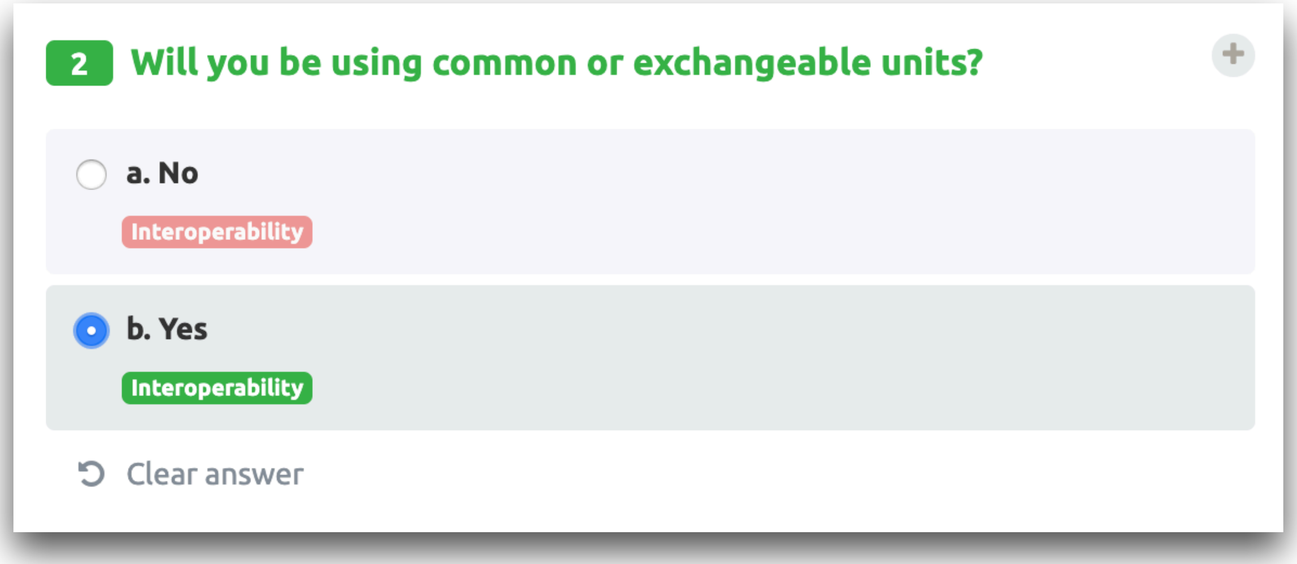
Follow-up Questions

- When an answer is selected, more questions can be unfolded
- Used for questions that are relevant only until certain conditions



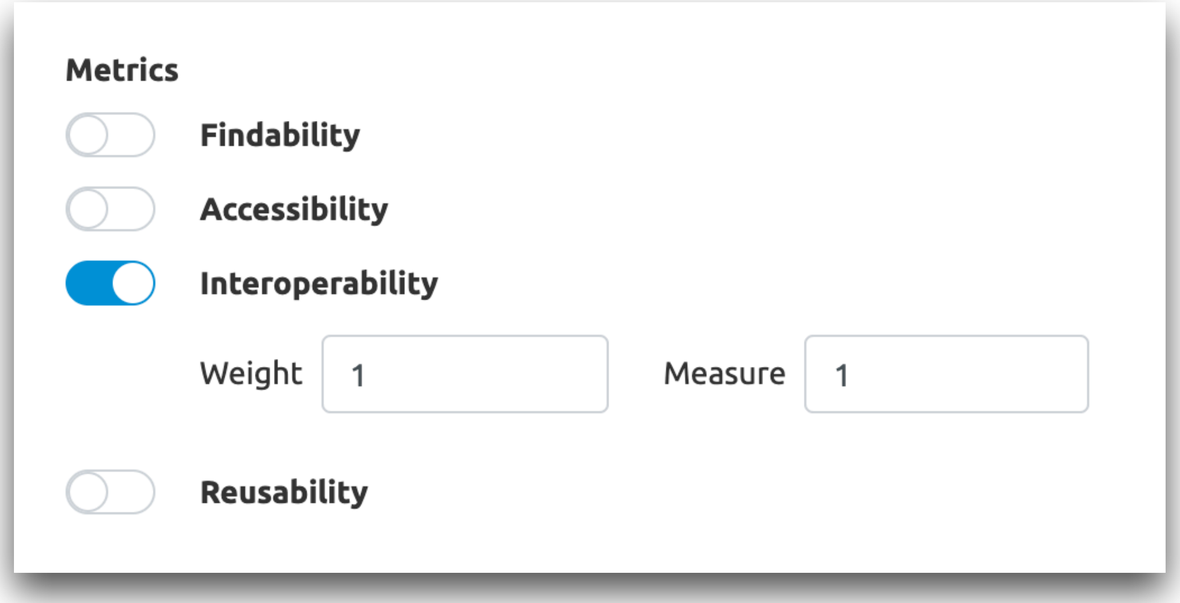
FAIR Metrics

- Each answer can affect different **FAIR metrics**
- Result is calculated as a weighted average of selected answers



FAIR Metrics

- Each answer can affect different **FAIR metrics**
- Result is calculated as a weighted average of selected answers



FAIR Metrics

- **Weight** — how important the answer is

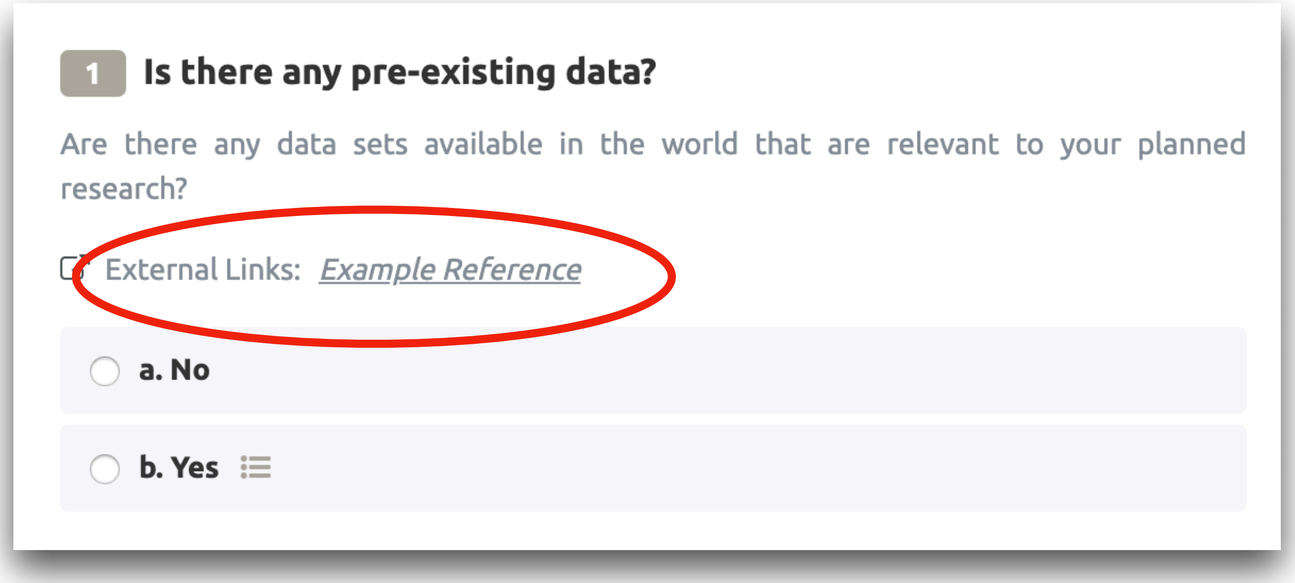
Not important **0**  **1** Very Important

- **Measure** — how does it affect the result

Bad **0**  **1** Good

Reference

- Can be used to provide links to **external resources**
- User can open links when filling the Questionnaire

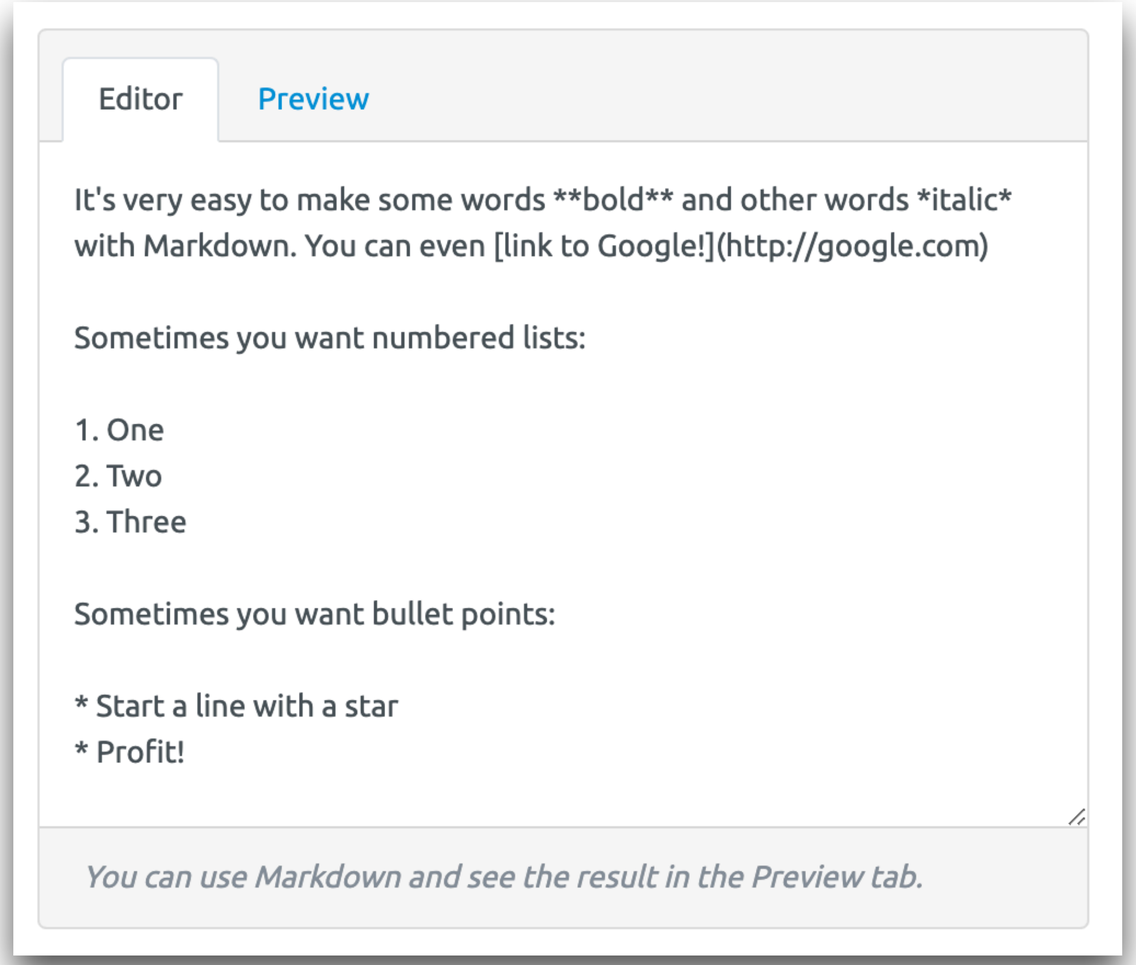


Expert

- Can be used to provide contacts to experts
- Users can see the contact by the question

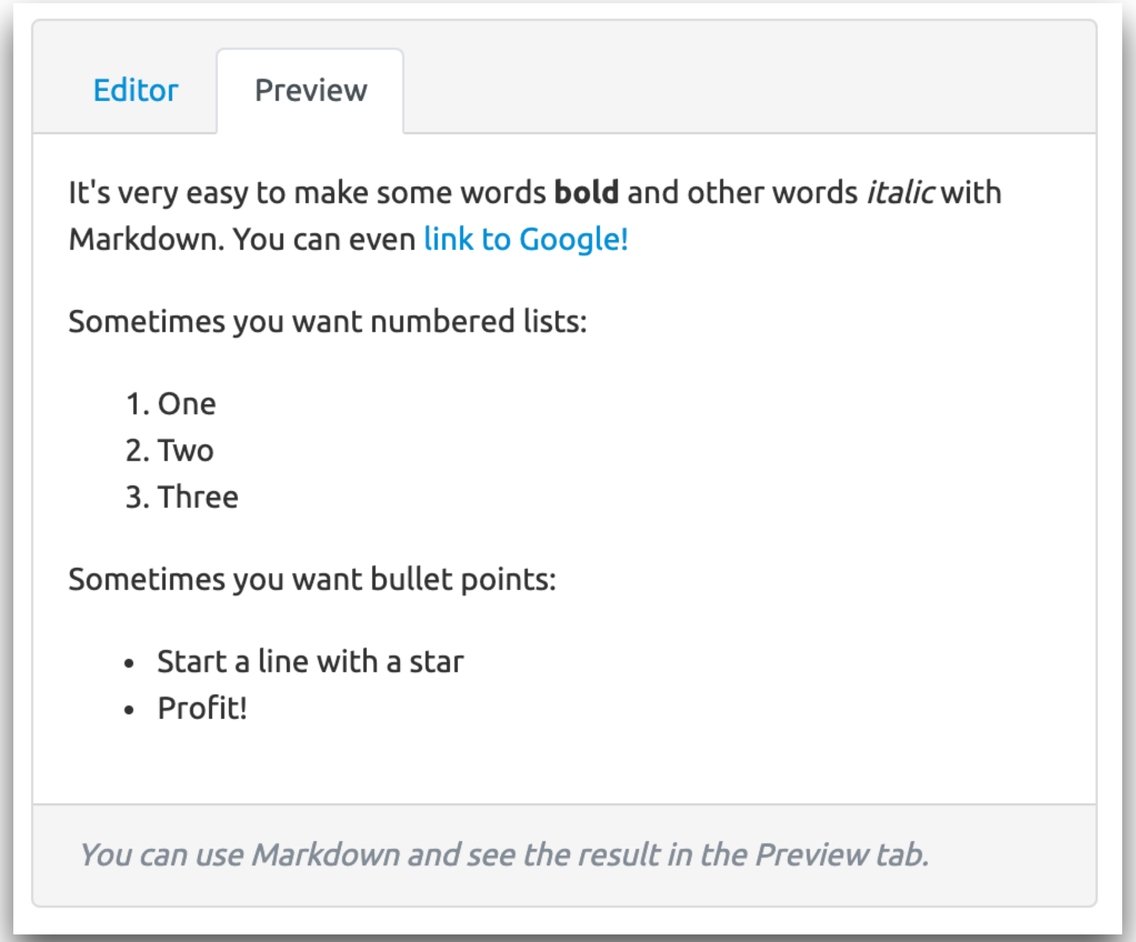
Using Markdown

- Markdown is a lightweight syntax for styling texts
- You can use basic features
  - bold or italic text
  - links
  - bullet or numbered lists
- More examples: https://guides.github.com/features/mastering-markdown/



Using Markdown

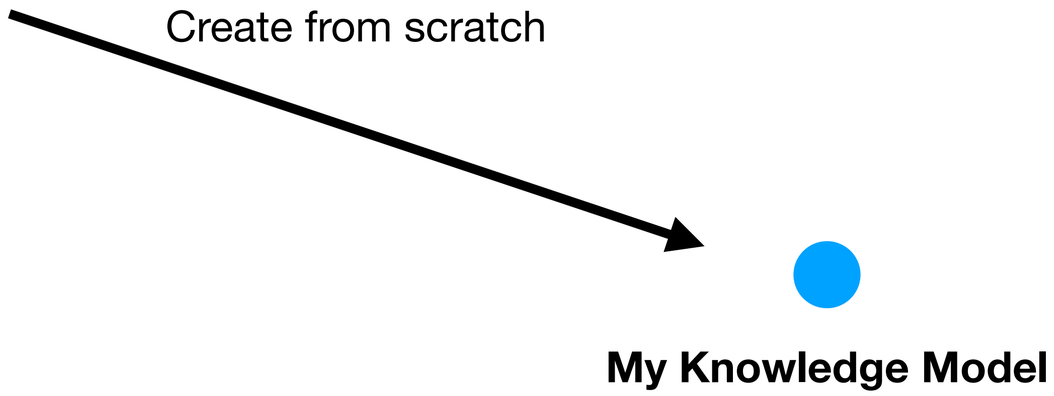
- Markdown is a lightweight syntax for styling texts
- You can use basic features
  - bold or italic text
  - links
  - bullet or numbered lists
- More examples: https://guides.github.com/features/mastering-markdown/



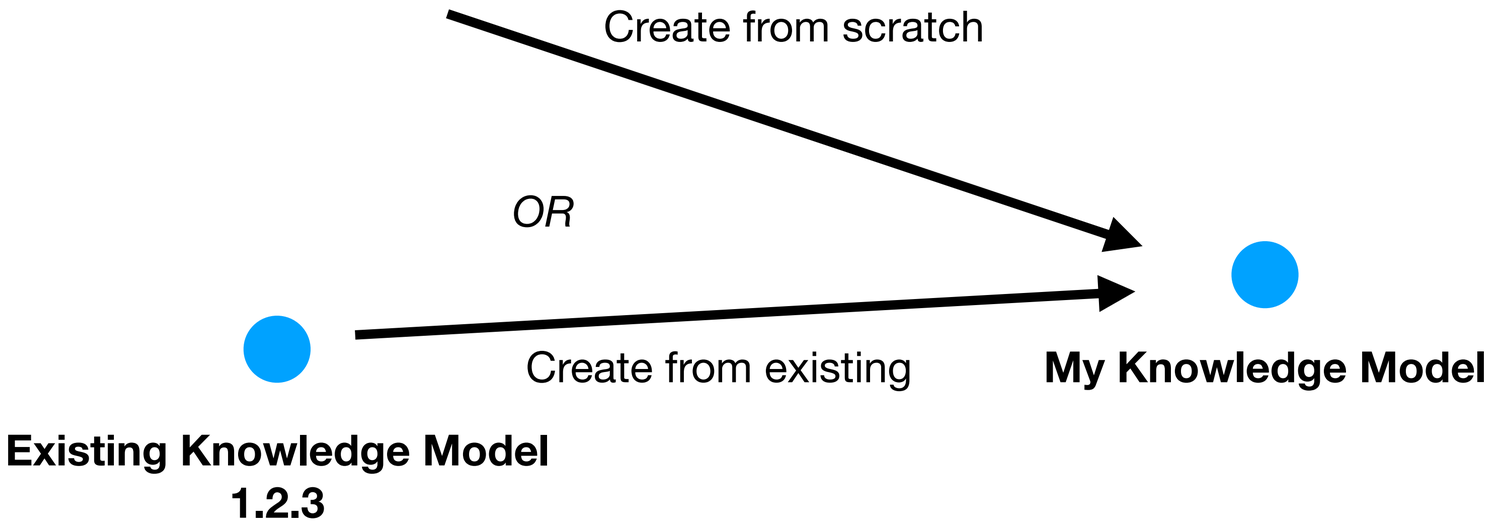
Publishing the Knowledge Model

- Before a Knowledge Model can be used, it has to be **published**
- **Version**, **description** and **readme** is assigned
- Once the version is published **it cannot be changed**
- New changes as a new version
- Published Knowledge Models can be used to create a Questionnaire

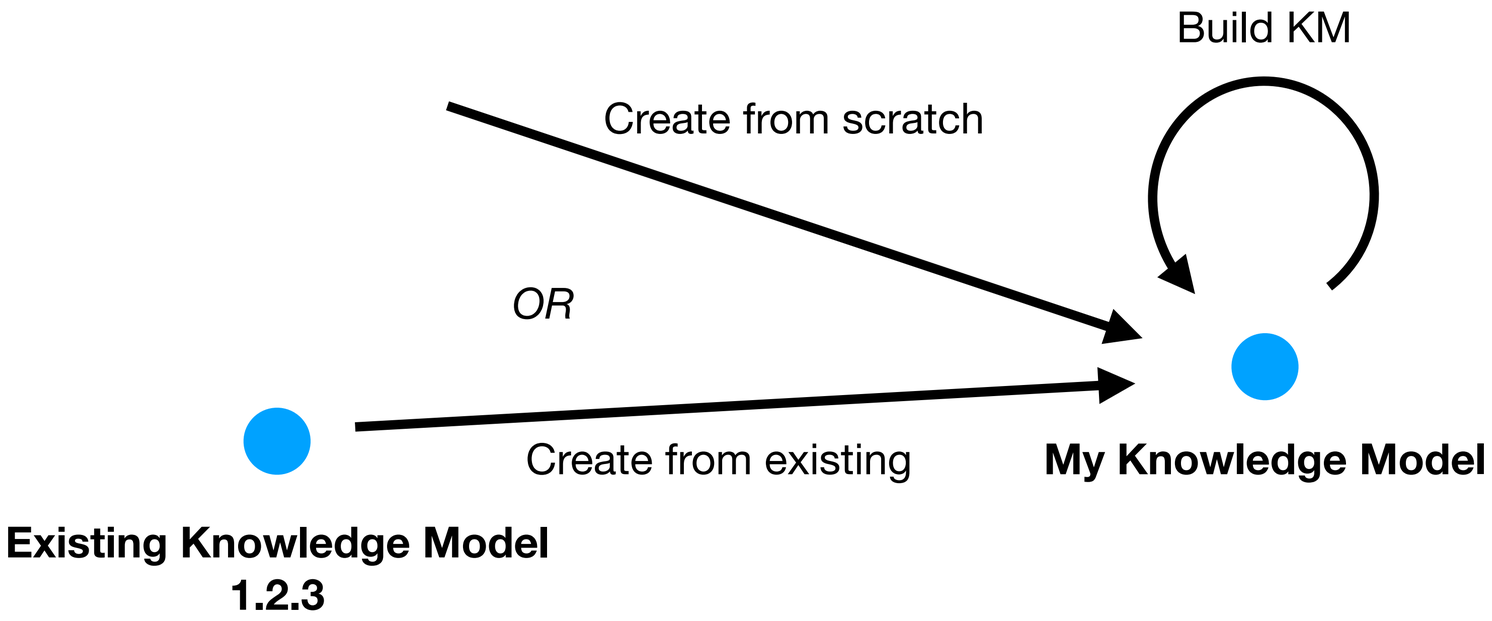
Knowledge Model Workflow



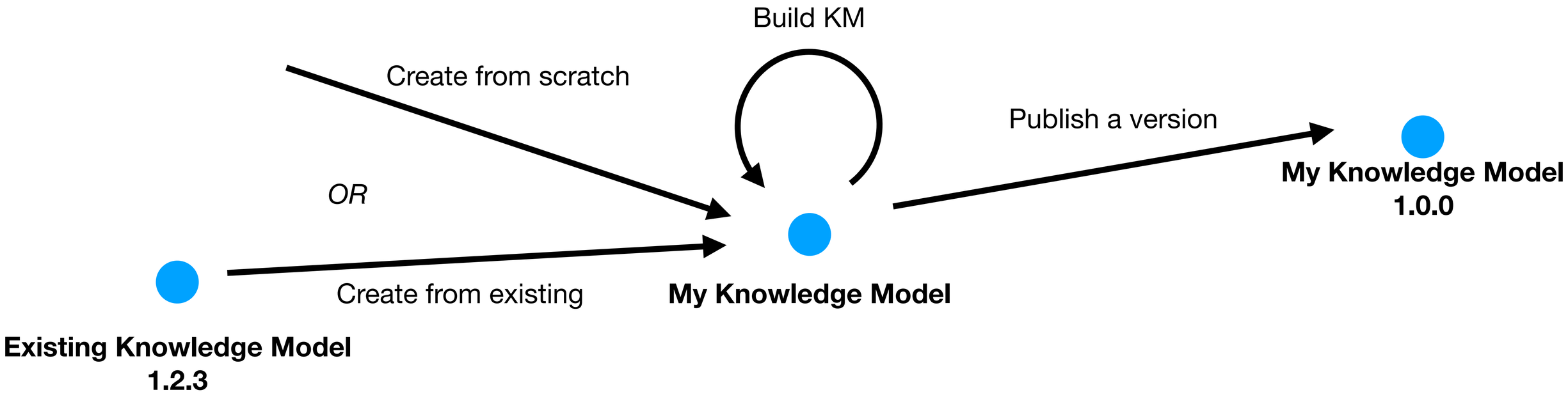
Knowledge Model Workflow



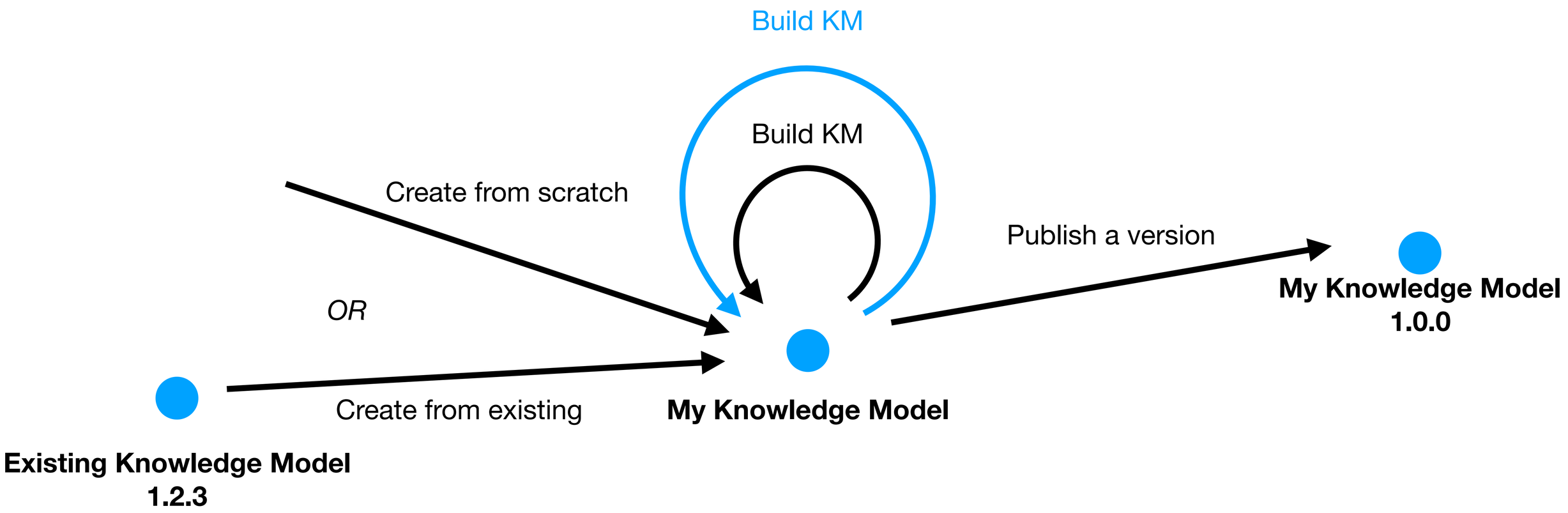
Knowledge Model Workflow



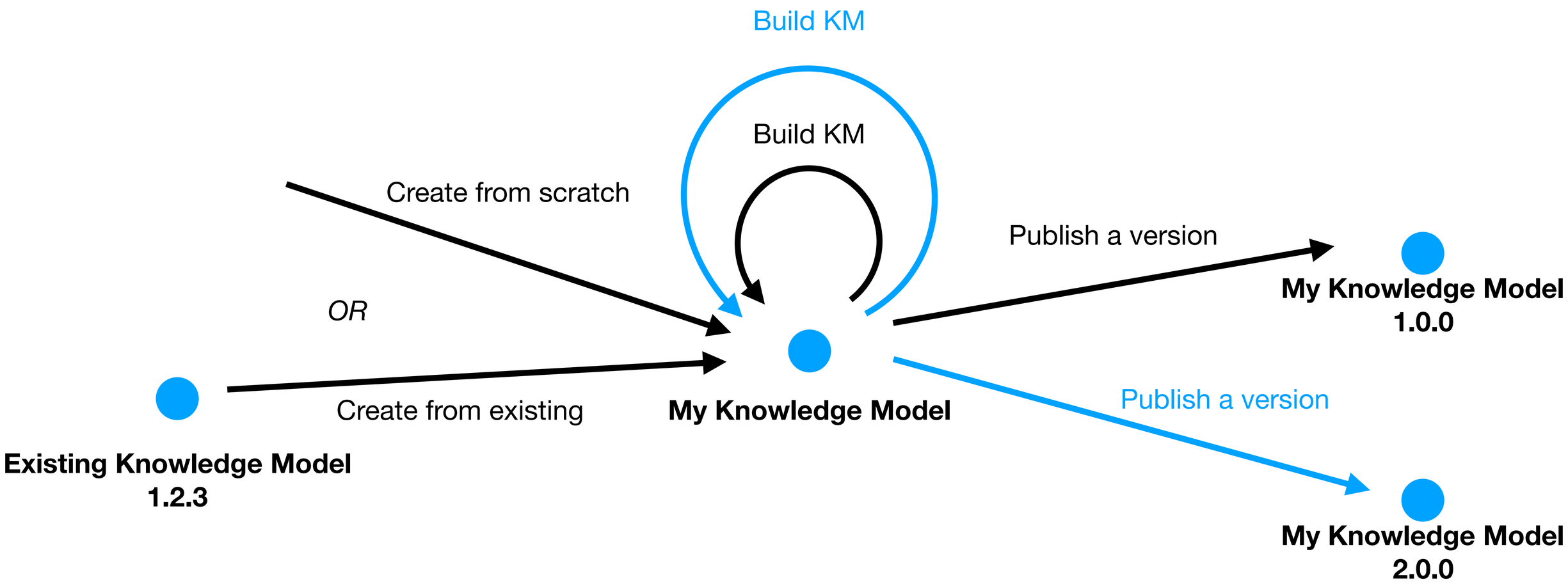
Knowledge Model Workflow



Knowledge Model Workflow



Knowledge Model Workflow

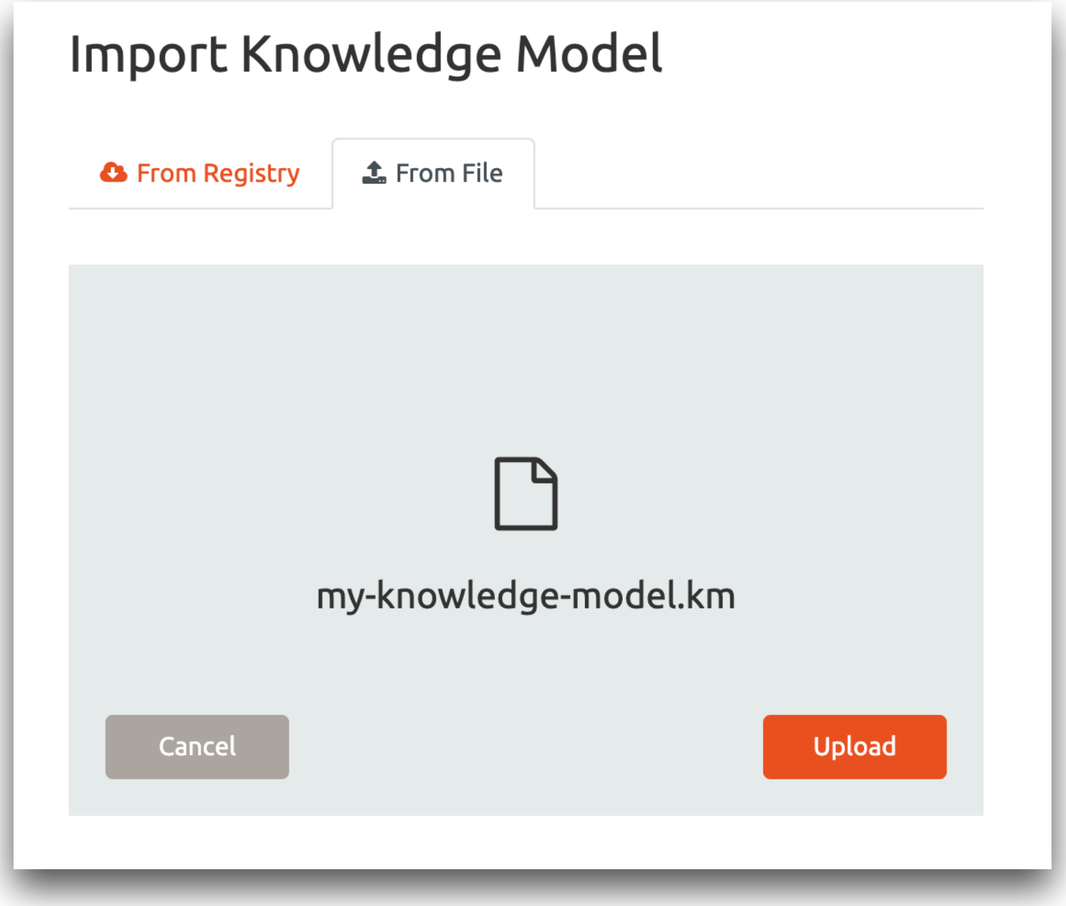


Knowledge Model Import & Export

- From file
- From Registry

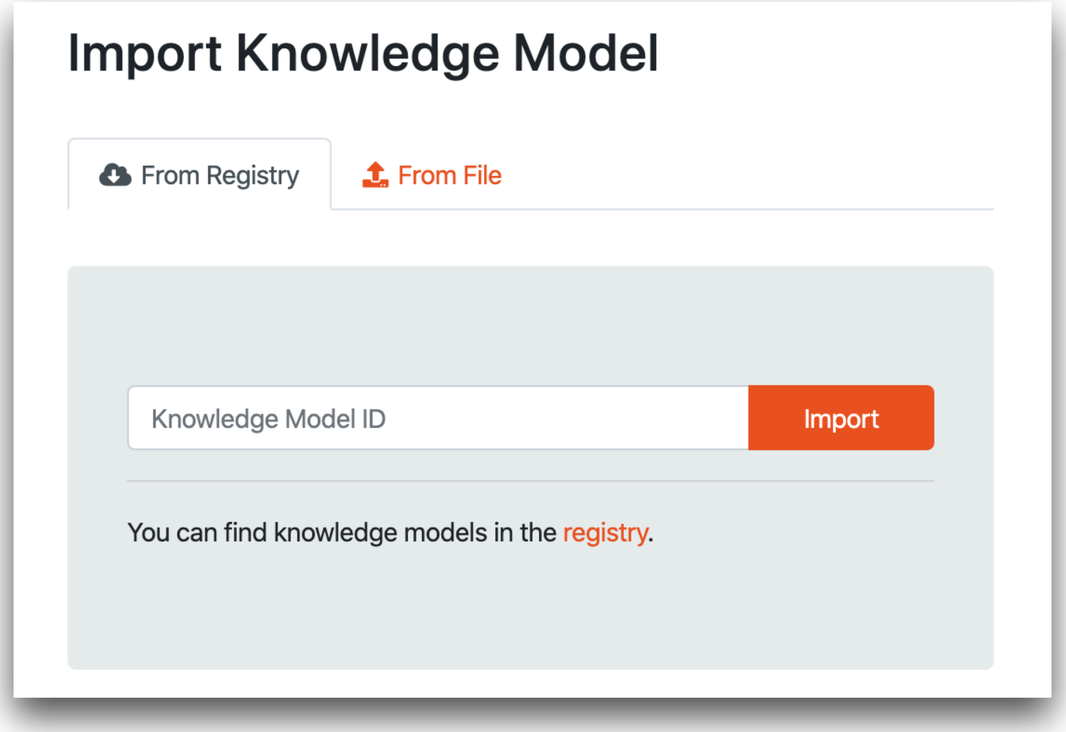
Import & Export with Files

- When a Knowledge Model was published it can be **exported**
- A *.km* file is downloaded from the DSW Instance
- It can be **imported** to a different DSW Instance

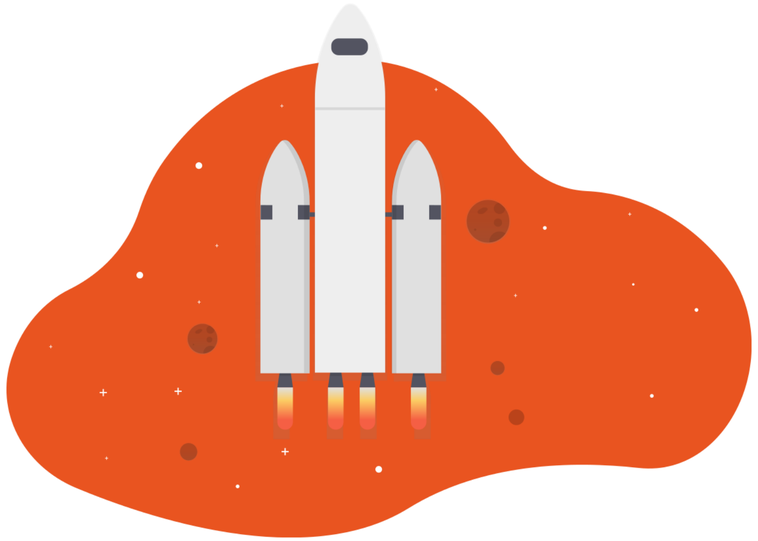


Import from Registry

- DSW Knowledge Models are publicly available in registry.ds-wizard.org
- Easy import into DSW instances using **Knowledge Model ID**



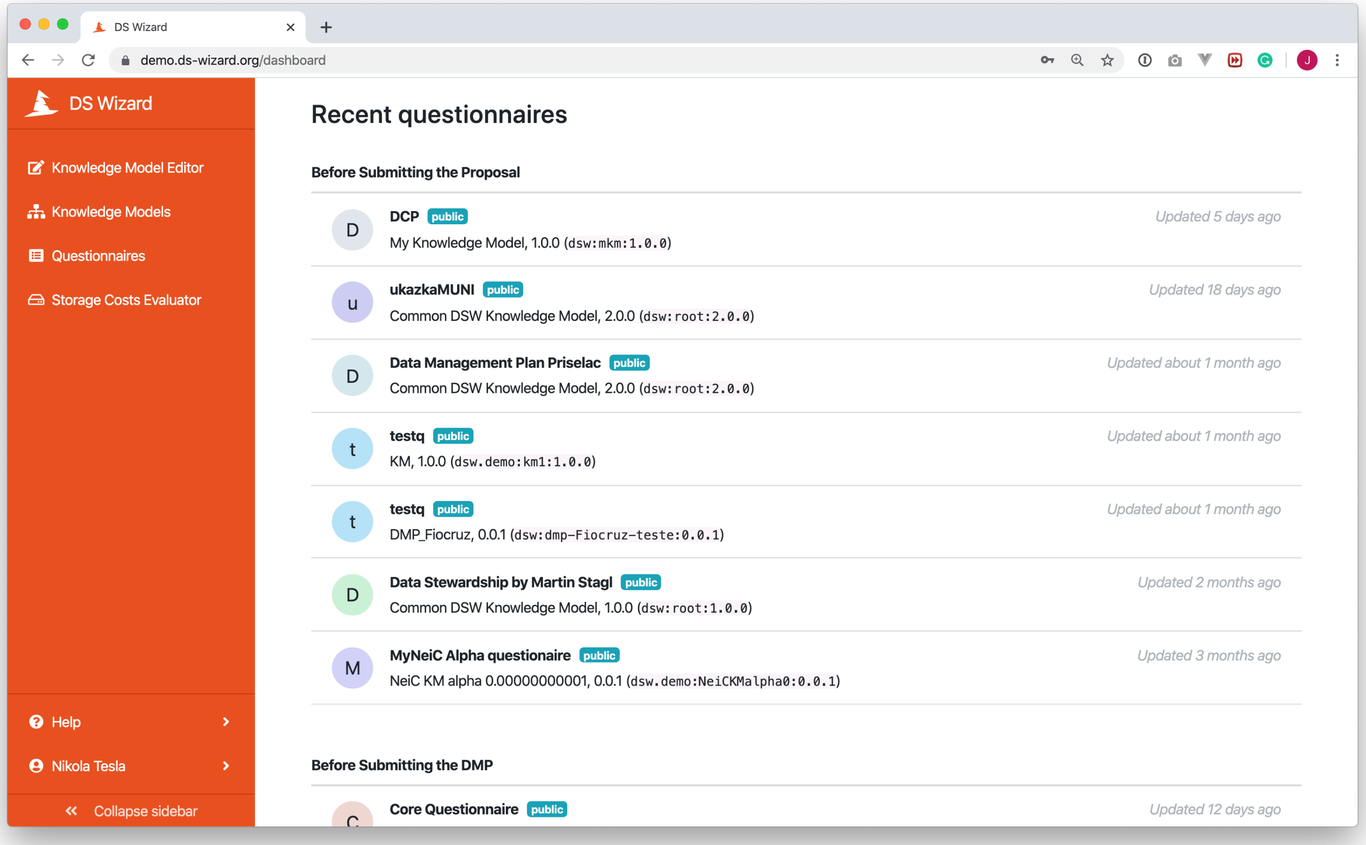
Data Stewards Demo



How to Get Started

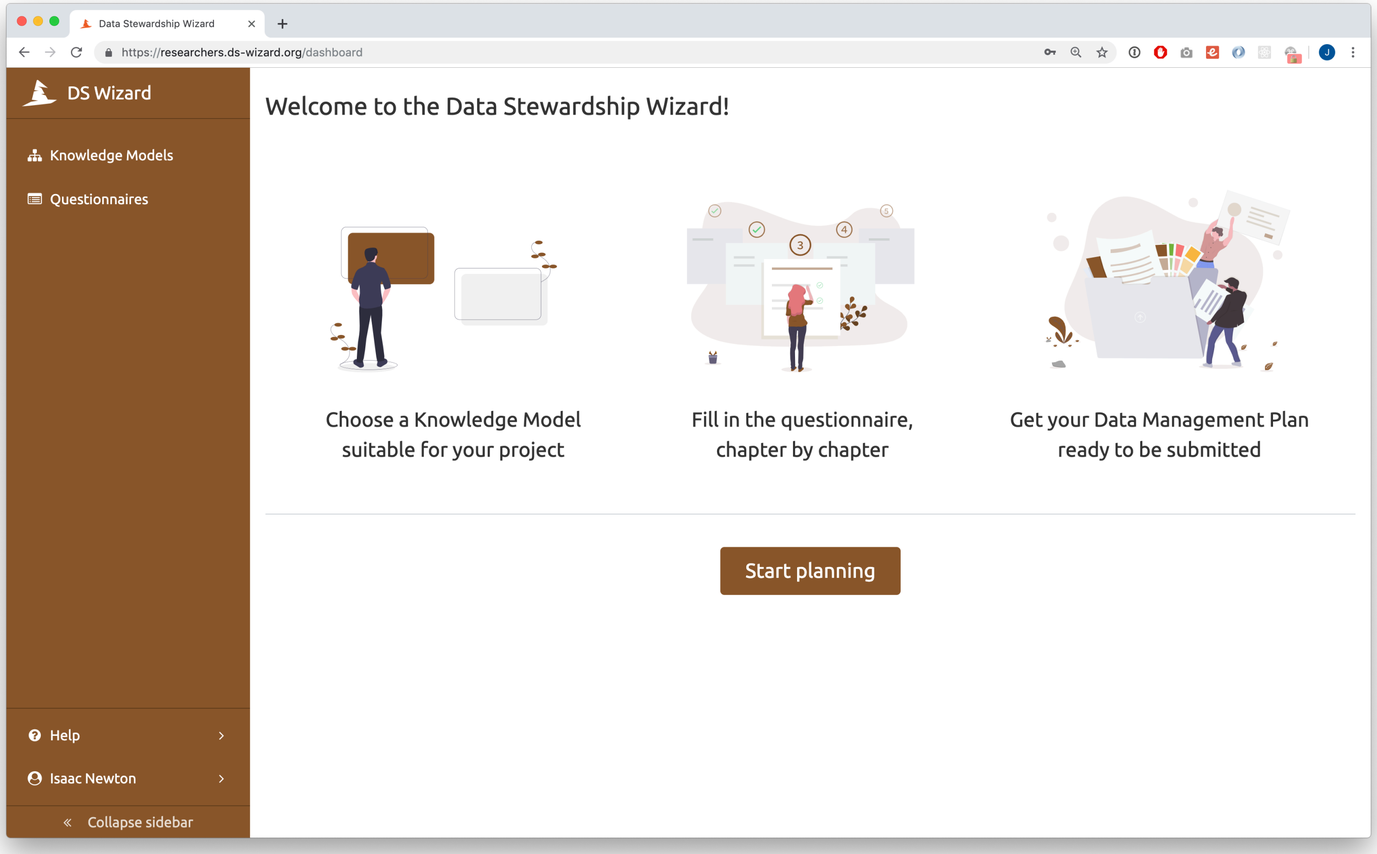
Demo

- demo.ds-wizard.org
- Playground
- Share instance with other users

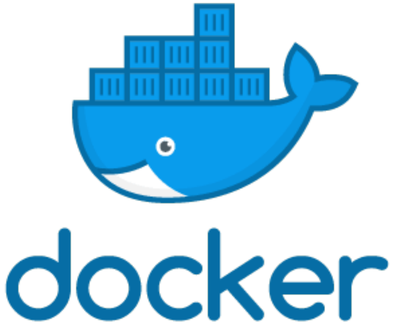


Researchers

- researchers.ds-wizard.org
- For individual researchers
- Easy to sign up and use
- Ready to use Knowledge Model



Self-Managed



- DS Wizard is distributed via Docker
- Documentation: https://docs.ds-wizard.org/
- Docker registry: https://hub.docker.com/u/datastewardshipwizard

DSW Cloud

- 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

Overview

	Demo	Researchers	Self-Managed	DSW Cloud
Production Ready		✔	✔	✔
Questionnaires & DMPs	✔	✔	✔	✔
Knowledge Models Management	✔		✔	✔
Custom Export Document Templates			✔	✔
User Management & Organization Settings			✔	✔
Features configuration & Visual style customizations			✔	✔

Questions & Discussion

Try DS Wizard yourself!

DSW

- ds-wizard.org
- info@ds-wizard.org
- https://twitter.com/dswizard_org