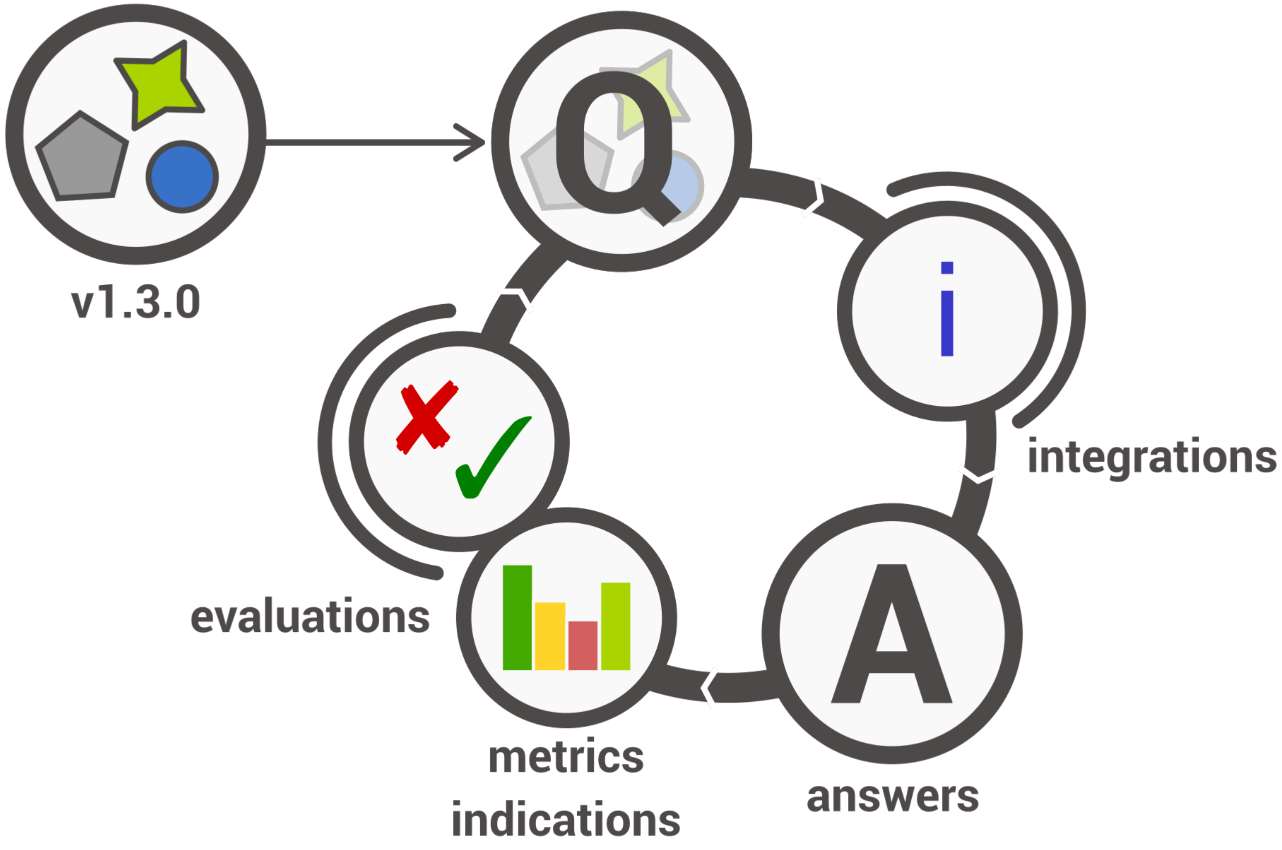
Data Steward Wizard

Korbinian Bösl
Data manager ELIXIR Norway
21 February 2020

DSW

DATA STEWARDSHIP WIZARD



v1.3.0

integrations

evaluations

metrics indications

answers

elixir-no.ds-wizard.org

Developed by

FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

DTL

GO FAIR

ÚOCHB IOCB PRAGUE

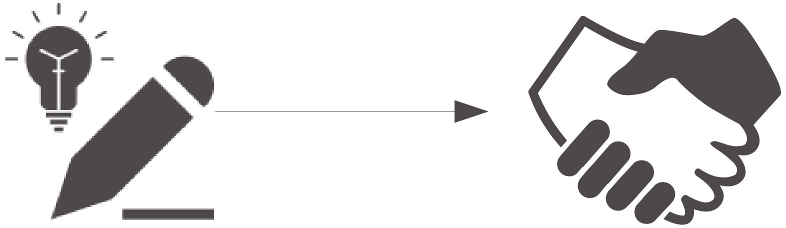
elixir NETHERLANDS

elixir CZECH REPUBLIC

ds-wizard.org
github.com/ds-wizard

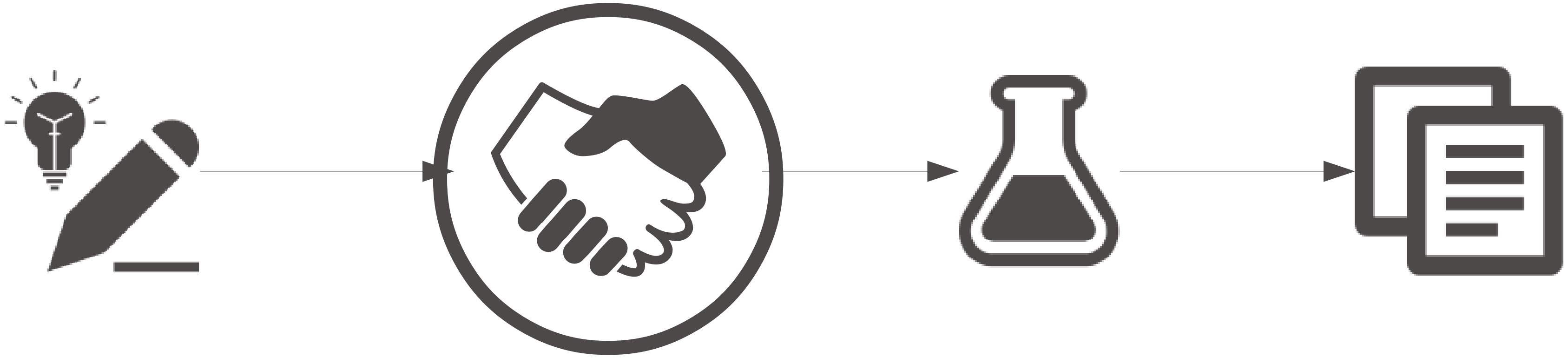
Forskningsrådet

Forskningsrådet



Forskningsrådet

Forskningsrådet



DMP

SCIENCE
EUROPE

Forskningsrådet

- Description & Re-use
- Documentation & Data quality
- Storage & Backups
- Legal & Ethic requirements
- Sharing & Preservation
- Responsibilities & Resources (FAIR principles)

SCIENCE EUROPE

https://www.scienceeurope.org/media/jezkhnoo/se_rdm_practical_guide_final.pdf

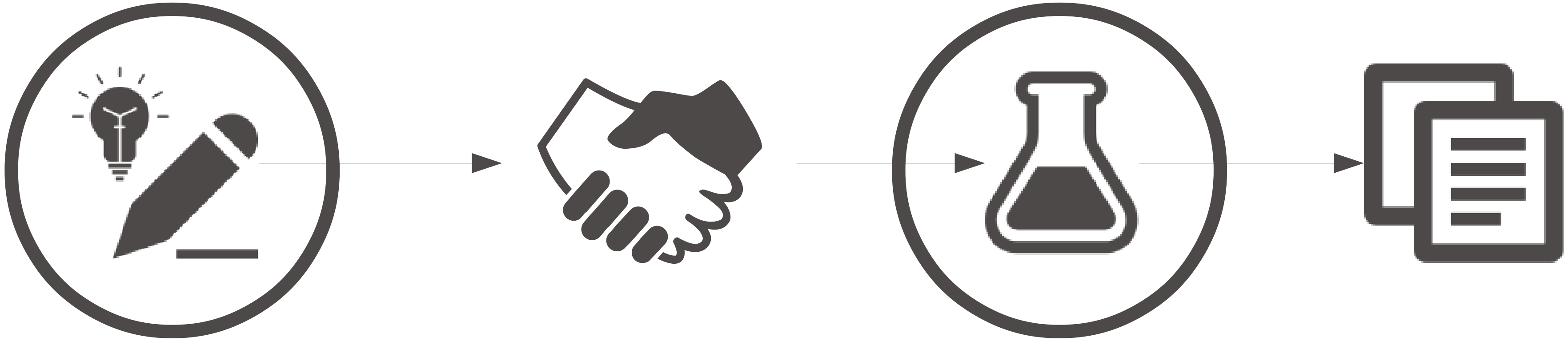
https://www.scienceeurope.org/media/nsxdyvqn/se_guidance_document_rdmps.pdf

https://www.forskningsradet.no/en/forskningspolitisk-radgivning/open-science/open-access-to-research-data/

European
Commission

Horizon 2020 → Horizon Europe

Open Research Data Pilot (ORD pilot)
Default since 2017 → mandatory 2021



DMP
outline

DMP

- After 6 month
- Update upon significant changes

European
Commission

Horizon 2020 → Horizon Europe

- Data summary
- FAIR data implementation
- Resources
- Data security
- Ethics

https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

DSW

DATA STEWARDSHIP WIZARD

Filled questionnaire → Template → DMP in various formats

once | per funding body | .docx, .tex, .html, .json

Adapted for Norwegian users

compatible

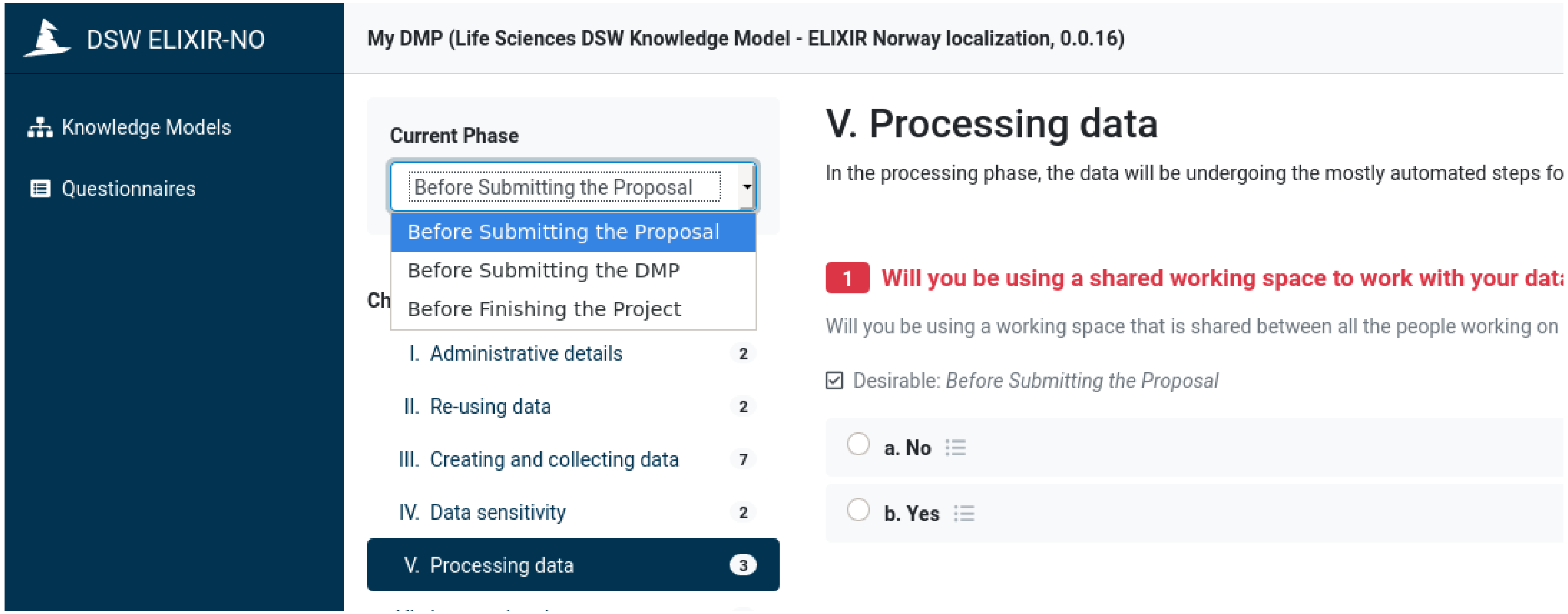
SCIENCE EUROPE

elixir-no.ds-wizard.org

ds-wizard.org
github.com/ds-wizard

DSW

DATA STEWARDSHIP WIZARD



DSW ELIXIR-NO

Knowledge Models

Questionnaires

My DMP (Life Sciences DSW Knowledge Model - ELIXIR Norway localization, 0.0.16)

Current Phase

Before Submitting the Proposal

- Before Submitting the Proposal
- Before Submitting the DMP
- Before Finishing the Project

Ch

I. Administrative details	2
II. Re-using data	2
III. Creating and collecting data	7
IV. Data sensitivity	2
V. Processing data	3

V. Processing data

In the processing phase, the data will be undergoing the mostly automated steps fo

1 Will you be using a shared working space to work with your dat

Will you be using a working space that is shared between all the people working on

☑ Desirable: *Before Submitting the Proposal*

○ a. No

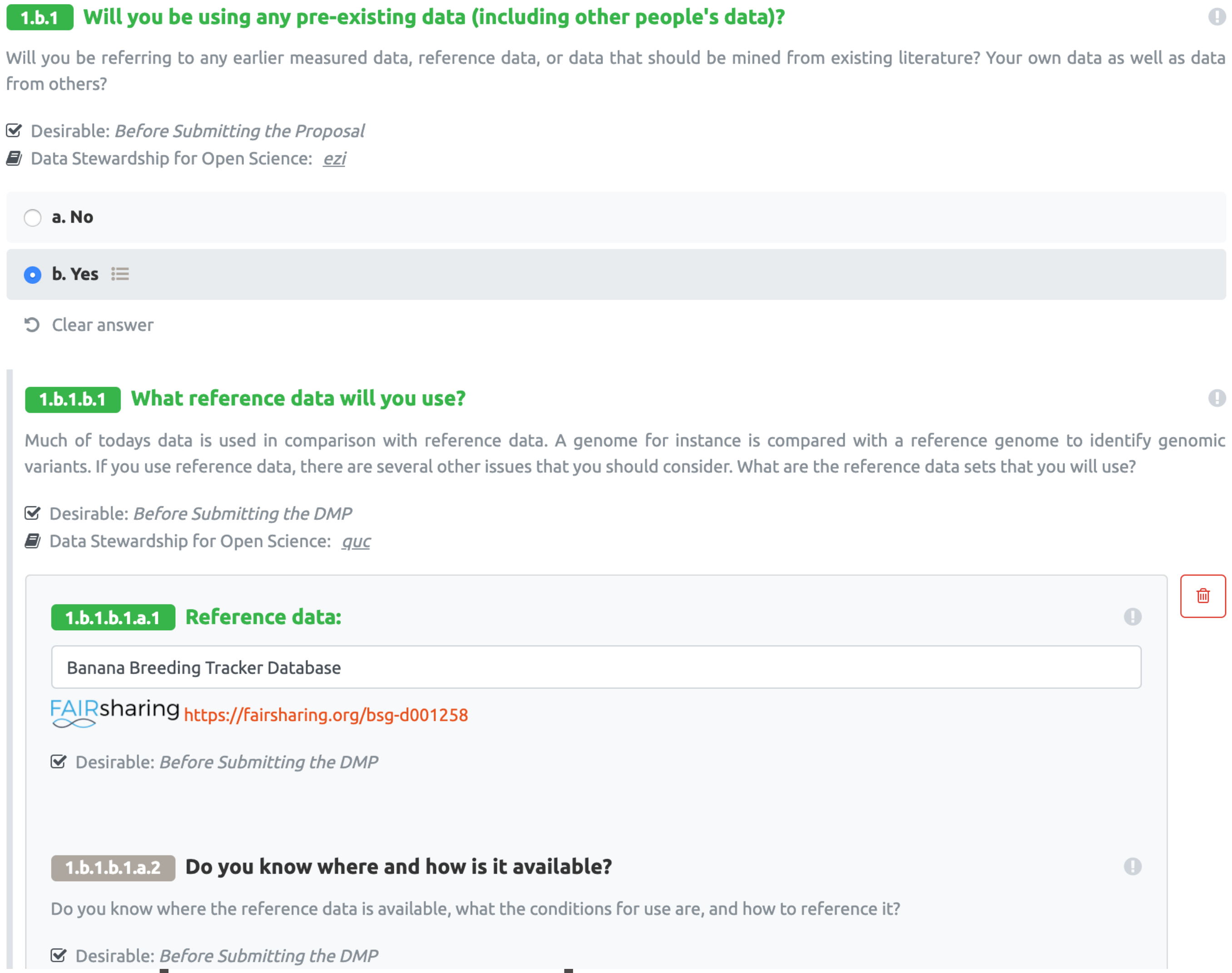
○ b. Yes

elixir-no.ds-wizard.org

ds-wizard.org
github.com/ds-wizard

DSW

DATA STEWARDSHIP WIZARD



1.b.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: *ezi*

○ **a. No**

● **b. Yes**

Clear answer

1.b.1.b.1 **What reference data will you use?**

Much of todays data is used in comparison with reference data. A genome for instance is compared with a reference genome to identify genomic variants. If you use reference data, there are several other issues that you should consider. What are the reference data sets that you will use?

☑ Desirable: *Before Submitting the DMP*

Data Stewardship for Open Science: *quc*

1.b.1.b.1.a.1 **Reference data:**

Banana Breeding Tracker Database

FAIRsharing https://fairsharing.org/bsg-d001258

☑ Desirable: *Before Submitting the DMP*

1.b.1.b.1.a.2 **Do you know where and how is it available?**

Do you know where the reference data is available, what the conditions for use are, and how to reference it?

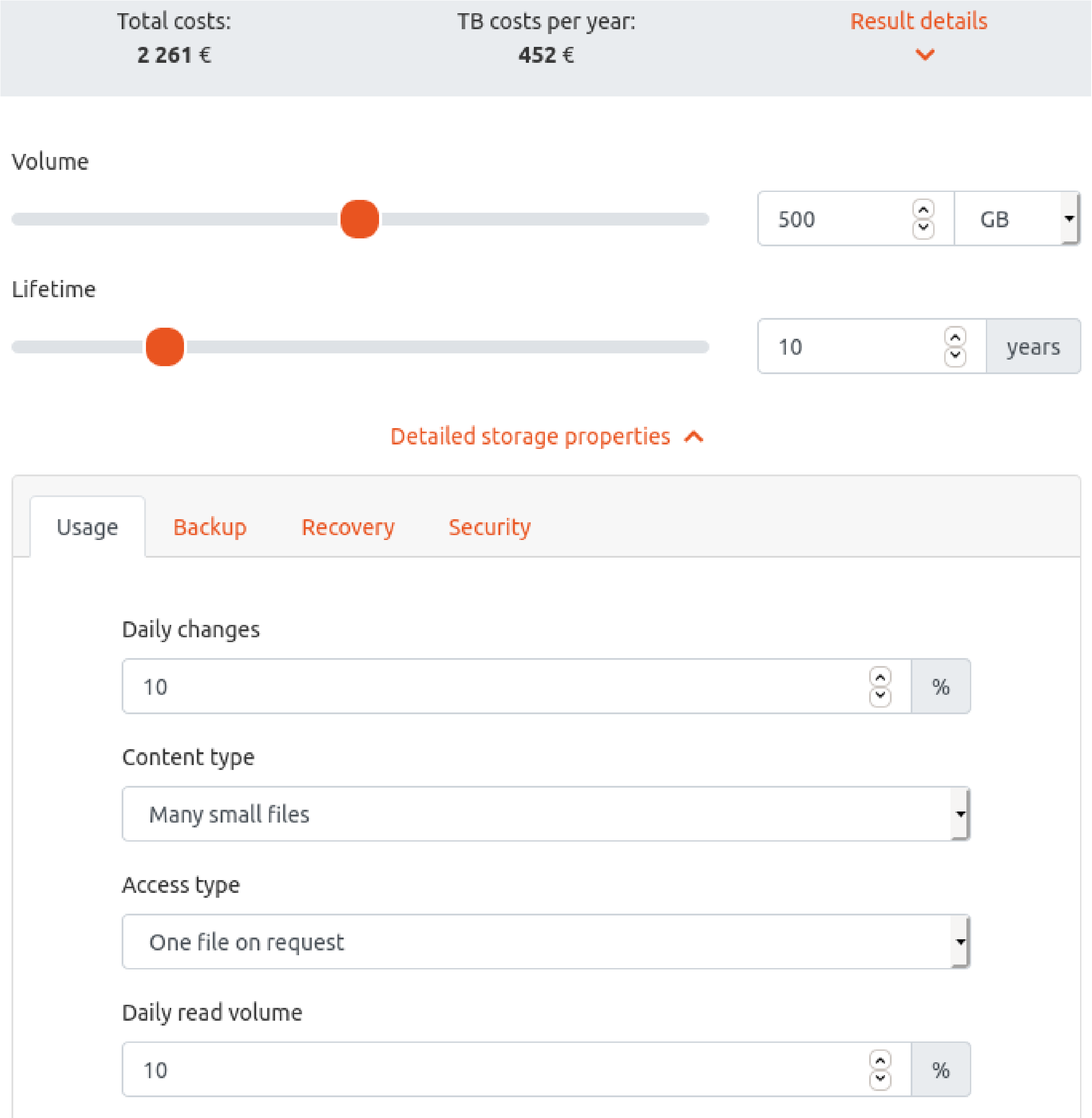
☑ Desirable: *Before Submitting the DMP*

elixir-no.ds-wizard.org

ds-wizard.org
github.com/ds-wizard

DSW

DATA STEWARDSHIP WIZARD



Total costs: 2 261 €

TB costs per year: 452 €

Volume: 500 GB

Lifetime: 10 years

Daily changes: 10 %

Content type: Many small files

Access type: One file on request

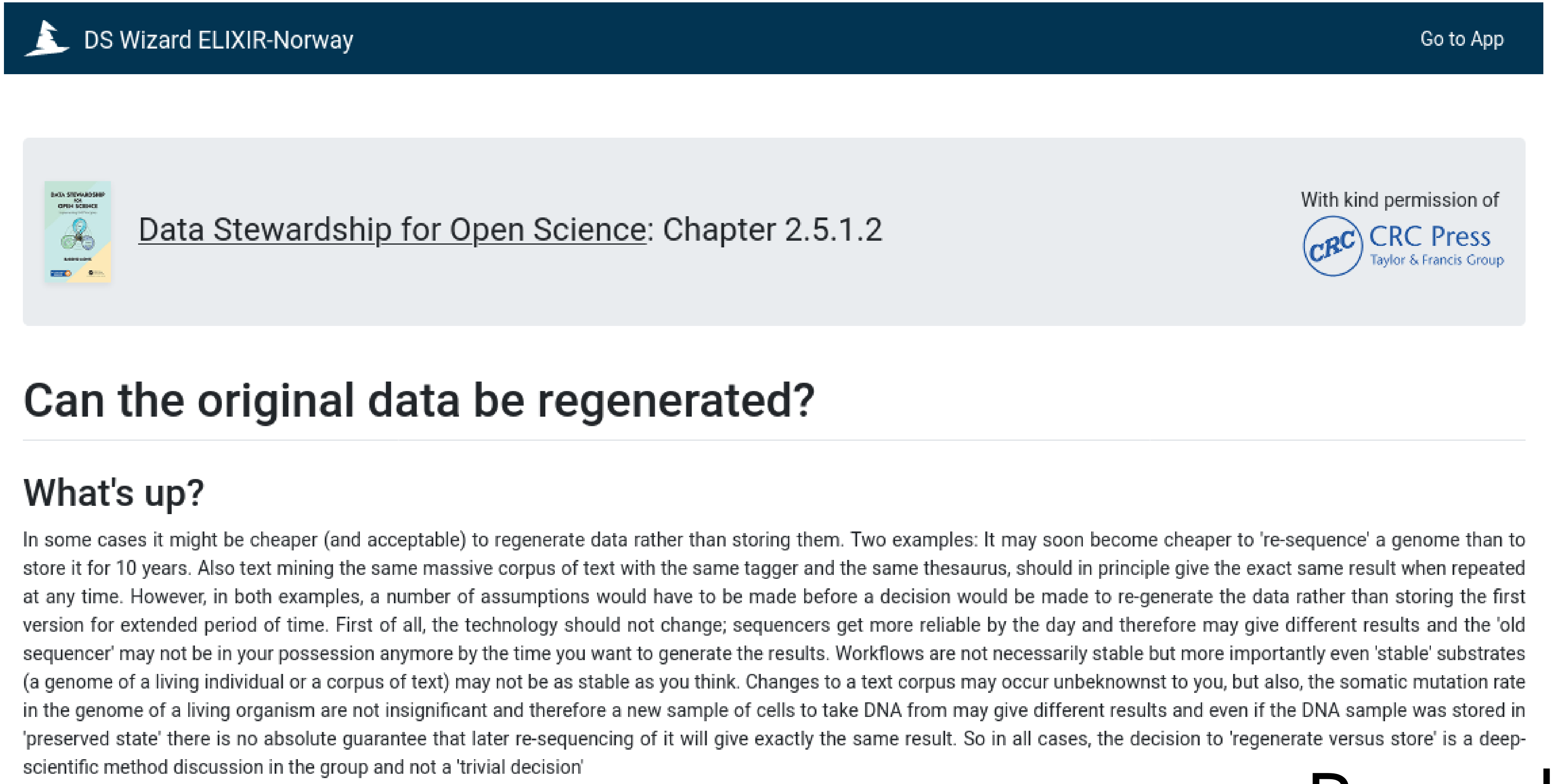
Daily read volume: 10 %

elixir-no.ds-wizard.org

ds-wizard.org
github.com/ds-wizard

DSW

DATA STEWARDSHIP WIZARD



Data Stewardship for Open Science: Chapter 2.5.1.2

With kind permission of CRC Press Taylor & Francis Group

Can the original data be regenerated?

What's up?

In some cases it might be cheaper (and acceptable) to regenerate data rather than storing them. Two examples: It may soon become cheaper to 're-sequence' a genome than to store it for 10 years. Also text mining the same massive corpus of text with the same tagger and the same thesaurus, should in principle give the exact same result when repeated at any time. However, in both examples, a number of assumptions would have to be made before a decision would be made to re-generate the data rather than storing the first version for extended period of time. First of all, the technology should not change; sequencers get more reliable by the day and therefore may give different results and the 'old sequencer' may not be in your possession anymore by the time you want to generate the results. Workflows are not necessarily stable but more importantly even 'stable' substrates (a genome of a living individual or a corpus of text) may not be as stable as you think. Changes to a text corpus may occur unbeknownst to you, but also, the somatic mutation rate in the genome of a living organism are not insignificant and therefore a new sample of cells to take DNA from may give different results and even if the DNA sample was stored in 'preserved state' there is no absolute guarantee that later re-sequencing of it will give exactly the same result. So in all cases, the decision to 'regenerate versus store' is a deep-scientific method discussion in the group and not a 'trivial decision'

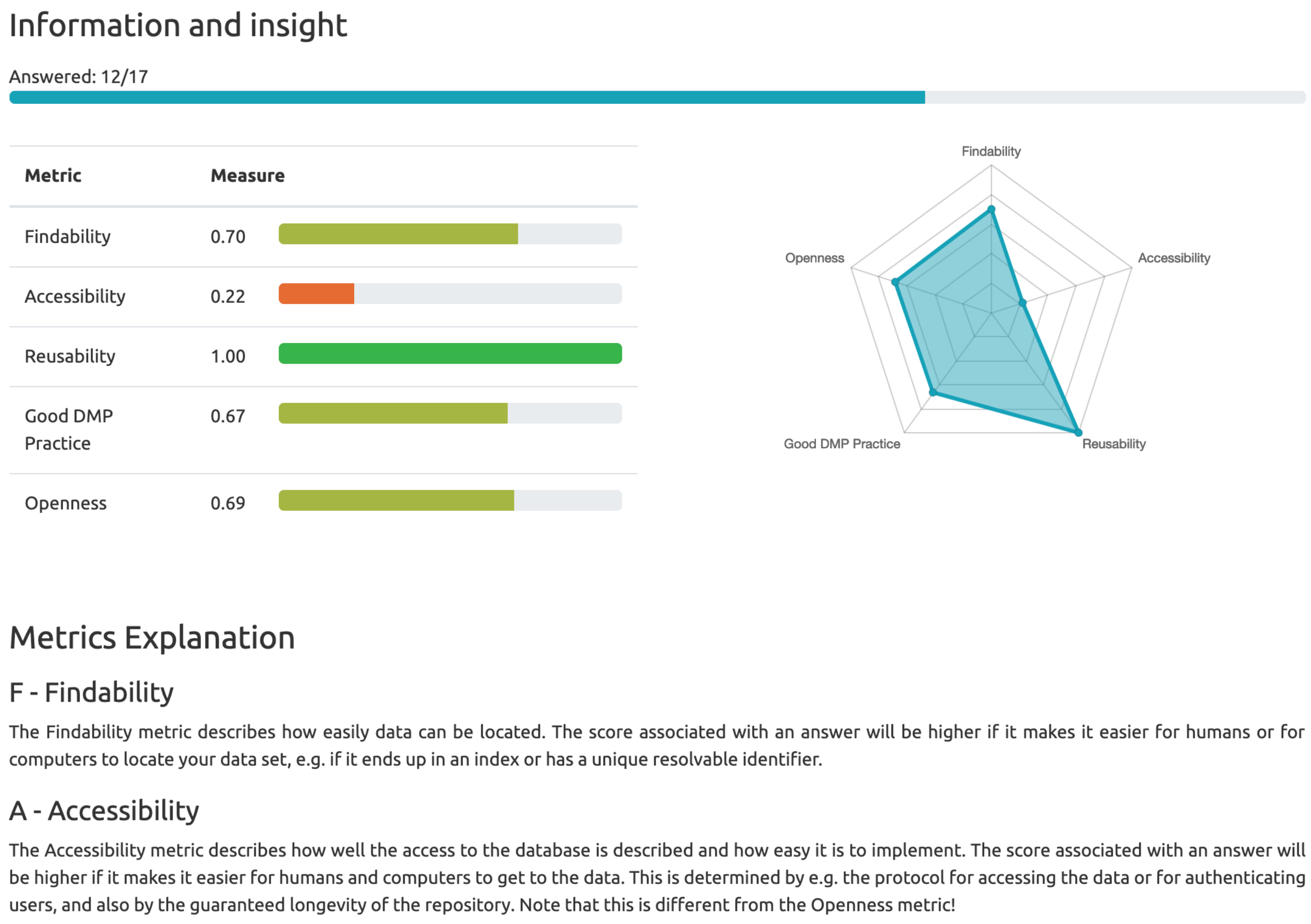
elixir-no.ds-wizard.org

Barend Mons

ds-wizard.org
github.com/ds-wizard

DSW

DATA STEWARDSHIP WIZARD



Information and insight

Answered: 12/17

Metric	Measure
Findability	0.70
Accessibility	0.22
Reusability	1.00
Good DMP Practice	0.67
Openness	0.69

Metrics Explanation

F - Findability

The Findability metric describes how easily data can be located. The score associated with an answer will be higher if it makes it easier for humans or for computers to locate your data set, e.g. if it ends up in an index or has a unique resolvable identifier.

A - Accessibility

The Accessibility metric describes how well the access to the database is described and how easy it is to implement. The score associated with an answer will be higher if it makes it easier for humans and computers to get to the data. This is determined by e.g. the protocol for accessing the data or for authenticating users, and also by the guaranteed longevity of the repository. Note that this is different from the Openness metric!

elixir-no.ds-wizard.org

ds-wizard.org
github.com/ds-wizard

elixir-norway.org
digitallifenorway.org

contact@bioinfo.no

@elixirnorway
@DigitaltLiv

European Commission

Horizon 2020
Research Infrastructure
Program

elixir NORWAY

UNIVERSITY OF BERGEN

UiT The Arctic University of Norway

UiO : University of Oslo

With funding from

The Research
Council of Norway

CENTRE FOR
DIGITAL LIFE
NORWAY

NTNU
Norwegian University of
Science and Technology

NMBU
Norwegian University
of Life Sciences

Data Steward Wizard development

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

DTL

GO FAIR

ÚOCHB
IOCB PRAGUE

elixir NETHERLANDS

elixir CZECH REPUBLIC