Konference ELIXIR CZ

Cesta k FAIR Data

Robert Pergl
Marek Suchánek

2.11.2016

FAIR Data

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

elixir CZECH REPUBLIC

FAIR Data

To be **Findable**:

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

Wilkinson, M.D. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018. DOI: 10.1038/sdata.2016.18

FAIR Data

To be **Accessible**:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
  - A1.1 the protocol is open, free, and universally implementable.
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

Wilkinson, M.D. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018. DOI: 10.1038/sdata.2016.18

FAIR Data

To be **Interoperable**:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

Wilkinson, M.D. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018. DOI: 10.1038/sdata.2016.18

FAIR Data

To be **Re-usable**:

- R1. meta(data) have a plurality of accurate and relevant attributes.
  - R1.1. (meta)data are released with a clear and accessible data usage license.
  - R1.2. (meta)data are associated with their provenance.
  - R1.3. (meta)data meet domain-relevant community standards.

Wilkinson, M.D. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018. DOI: 10.1038/sdata.2016.18

FAIR Data Challenges

- Technical
- Organisational and ensuring Data Quality

elixir CZECH REPUBLIC

FAIR Data Challenges: Technical

- H2020 Excelerate WP5 – Interoperability: https://www.elixir-europe.org/excelerate/interoperability
- DTL Projects: http://www.slideshare.net/lolavo/fair-data-overview (next slide)

elixir CZECH REPUBLIC

FAIR Data Challenges: Technical

FAIR Data applications
FAIR Data Point
FAIRifier
Data FAIRport
FAIR Data search engine
…

FAIR Data Protocol
FAIR Profiles
FAIR accessor
API spec
PIDs
LDP
…

FAIR Data Resources
datasets
ontologies
controlled vocabularies
taxonomies
…

http://www.slideshare.net/lolavo/fair-data-overview

elixir CZECH REPUBLIC

FAIR Data Challenges: Organisational and Data Quality

- ELIXIR-CZ: Data stewardship action team (UOCHB + CVUT)
  - Raise local awareness of the data produced and its value
  - Help life science data producers make their data resources FAIR: consultations, BYODs, trainings

elixir CZECH REPUBLIC

http://ccmi.fit.cvut.cz/elixir-questionnaire/

Data stewardship action team

elixir CZECH REPUBLIC

Vision | Action steps | Lifecycle | Data | Roles | Managerial Questionnaire | Technical Questionnaire

Let's dig the gold mine!

Bioinformatics produces a lot of data that is very valuable and that's our gold mine.

In our working group, we realize this value of data. We set ourselves the following goals, by which we want to help BioMed researchers mine their gold:

- Collect and provide the information about bioinformatics data produced.
- Help the data producers to take care about their data (a.k.a. Data Stewardship).
- Help the data producers share their data with others.
- Connect and help interested parties to use the available data sources.

Clip art image by Cliparts.co

Sharing the data sources contents poses challenges regarding technical solutions, organisational set up, security and legal issues. The ultimate goal is making the data F.A.I.R., i.e. *Findable, Accessible, Interoperable, Re-usable* while maintaining all the necessary constraints.

Check our action steps.

FAIR Data Challenges: Organisational and Data Quality

- Ontological Analyses of ELIXIR Core Resources (ELIXIR-NL, ELIXIR-CZ, ELIXIR-SE, ELIXIR-CH, with support of the author of Unified Foundational Ontology Dr. Giancarlo Guizzardi)
  - Initiative to use foundational ontologies to improve the core resources schemas
  - See Martínez Ferrandis, A.M., Pastor López, O., Guizzardi, G., 2013. **Applying the Principles of an Ontology-Based Approach to a Conceptual Schema of Human Genome**, in: Ng, W., Storey, V.C., Trujillo, J.C. (Eds.), Conceptual Modeling. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 471\u2013478.

elixir CZECH REPUBLIC

FAIR Data Challenges: Organisational and Data Quality

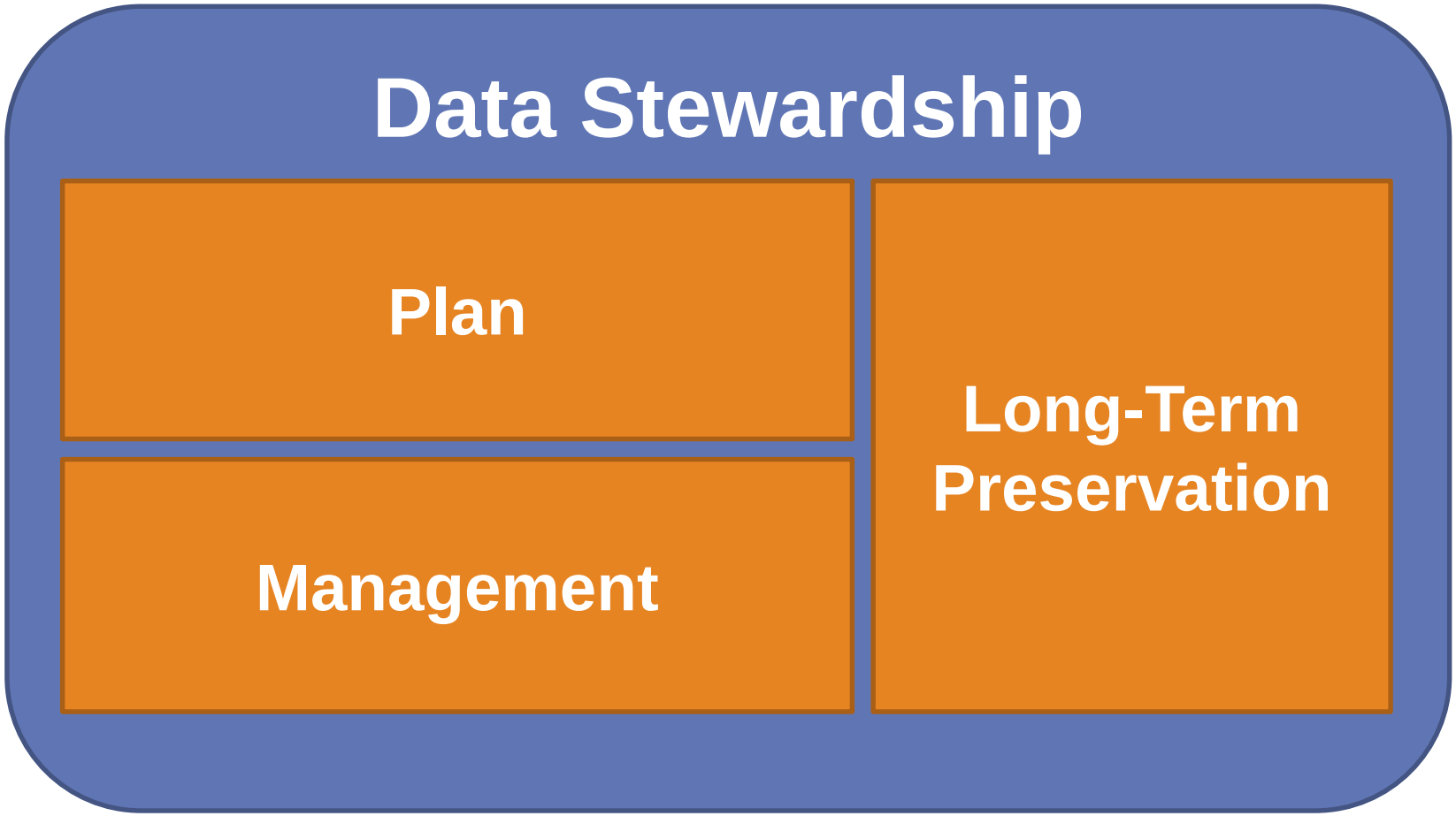
- Data Stewardship Planning Portal (ELIXIR-NL, ELIXIR-CZ, ELIXIR-SE, ELIXIR-SI)
  - Long-term community-driven project
  - First stage: architecture and prototype: joint project **ELIXIR-CZ** (Robert Pergl, Marek Suchánek) and **ELIXIR-NL** (Rob Hooft)

FAIR Data Challenges: Organisational and Data Quality

- Data Stewardship ⊃ Data Management
- „Data Management Plans (DMPs) are a **key element** of good data management."
- Horizon 2020: „A **DMP is required** for all projects participating in the extended ORD pilot, unless they opt out of the ORD pilot. However, projects that opt out are **still encouraged to submit a DMP** on a voluntary basis."
- European Open Science Cloud: Recommendations of the High Level Expert Group

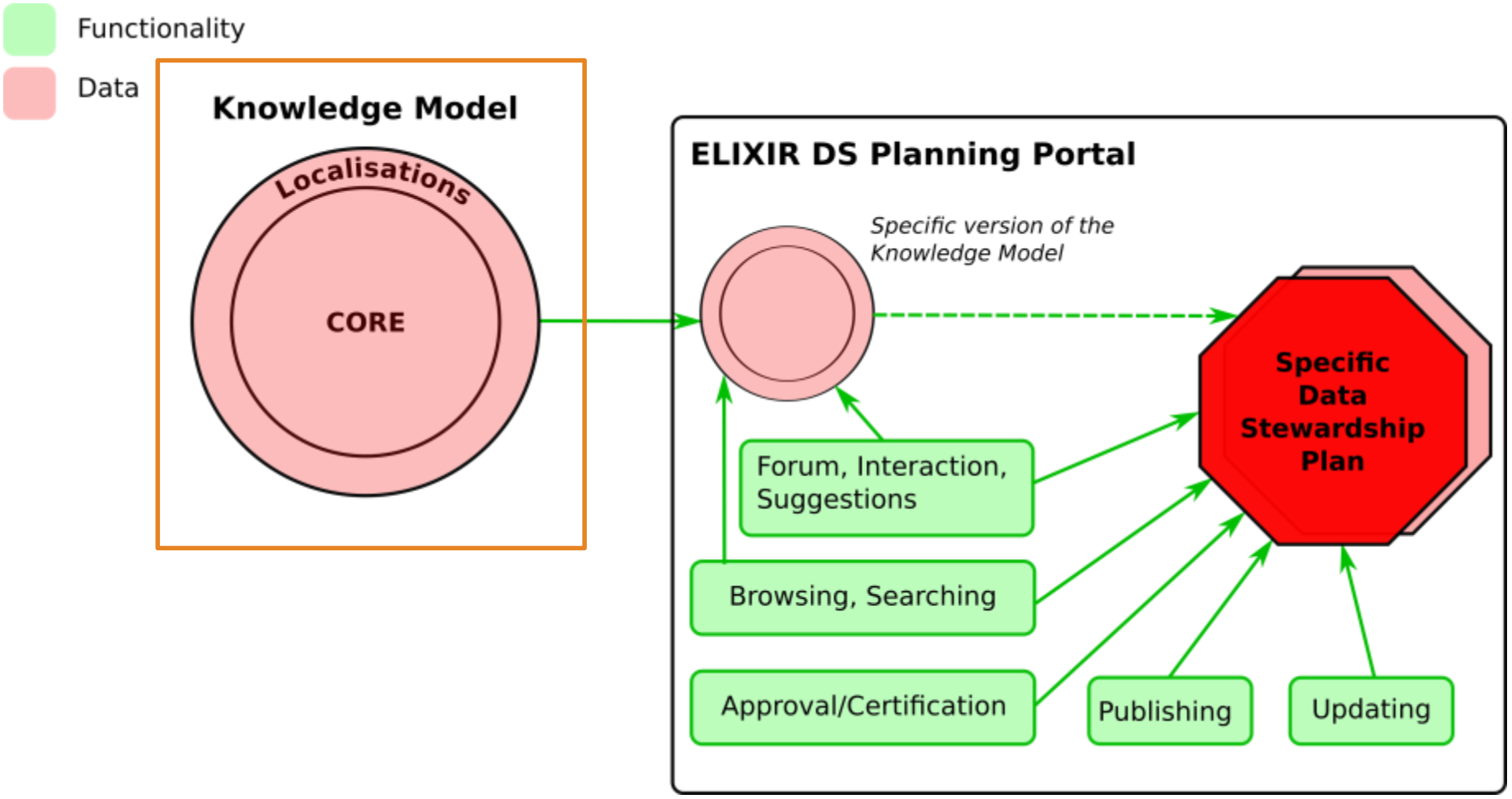
European Commission
Horizon 2020
European Union funding
for Research & Innovation
elixir CZECH REPUBLIC

FAIR Data Challenges: Organisational and Data Quality



elixir CZECH REPUBLIC

Data Stewardship Planning Portal Architecture

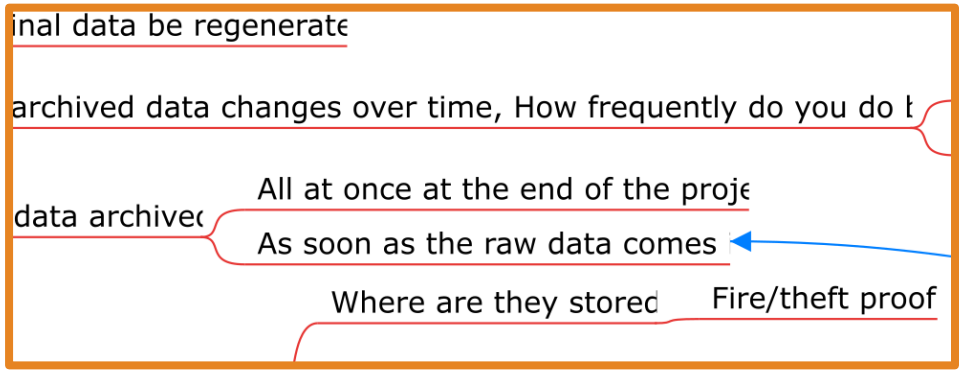


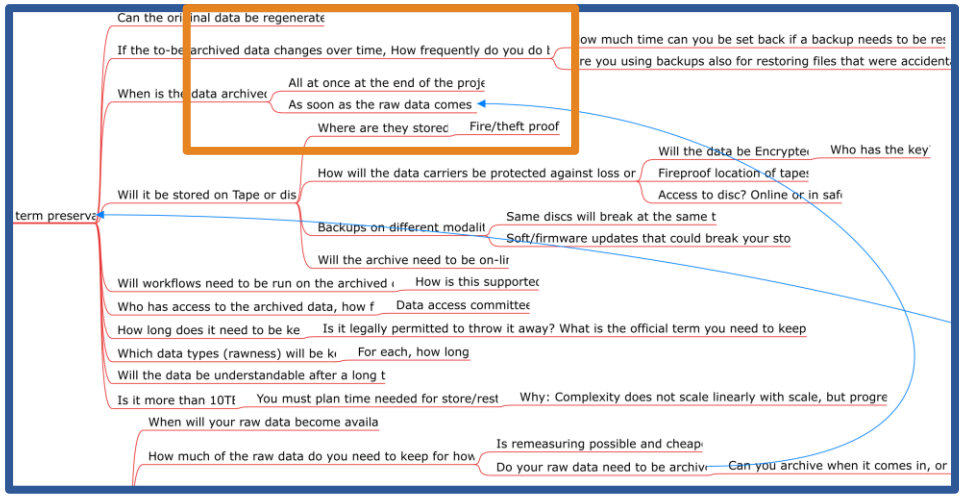
https://f1000research.com/posters/5-2420

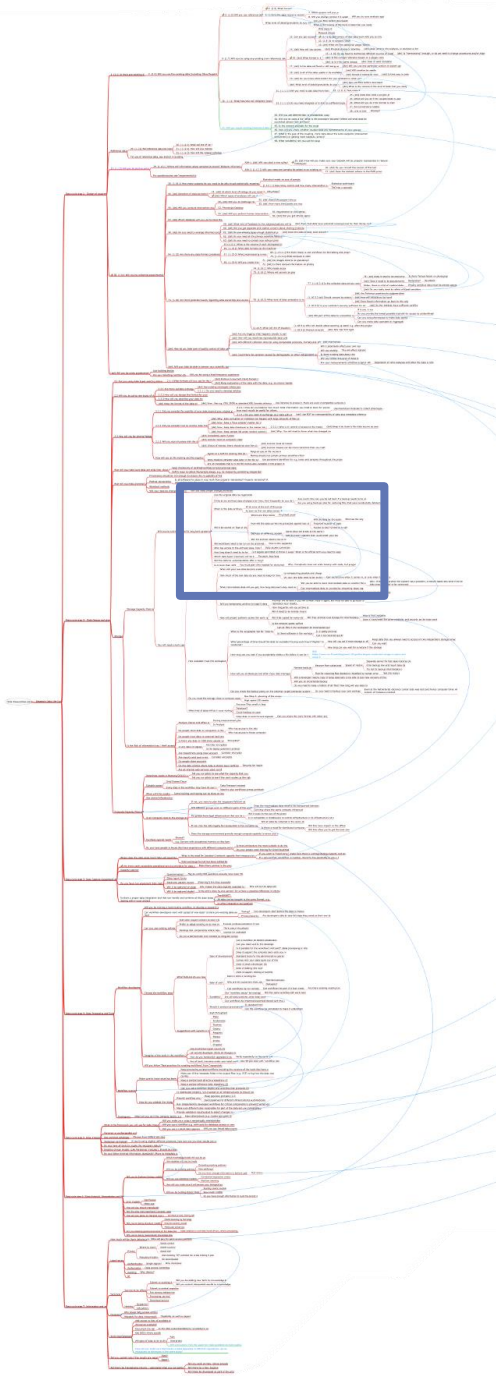
elixir CZECH REPUBLIC

Data Stewardship Knowledge Model (DS-KM)

- Mind map by Rob Hooft (ELIXIR-NL)
  - Over 600 nodes
  - Up to 9 levels
  - Yet just a fragment of all knowledge







elixir CZECH REPUBLIC

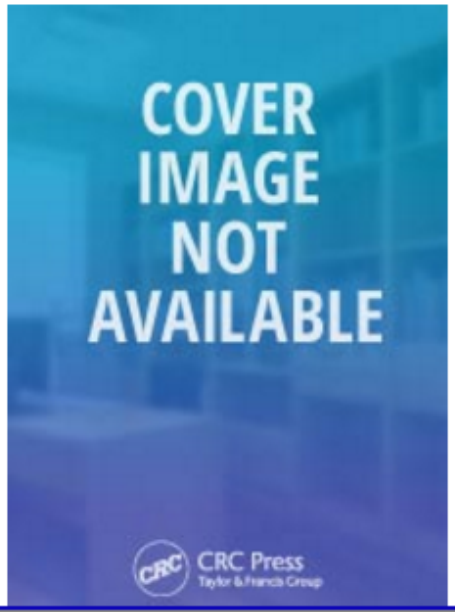
--> Data Stewardship Book

CRC Press
Taylor & Francis Group

Czech Republic

Search or Browse by Subject

Home / Computer Science & Engineering / Data Mining and Knowledge Discovery / Data Stewardship for Discovery: A Practical Guide for Data Experts



Data Stewardship for Discovery: A Practical Guide for Data Experts

Barend Mons

VitalSource eBook access code and instructions will be provided within the print book.

June 15, 2017 **Forthcoming** by Chapman and Hall/CRC
Reference - 200 Pages - 50 B/W Illustrations
ISBN 9781498753173 - CAT# K27337

elixir CZECH REPUBLIC

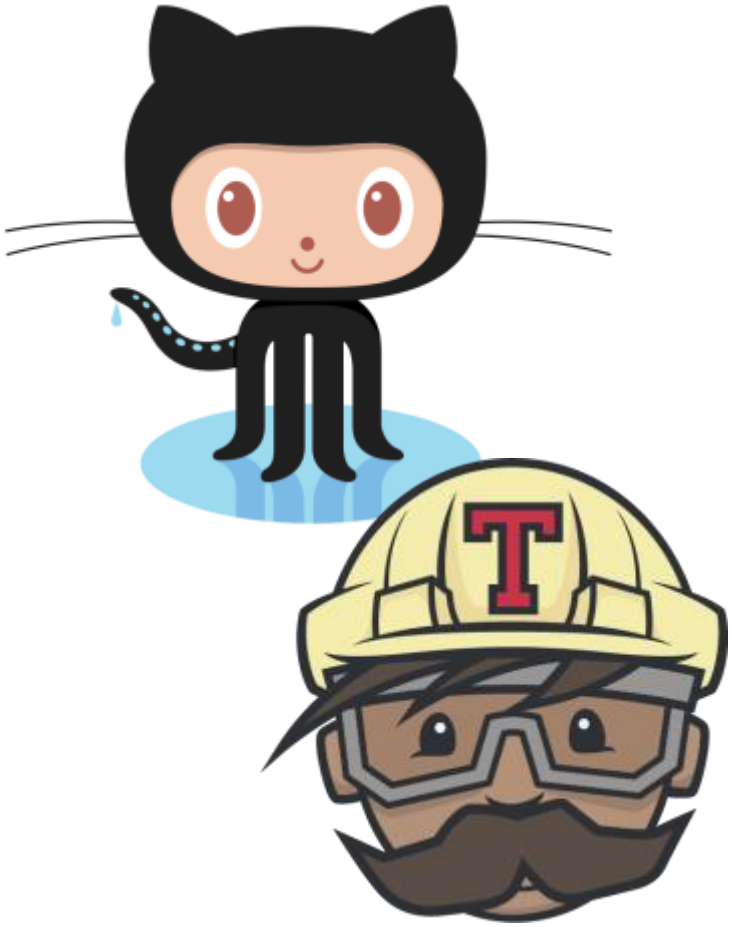
DS-KM – A FAIR Project for FAIR Data

- Mind Map → JSONs
- Core + Localizations
  - National + Domain
  - Extend/Override
  - Preconditions
  - Cross references
  - Reordering
- Include the book contents
- JSON schema (http://json-schema.org)

```
{
    "chapterid": 1,
    "namespace": "core",
    "formatversion": 1,
    "title": "Design of experiment",
    "text": "Before you decide to embark on any new study, it is nowad
    "questions" :
    [
        {
            "questionid": 1, "type": "option",
            "title": "Is there any pre-existing data?",
            "text": "Are there any data sets available in the world th
            "answers":
            [
                {
                    "id": 0, "label": "No",
                    "advice": "You know that this is very unlikely? Th
                },
```

elixir CZECH REPUBLIC

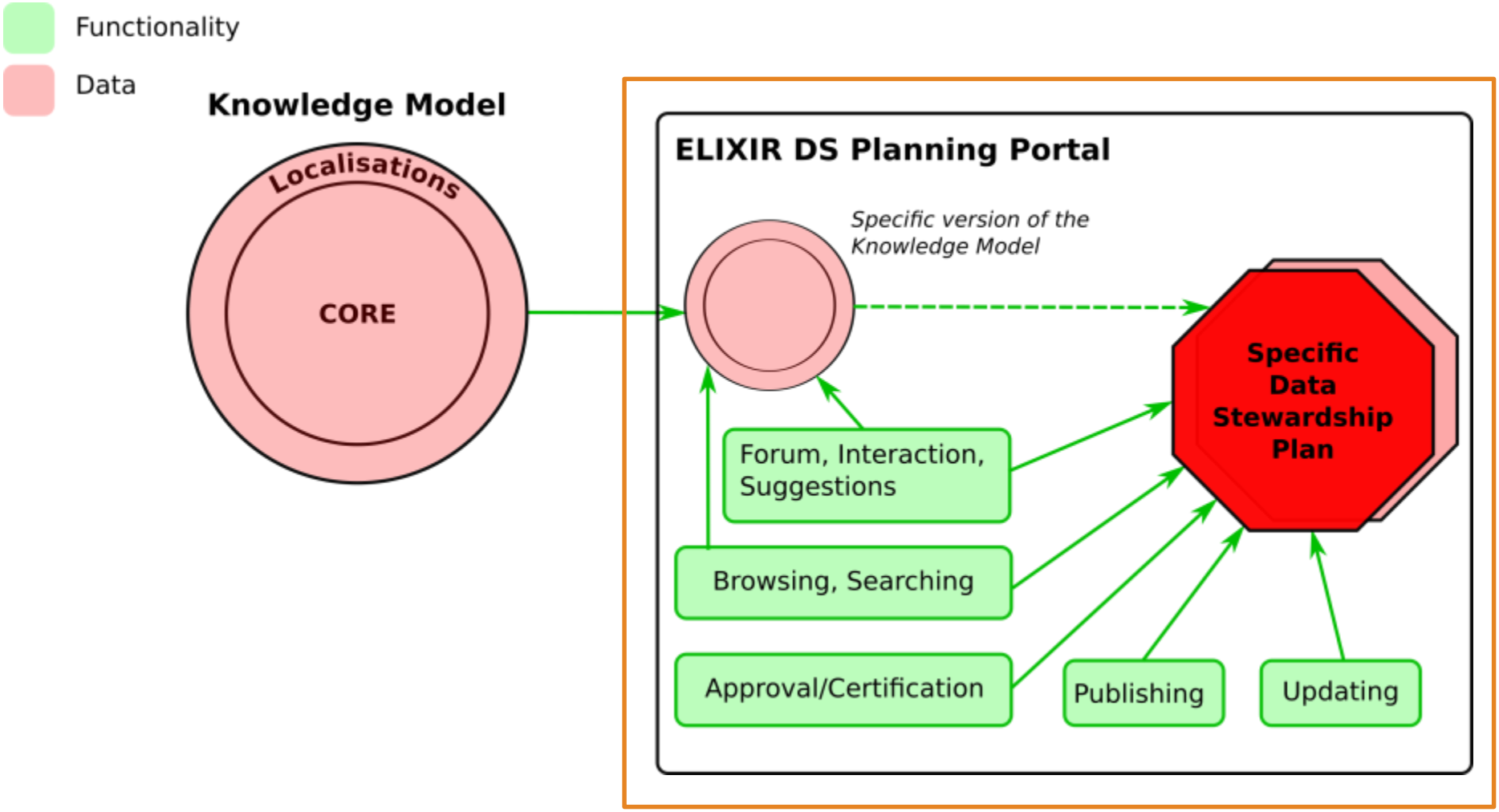
DS-KM – A FAIR Project for FAIR Data



- GitHub
  - https://github.com/CCMi-FIT/ds-km
  - Open source project (Apache 2 License)
  - History of changes (commits), branches
  - Issues, projects, comments, wiki
  - Forks & contribution
- Travis CI
  - Validation of DS KM content
- Precompiler
  - Compile chapter JSONs of core + desired localizations
  - Rules, dependency resolution, restructuralization

elixir CZECH REPUBLIC

Data Stewardship Planning Portal Architecture

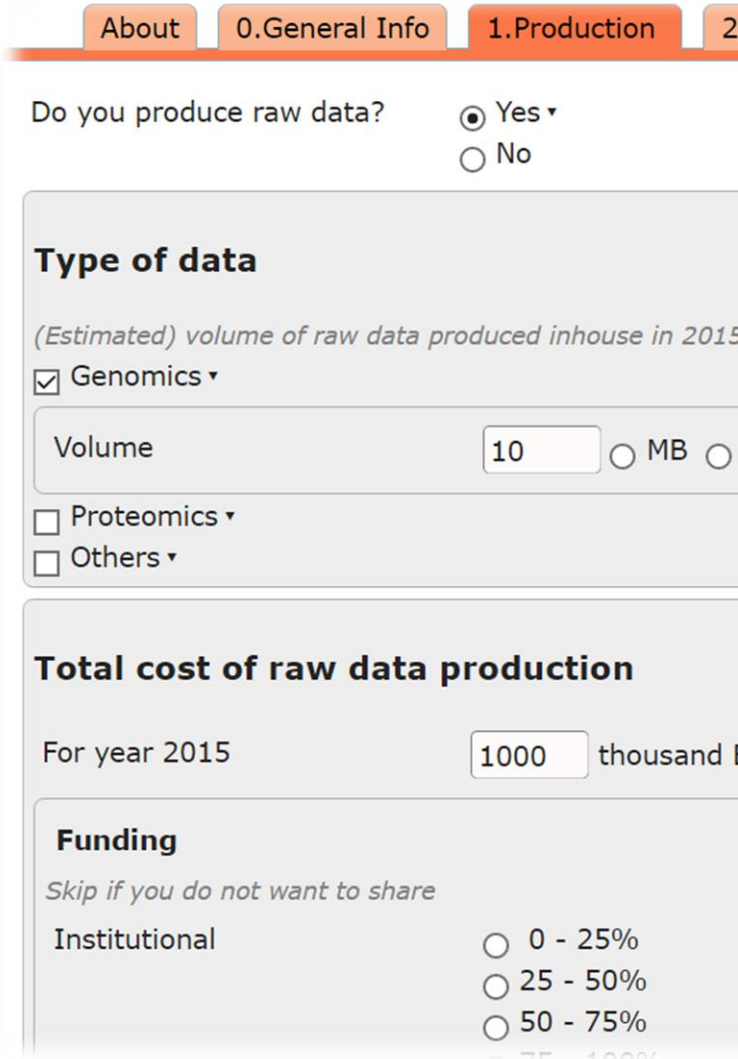


https://f1000research.com/posters/5-2420

elixir CZECH REPUBLIC

DS Wizard

- A prototype for interactive browsing of DS-KM (questionnaire)
  - Guiding through all the questions
  - Web application (Haskell GHC+Haste)
- The wizard is the basis for helping you with creating data management plans.
- Future: all the remaining functionality ☺



Roads to Rome (http://roadstorome.moovellab.com)

elixir CZECH REPUBLIC

Invitation

- Interested in our project?
- Want to become **FAIR**?

- Contact us:
  perglr@fit.cvut.cz
  suchama4@fit.cvut.cz

- Come to **ELIXIR All Hands 2017** & visit our workshop there!

elixir CZECH REPUBLIC

Konference uspořádána za podpory MŠMT, projekt velké infrastruktury ELIXIR-CZ (Grant LM2015047).

elixir CZECH REPUBLIC