Side Event: Towards international RDM harmonization in agricultural research

Embedding FAIRagro into the international agricultural RDM community

Funded by
Deutsche Forschungsgemeinschaft
German Research Foundation
project number 501899475
In cooperation with
nfdi

June 18, 2024

Agenda

11:00 - 11:10	**Welcome**	Xenia Specka, Nikolai Svoboda
11:10 - 11:25	**CGIAR's GARDIAN ecosystem: Enabling data-driven insights**	Medha Devare
11:25 - 11:40	**Internationalisation of plant research by de.NBI and ELIXIR**	Sebastian Beier
11:40 - 12:05	**FAIRagro impulses on (planned/used) standards, tools and outcomes** 1. Metadata and PID Concepts and Ontology Usage in FAIRagro 2. A technical glimpse into the FAIRagro Middleware 3. Building the FAIRagro search portal	 Daniel Martini Daniel Arend Julian Schneider
12:05 - 12:25	**Discussion on cross-cutting topics between FAIRagro, CGIAR, de.NBI & ELIXIR** How can we collaborate and use synergies?	all (Moderation: Nikolai Svoboda)
12:25 - 12:30	**Wrap up** Activities, milestones for 2024/2025 and responsible persons	Xenia Specka

Welcome to this session

Objectives

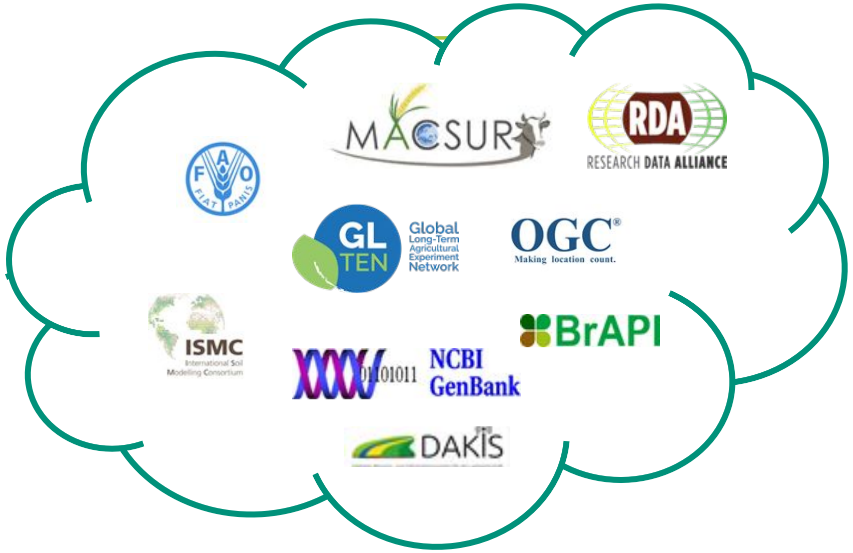
Bring together representatives from international organizations and FAIRagro

Getting to know each other and each other's work

Identification of topics for future collaboration and planning the next steps

BY

FAIRagro internationalization strategy



- **FAIRagro partners have many international contacts**
  with potential cooperation with FAIRagro
- **Internationalization strategy** → aims to coordinate those efforts
- **Priorities** given to:
  - WUR / Wageningen Data Competence Center (WDCC)
  - CGIAR Consultative Group on International Agricultural Research
  - AgMIP → Agricultural Model Intercomparison and Improvement Project

Wageningen Data Competence Center

- Data Steward Service Center (DSSC)
  - Data stewardship
  - Institutional data management
- Workshop planned 2024/2025

CGIAR
de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE
elixir
AgMIP

cc BY

CGIAR's GARDIAN ecosystem: Enabling data-driven insights

Medha Devare

CGIAR

Internationalisation of plant research by de.NBI and ELIXIR

Sebastian Beier

FAIRagro

FAIRagro impulse 1

Metadata and PID Concepts and Ontology Usage in FAIRagro

Daniel Martini

cc BY

Improving FAIRness

To be Findable:

🟨 F1. (meta)data are assigned a globally unique and persistent identifier

⬜ F2. data are described with rich metadata (defined by R1 below)

🟨 F3. metadata clearly and explicitly include the identifier of the data it describes

F4. (meta)data are registered or indexed in a searchable resource

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

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 ⬜

To be Reusable:

🟦 R1. meta(data) are richly described with 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 detailed provenance ⬜

🟦 R1.3. (meta)data meet domain-relevant community standards ⬜

⬜ Metadata

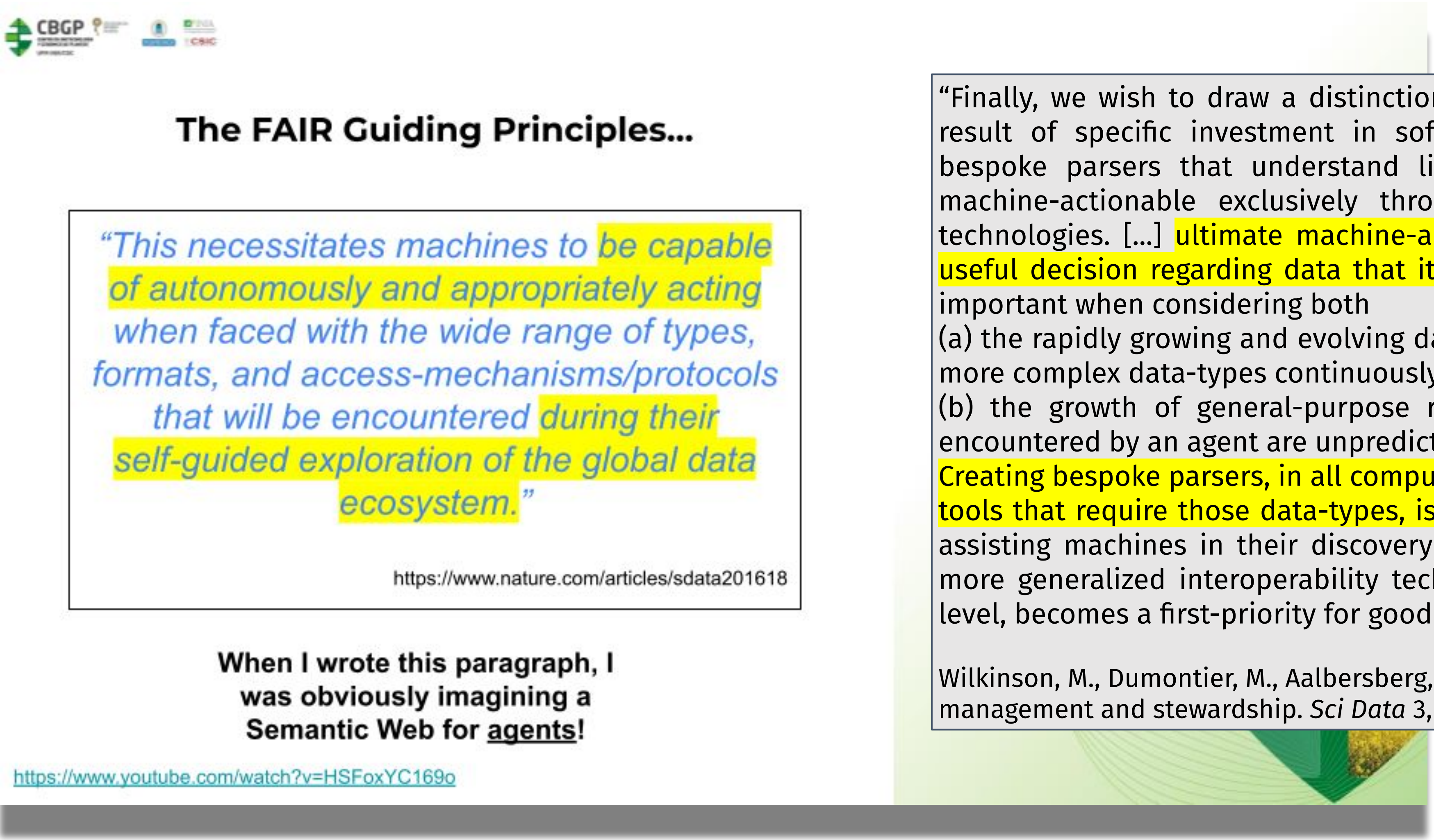
🟨 PIDs

🟦 Ontologies

Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016).
see also: https://go-fair.org/principles

cc BY

Machine-Actionability



The FAIR Guiding Principles...

> "This necessitates machines to be capable of autonomously and appropriately acting when faced with the wide range of types, formats, and access-mechanisms/protocols that will be encountered during their self-guided exploration of the global data ecosystem."
>
> https://www.nature.com/articles/sdata201618

When I wrote this paragraph, I was obviously imagining a Semantic Web for agents!

https://www.youtube.com/watch?v=HSFoxYC169o

"Finally, we wish to draw a distinction between data that is machine-actionable as a result of specific investment in software supporting that data-type, for example, bespoke parsers that understand life science wwPDB files [...], and data that is machine-actionable exclusively through the utilization of general-purpose, open technologies. [...] ultimate machine-actionability occurs when a machine can make a useful decision regarding data that it has not encountered before. This distinction is important when considering both

(a) the rapidly growing and evolving data environment, with new technologies and new, more complex data-types continuously being developed, and

(b) the growth of general-purpose repositories, where the data-types likely to be encountered by an agent are unpredictable.

Creating bespoke parsers, in all computer languages, for all data-types and all analytical tools that require those data-types, is not a sustainable activity. As such, the focus on assisting machines in their discovery and exploration of data through application of more generalized interoperability technologies and standards at the data/repository level, becomes a first-priority for good data stewardship."

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Mark Wilkinson, CoRDI talk slide 23:
https://docs.google.com/presentation/d/1WJxGnfoAXbuKj23OE_1Qgl9cqksoq7Vlpfy4hT90JmQ/edit

Metadata

- Library Science:
  "Metadata is bibliographic data (author, title, abstract…)"
- Broadly accepted:
  "Metadata is data about data (data format, download URL…)"

Neither gives us machine-actionability

Metadata describes data and
is expressed in a *metalanguage*
that allows for *metaprogramming*
to process and interpret the data it is about

"how to say things, not what to say"

addresses "Creating bespoke parsers…is not a sustainable activity"

cc BY

PIDs

F1. (Meta)data are assigned a globally unique and persistent identifier

I2. (Meta)data use vocabularies that follow FAIR principles

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

What do we need to identify?

Anything that we want to make statements about in metadata:

- datasets
- publications
- samples
- observed variables/traits
- columns and records *in* datasets
- classes and datatypes
- attributes and relations in metadata formats

taking ultimate machine-actionability serious...

- ...the identifier has to convey the information which protocol to use, so that a machine can determine that on its own
- it thus has to be read as: “...are retrievable by their identifier *and their identifier alone*...”
  → **dereferentiability**

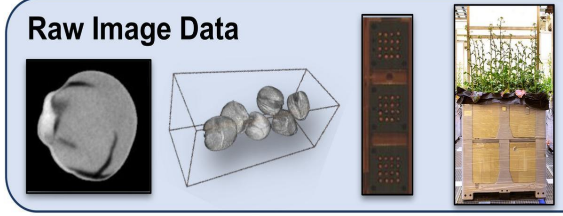
the only feasible option for specifying PIDs currently are URIs (as specified by RFCs 3986 and 8820)

Digital Twins in plant research data ecosystem

FAIRagro

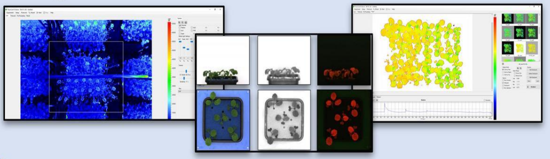
NPZ INNOVATION
Wir forschen für Qualität.

Raw Image Data



Spectrum: images taken at visible light, static fluorescence, near-infrared wavelengths, NMR images, CT images
Angles: top, several side views

Image-Derived Traits

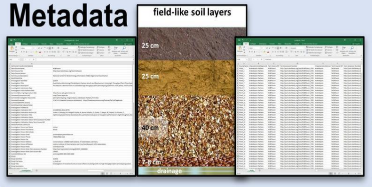


Architecture: plant height, projected leaf area, leaf angles, growth rate
Color: average leaf hue, green to brown ratio, variance in leaf color
Intensity: static fluorescence, near-infrared emitted radiation

Environmental Data

Shoot environment: air temperature, humidity, light intensity, CO_2 concentration
Root environment: soil temperature, water content, nutrition levels, pH

Metadata



isatab

miappe

Plant: species, genotype, seed origin
Conditions: soil and container type, watering regime, experiment location
Measurements: observation units, measurement methods, sensor types

(Arend et al. 2022 The Plant Journal; DOI: 10.1111/tpj.15804)

Sample flow

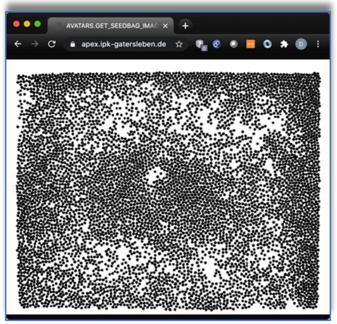
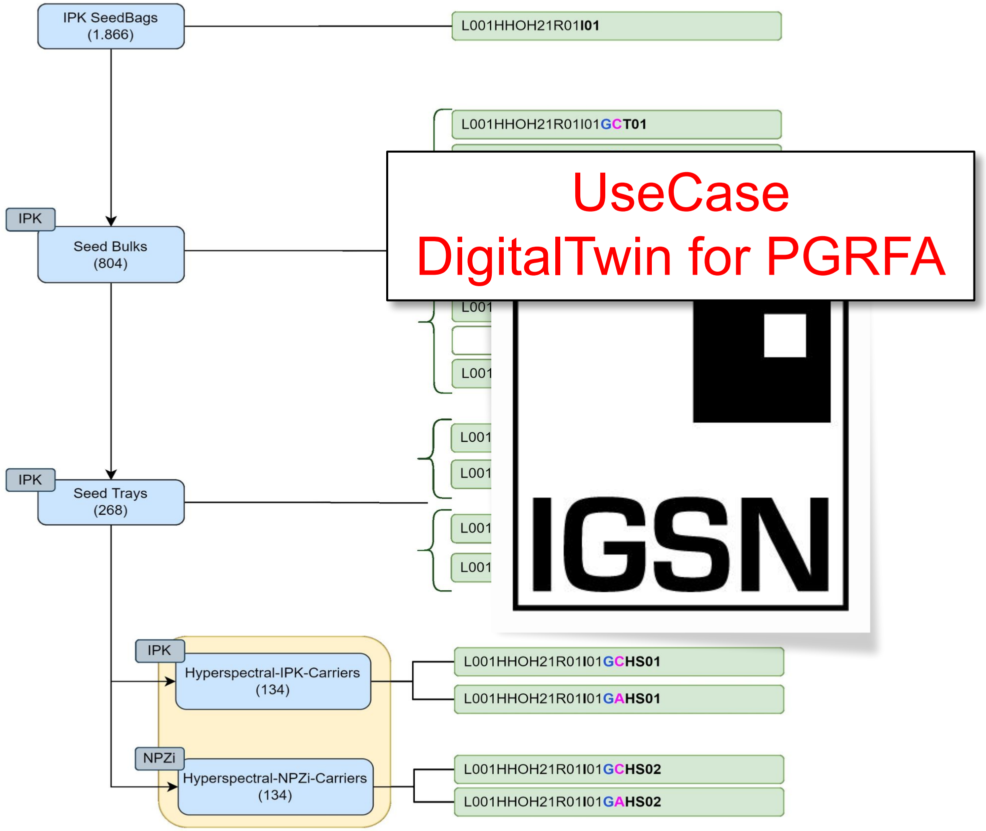


image source: AVATARS project IPK, NPZi

Digital Twins



UseCase
DigitalTwin for PGRFA

C004HMOO20R06I01ST01

genotype* | *field plot* | *season* | *physical object

(Rey-Mazon, NPZi; Plant 2030 status seminar; 2023)

CC BY

Inventory of Standards

Compile an inventory of all meaningful data standards
for representing (meta)data in agrosystems research

- Generic Ontologies and Vocabularies:
  - SSN/SOSA
  - PROV
  - ODRL
  - DQV
- Domain Specific Ontologies and Vocabularies:
  - Crop Ontology
  - AGROVOC
- Geospatial Data and Metadata:
  - INSPIRE
  - ISO19115
  - GML
- Plant Phenotyping Data and Metadata
  - MIAPPE
  - ISA-Tab
- Modeling Data:
  - ICASA
  - AgMIP
- Protocol Standards:
  - OGC WMS/WFS
  - BrAPI
  - OAI-PMH
- Informal Terminologies/Codesystems:
  - EPPO
  - Pesticide Registration Database
- …

➜ different domain specificity

➜ different interoperability level:
syntax, semantics, protocol standards

➜ different level of formalization:
simple term list vs. full-fledged ontologies

FAIR and Standardization

R1.3. (Meta)data meet domain-relevant community standards

➡ This is ***not*** a call for standardization!

for something to be “domain-relevant”, it has to exist for quite some time already

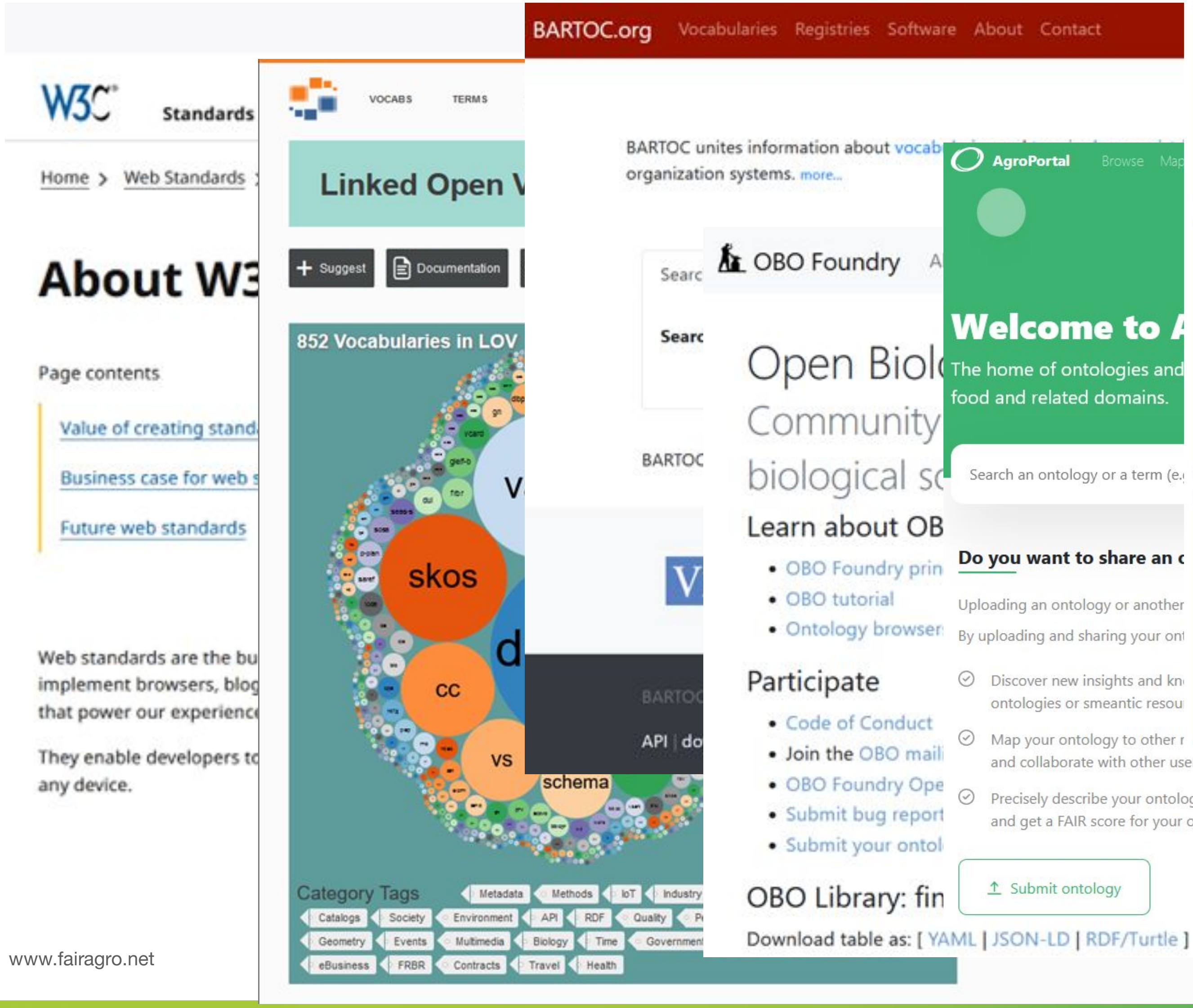
> “...application of more generalized interoperability technologies and standards at the data/repository level, becomes a first-priority for good data stewardship.”
>
> ...
>
> “...when community-endorsed vocabularies or other (meta)data standards do not include the attributes necessary to achieve rich annotation, there are two possible solutions: either publish an extension of an existing, closely related vocabulary, or—in the extreme case—create and explicitly publish a new vocabulary resource, following FAIR principles (‘I2’).”
>
> Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

➡ it is a call for using and applying standards...

- ...that exhibit certain formal properties / characteristics
- ...that do mostly already exist

1 cc BY

Ontology Resources Awareness



W3C Standards

Home > Web Standards >

About W3

Page contents

- Value of creating stand
- Business case for web s
- Future web standards

Web standards are the bu implement browsers, blog that power our experience

They enable developers to any device.

VOCABS TERMS

Linked Open V

+ Suggest | Documentation

852 Vocabularies in LOV

Category Tags

Metadata | Methods | IoT | Industry | Catalogs | Society | Environment | API | RDF | Quality | Geometry | Events | Multimedia | Biology | Time | Government | eBusiness | FRBR | Contracts | Travel | Health

BARTOC.org Vocabularies Registries Software About Contact

BARTOC unites information about vocab organization systems. more...

Searc

Searc

BARTOC

BARTOC

API | do

OBO Foundry

Open Biol Community biological sc

Learn about OB

- OBO Foundry prin
- OBO tutorial
- Ontology browser

Participate

- Code of Conduct
- Join the OBO mail
- OBO Foundry Ope
- Submit bug report
- Submit your ontol

OBO Library: fin

Download table as: [YAML | JSON-LD | RDF/Turtle]

AgroPortal Browse Map

Welcome to A

The home of ontologies and food and related domains.

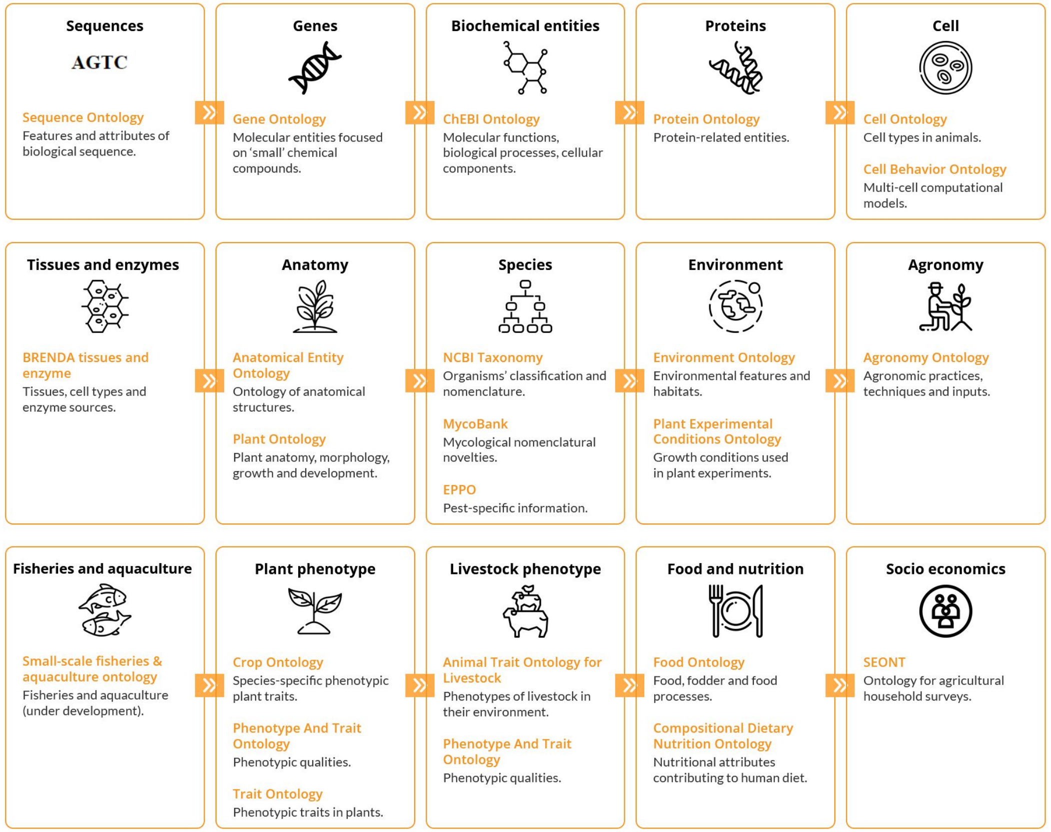
Search an ontology or a term (e.

Do you want to share an o

Uploading an ontology or another By uploading and sharing your ont

- Discover new insights and kn ontologies or smeantic resou
- Map your ontology to other r and collaborate with other users.
- Precisely describe your ontology with relevant metadata and get a FAIR score for your ontology.

Submit ontology



Sequences
AGTC
Sequence Ontology
Features and attributes of biological sequence.

Genes
Gene Ontology
Molecular entities focused on 'small' chemical compounds.

Biochemical entities
ChEBI Ontology
Molecular functions, biological processes, cellular components.

Proteins
Protein Ontology
Protein-related entities.

Cell
Cell Ontology
Cell types in animals.
Cell Behavior Ontology
Multi-cell computational models.

Tissues and enzymes
BRENDA tissues and enzyme
Tissues, cell types and enzyme sources.

Anatomy
Anatomical Entity Ontology
Ontology of anatomical structures.
Plant Ontology
Plant anatomy, morphology, growth and development.

Species
NCBI Taxonomy
Organisms' classification and nomenclature.
MycoBank
Mycological nomenclatural novelties.
EPPO
Pest-specific information.

Environment
Environment Ontology
Environmental features and habitats.
Plant Experimental Conditions Ontology
Growth conditions used in plant experiments.

Agronomy
Agronomy Ontology
Agronomic practices, techniques and inputs.

Fisheries and aquaculture
Small-scale fisheries & aquaculture ontology
Fisheries and aquaculture (under development).

Plant phenotype
Crop Ontology
Species-specific phenotypic plant traits.
Phenotype And Trait Ontology
Phenotypic qualities.
Trait Ontology
Phenotypic traits in plants.

Livestock phenotype
Animal Trait Ontology for Livestock
Phenotypes of livestock in their environment.
Phenotype And Trait Ontology
Phenotypic qualities.

Food and nutrition
Food Ontology
Food, fodder and food processes.
Compositional Dietary Nutrition Ontology
Nutritional attributes contributing to human diet.

Socio economics
SEONT
Ontology for agricultural household surveys.

https://www.w3.org
https://lov.linkeddata.es
https://bartoc.org/
https://obofoundry.org/
https://agroportal.lirmm.fr/
https://bigdata.cgiar.org/ontologies-for-agriculture/
...

Ontology Usage in FAIRagro = Ontology Reuse

Why?

User requirements:

- “I want to see where data came from”
- “I want to see how data was generated”
- “I need to communicate usage restrictions”
- “we need additional domain specific keywords to find what we are looking for”
- “I want to search in variable descriptions”
- “I need to use data from different sources”

How-to *practically*…

- …make use of all the ontology richness?
- …derive from existing ontology terms?
- …program against data that is not known at program development time?
- …”retro-fit” this on legacy infrastructure?
- …build user interfaces in such a setting?

Approach:

- schema.org for warming up with RDF
- minimum metadata profiles as starters for implementation
- ODRL and PROV as case examples:
  how-to modularly combine vocabularies
- reuse design patterns as recipes for extending existing ontologies instead of reinventing from scratch:
  `rdfs:subClassOf`, `rdfs:subPropertyOf`, `skos:broader`…
- mappings for converting legacy vocabularies into ontology representations and for converting different knowledge representations (OBO, OWL, RDF, SKOS…)
- alignments for metadata “translation”

FAIRagro impulse 2

A technical glimpse into the FAIRagro Middleware

Daniel Arend

CC BY

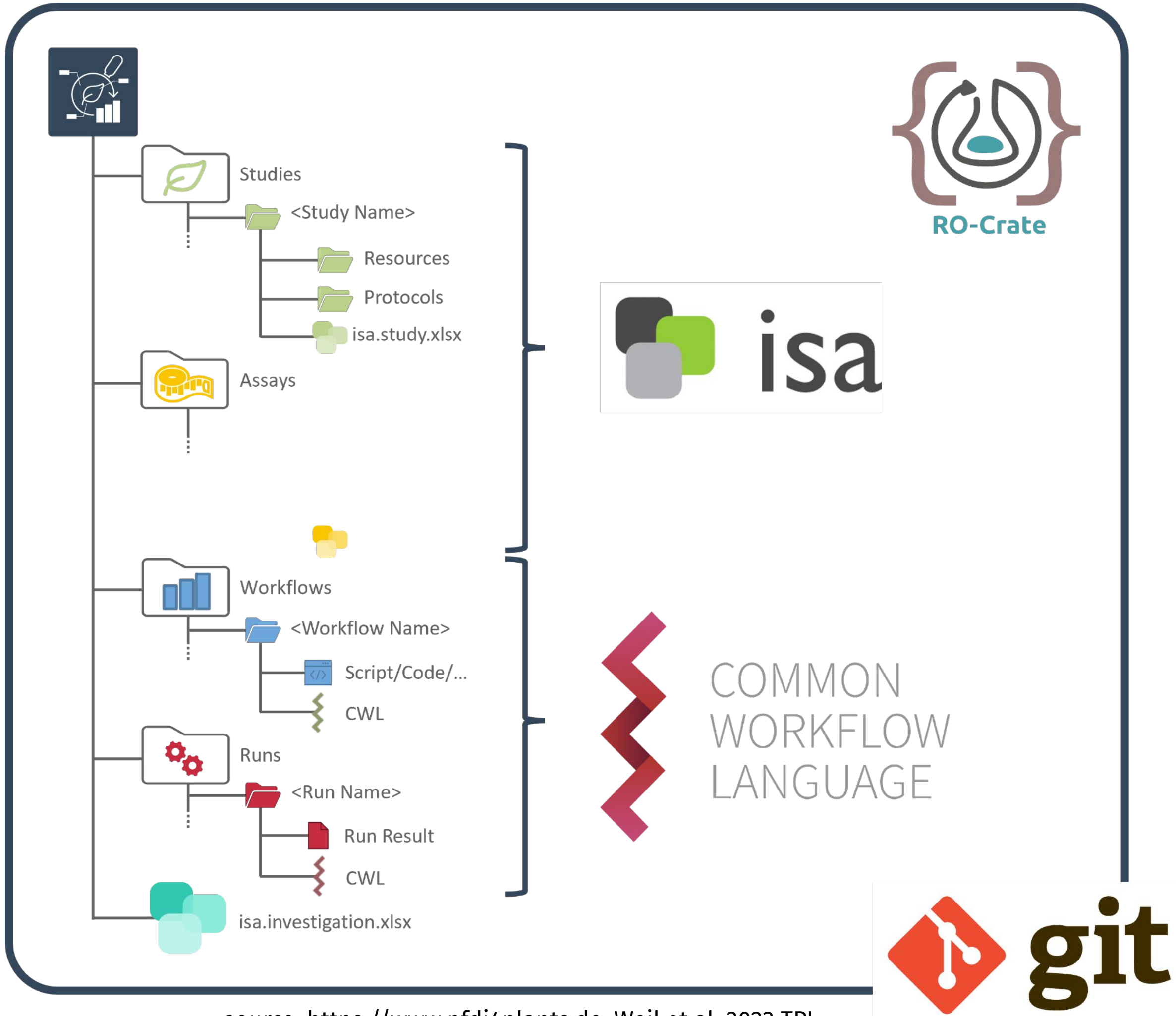
Middleware Approach

- ⇨ two-step implementation (see presentation on Monday)
- ⇨ based on Schema.org/BioSchema metadata & FAIR Digital Objects (FDOs) → basis for AI-Readiness
- ⇨ adapt concept of ARCs (Annotated Research Context)
- ⇨ initially designed by DataPLANT
- ⇨ also in discussion/adoption by other consortia (NFDI4Bioimage, NFDI4Biodiversity...)

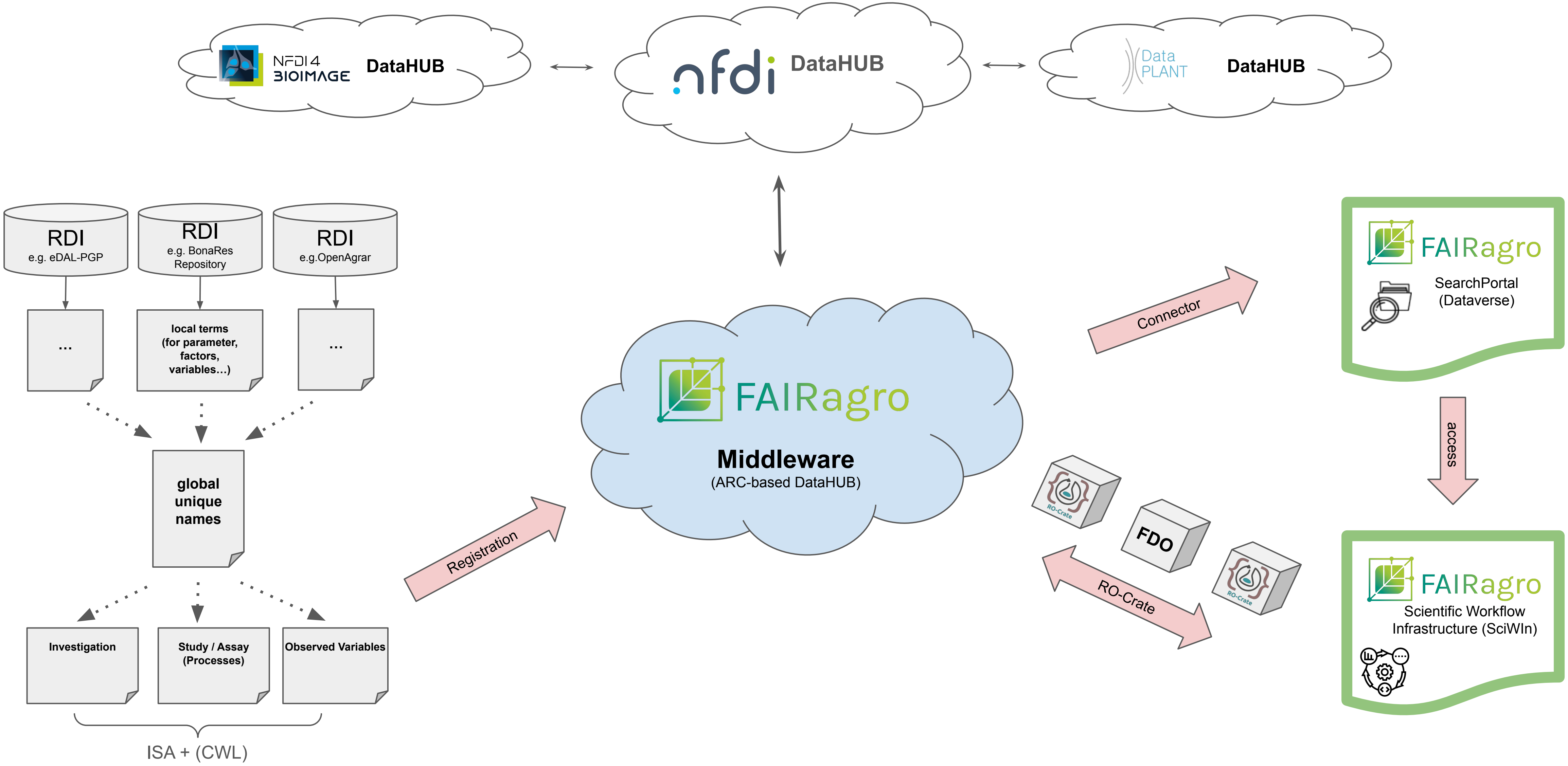
ISA → homogeneous & interoperable metadata handling

CWL → reproducible workflow handling

Git → management & provenance



source: https://www.nfdi4plants.de, Weil et al. 2023 TPJ



www.fairagro.net

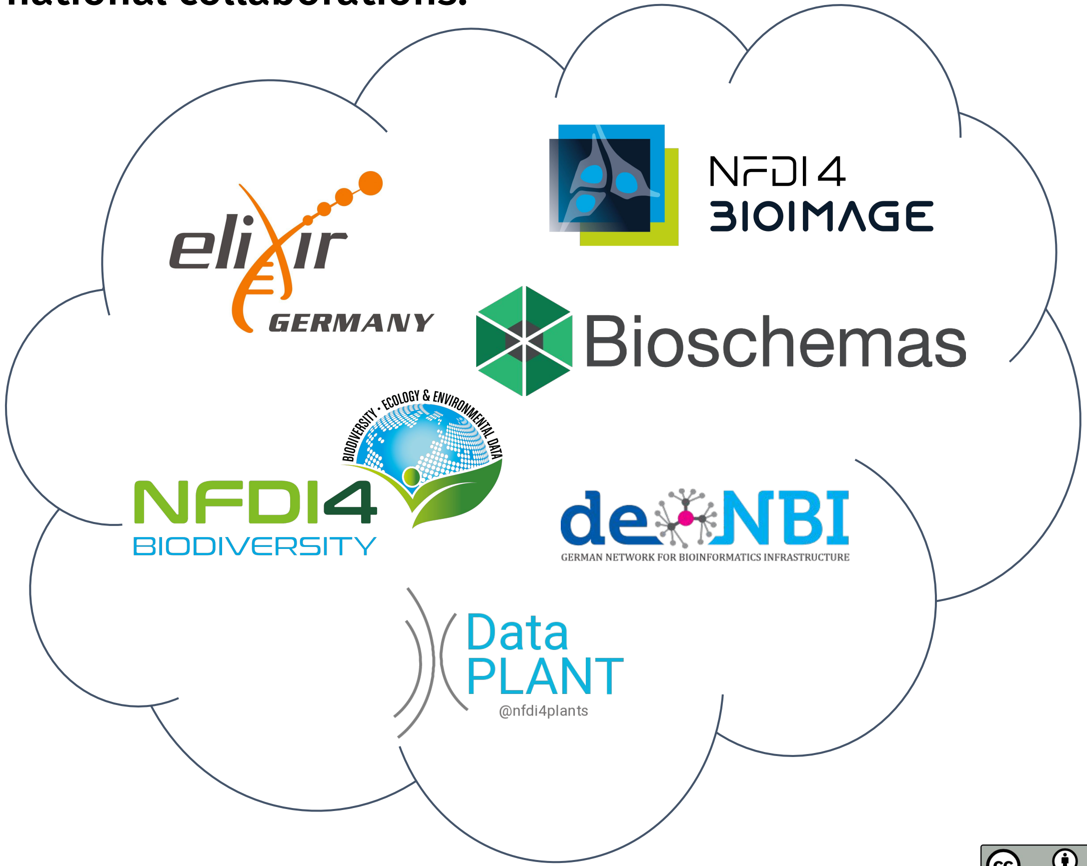
source: Brizuela et al. 2024 *JIB in press*

Summary

The FAIRagro middleware as technical backbone and central infrastructure for providing comprehensive services is design using state-of the art concepts and technologies. It is inspired and designed using synergies from thigh intra/inter-consortial, national and international collaborations.

- ⇨ **ARC concept:** *DataPLANT, NFDI4Biodiversity, NFDI4Bioimage…*
- ⇨ **FDOs/RO-Crate:** *ELIXIR Interoperability + Data Platform, ELIXIR Plant Community, de.NBI*
- ⇨ **Schema.org/BioSchema:** *BioSchema SC & Community, NFDI4Chem, NFDI4Microbiota*

- organise on-demand/regular meetings & tech deep dives
- collaborative work on several projects during different Hackathons/Symposium & released several preprints
- initiated two additional working groups for Bioschema extension


elixir GERMANY
NFDI4 BIOIMAGE
Bioschemas
BIODIVERSITY · ECOLOGY & ENVIRONMENTAL DATA
NFDI4 BIODIVERSITY
de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE
Data PLANT
@nfdi4plants

cc BY

FAIRagro

FAIRagro impulse 3

Building the FAIRagro search portal

Julian Schneider

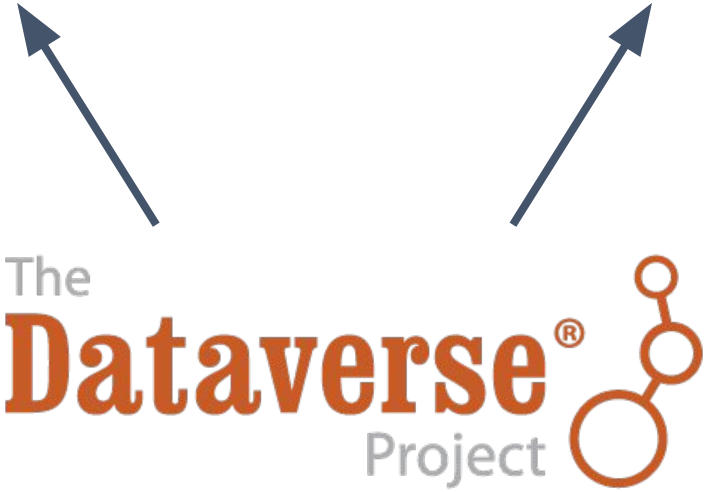
Services in the FAIRagro search portal

Central Search Service

- Searchability of datasets
  - metadata from Middleware
  - → datasets from all RDIs

Infrastructure Registry

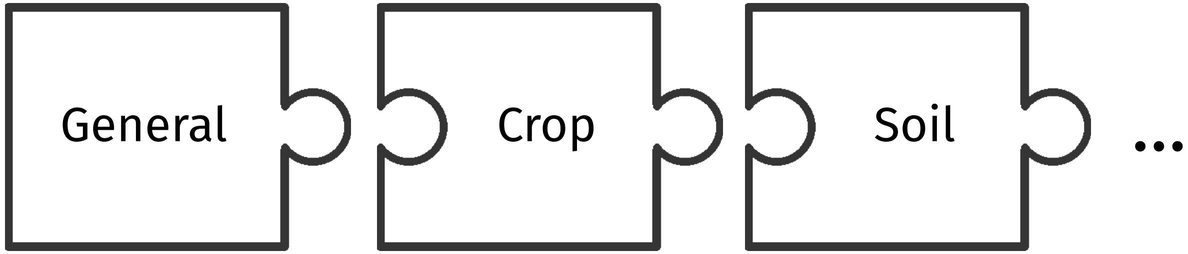
- separate from the Central Search Service
- Findability of RDIs
  - aggregates info from Middleware



Metadata powering the FAIRagro search portal

Central Search Service

- Specialized MDS
  - modularity covers domains
  - → flexible metadata blocks in Dataverse
  - could be used to generate Bioschemas markup



Infrastructure Registry

- Interoperable Metadata Standard
  - e.g. *re3data, DCAT, ...*

Connections for FAIRagro search portal development

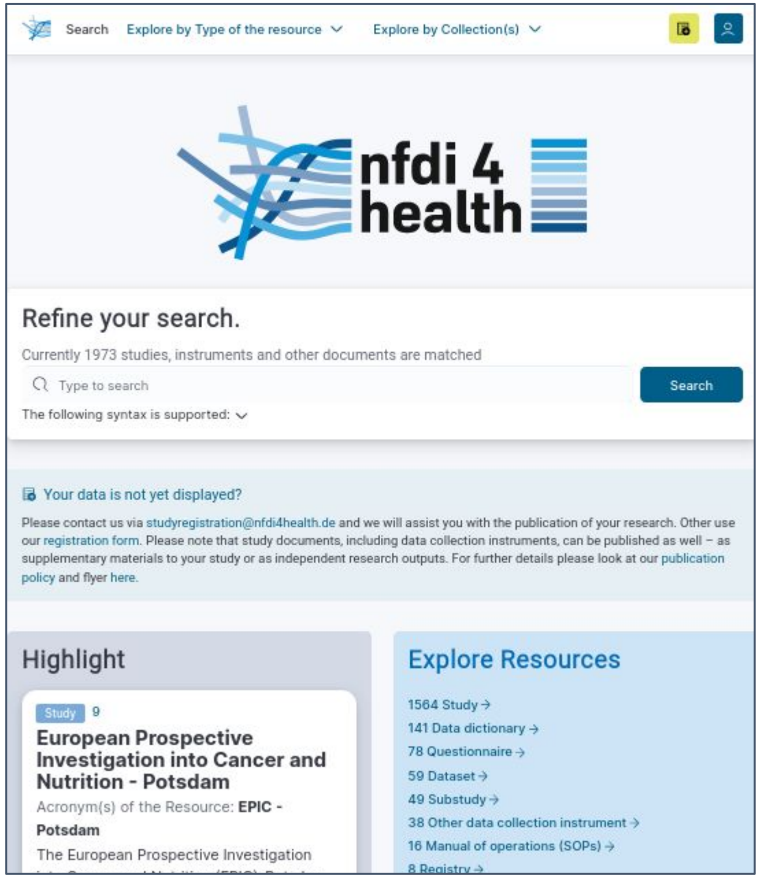
NFDI4Health

- Dataverse-based Study Hub
  - UI serves as foundation for FAIRagro: →

Dataverse community

DataPLANT

- Connection ARC → Dataverse



Refine your search.

Currently 1973 studies, instruments and other documents are matched

The following syntax is supported:

Your data is not yet displayed?

Please contact us via studyregistration@nfdi4health.de and we will assist you with the publication of your research. Other use our registration form. Please note that study documents, including data collection instruments, can be published as well – as supplementary materials to your study or as independent research outputs. For further details please look at our publication policy and flyer here.

Highlight

Study 9

European Prospective Investigation into Cancer and Nutrition - Potsdam

Acronym(s) of the Resource: **EPIC - Potsdam**

The European Prospective Investigation

Explore Resources

- 1564 Study →
- 141 Data dictionary →
- 78 Questionnaire →
- 59 Dataset →
- 49 Substudy →
- 38 Other data collection instrument →
- 16 Manual of operations (SOPs) →
- 8 Registry →

https://csh.nfdi4health.de/

Discussion on cross-cutting topics between FAIRagro, CGIAR, de.NBI & ELIXIR

Discussion on cross-cutting topics between FAIRagro, CGIAR, de.NBI and/or ELIXIR

- What could be a possible topic for collaboration?
- How can we collaborate and use synergies?
- What are the next steps?
- Who is responsible for the next steps?

Wageningen Data Competence Center

- Data Steward Service Center (DSSC)
  - Data stewardship
  - Institutional data management
- Workshop planned 2024/2025

Topic: The human site of RDM

CC BY

FAIRagro

FAIRagro Community Summit Berlin 2024