HOW TO FAIRIFY DATA IN YOUR OWN LAB

Marta Dembska, DLR Institute of Data Science

MSE Day, 14.11.2023

DLR

Reality of laboratory data

Data are diverse…

Image Ernesto Enslava on Pixabay

Image by Tatiana on Pixabay

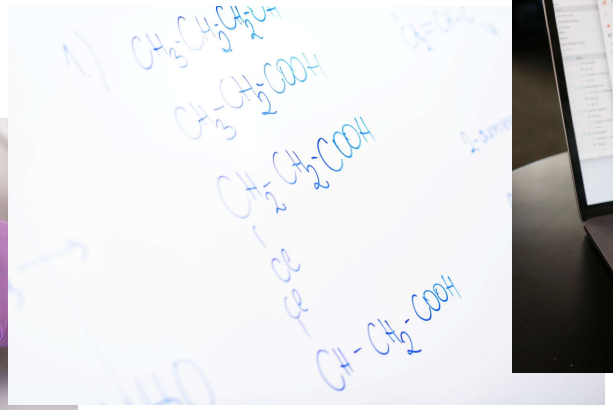


Image by Benis Arapovic on Vecteezy

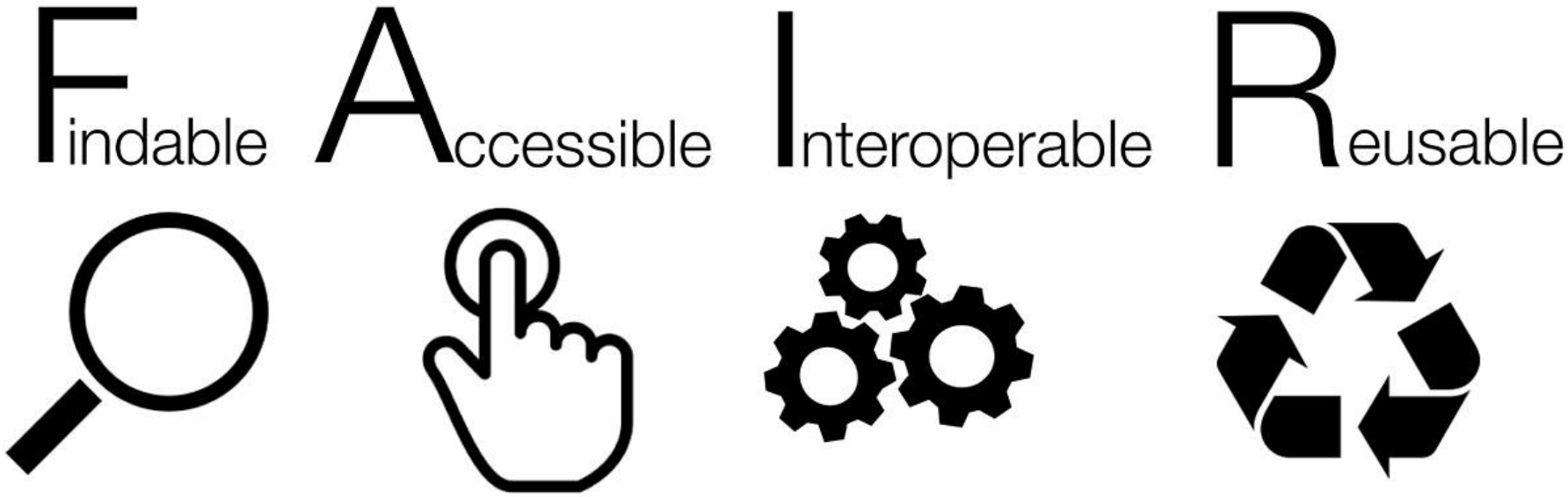
Photo by Campain Creators on Unslpash

Image PublicDomainPictures on Pixabay

… and scientist are creative

Image by storyset on Freepik

Can someone else make use of your data?



By SangyaPundir - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=53414062

FAIRification of your laboratory data starts with making them (internally or externally) available

Findability

F1

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

F2

Data are described with rich metadata (defined by R1)

F3

Metadata clearly and explicitly include the identifier of the data they describe

F4

(Meta)data are registered or indexed in a searchable resource

Photo by Tobias Fischer on Unsplash

Findability

F1 Globally unique and persistent identifier

- Unique identifiers are the key for publishing FAIR data

If your data cannot be found, it‘s like it didn‘t exist!

- Keeping a web link active is not for free

F2 Rich metadata

- Generous and extensive – incl. information on context, quality and conditions, or characteristics of the data
- Make data location based on their metadata possible

F3 Connection between data and their metadata

- Metadata and datasets are often separate files!
- Connection should be annotated in a formal manner – there are tools that can be used

F4 Metadata searchable

- Findability not guaranteed by rich metadata and identifiers

If nobody knows that your data resources exist, they may go unused!

- Domain specific registries exist and are ready to use

Accessibility

A1

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

A1.1

The protocol is open, free, and universally implementable

A1.2

The protocol allows for an authentication and authorisation procedure, where necessary

Photo by Jonathon Young on Unsplash

A2

A2. Metadata are accessible, even when the data are no longer available

Accessibility

A1 Standard communication protocols retrieval

- Avoid barriers to access (e.g. unusual protocols, poor documentation)
- Provide contact person for sensitive data
- Access to at least metadata for (almost) every user
- Clear instructions on authentication

Accessible does not automatically mean open or free!

- Machine actionability of access requirements

A1.1 Open and free protocol

A1.2 Authentication and authorisation where necessary

A2 Metadata accessible after data are not

- Metadata easier and cheaper to store
- Metadata important for replication

Interoperability

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

Photo by Kaleidico on Unsplash

Interoperability

I1 Common language for knowledge representation

(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

- Precise syntax and grammar
- Language specifications accessible
- More than one scenario in mind in design process

I2 FAIR vocabularies

- Controlled vocabularies documented and using unique and persistent identifiers
- Documentation findable and accessible

I3 Meaningful links between (meta)data resouces

- Cross-reference between datasets
- Connection to enrich the contextual knowledge about them

Reusability

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

Photo by Ravin Rau on Unsplash

Reusability

R1 Labels describe the context

R1.1 Clear usage rights

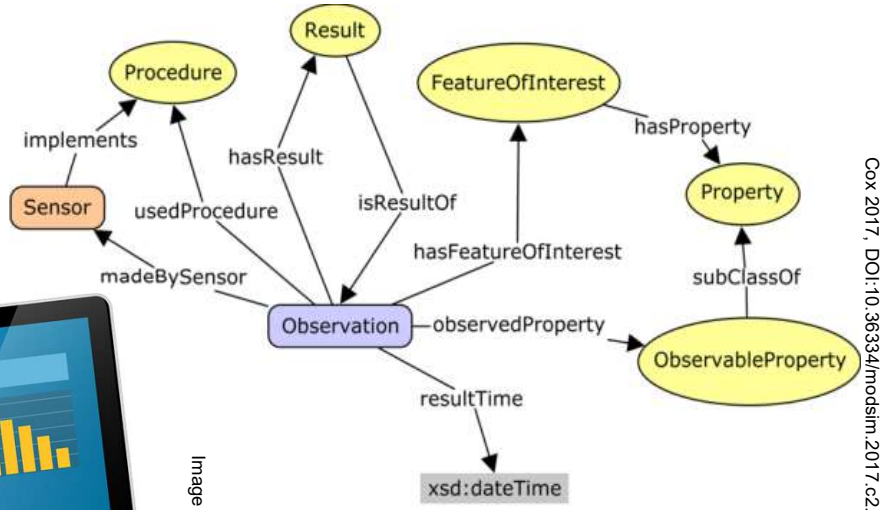
R1.2 Detailed origin of data

R1.3 Alignment with domain community standards

- Enable user to decide if the data is useful for them (incl. experimental protocols, brand of equipment, laboratory conditions etc.)
- Metadata as general as possible
- Clear and commonly used licence attached to data
- Provenance of data (e.g. authorship and other roles in creation, workflow description, processing details)
- Similar datasets are easier to reuse
- Follow best practise standards of the community

Enabling interoperability with domain level ontologies

- Creating semantic content for domain experts and machines
- Digitalisation of (meta)data acquisition (e.g. ELNs)
- Knowledge representation in a given domain
- Formalisation of procedures
- Reproducibility of experiments
- Data life cycle
- Quality measures



Cox 2017, DOI:10.36334/modsim.2017.c2.cox

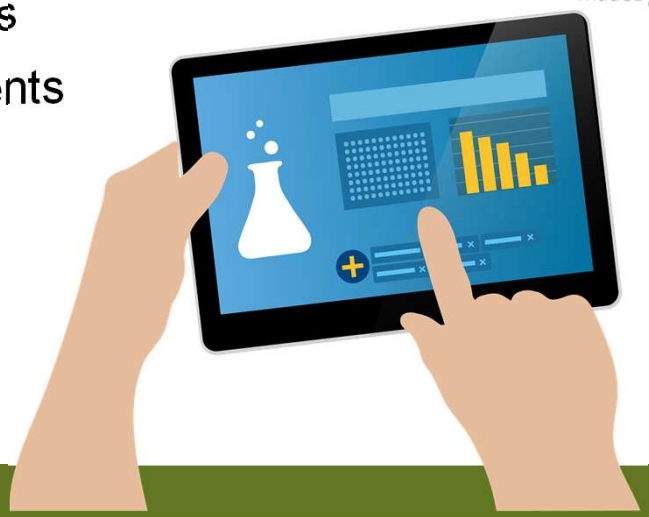
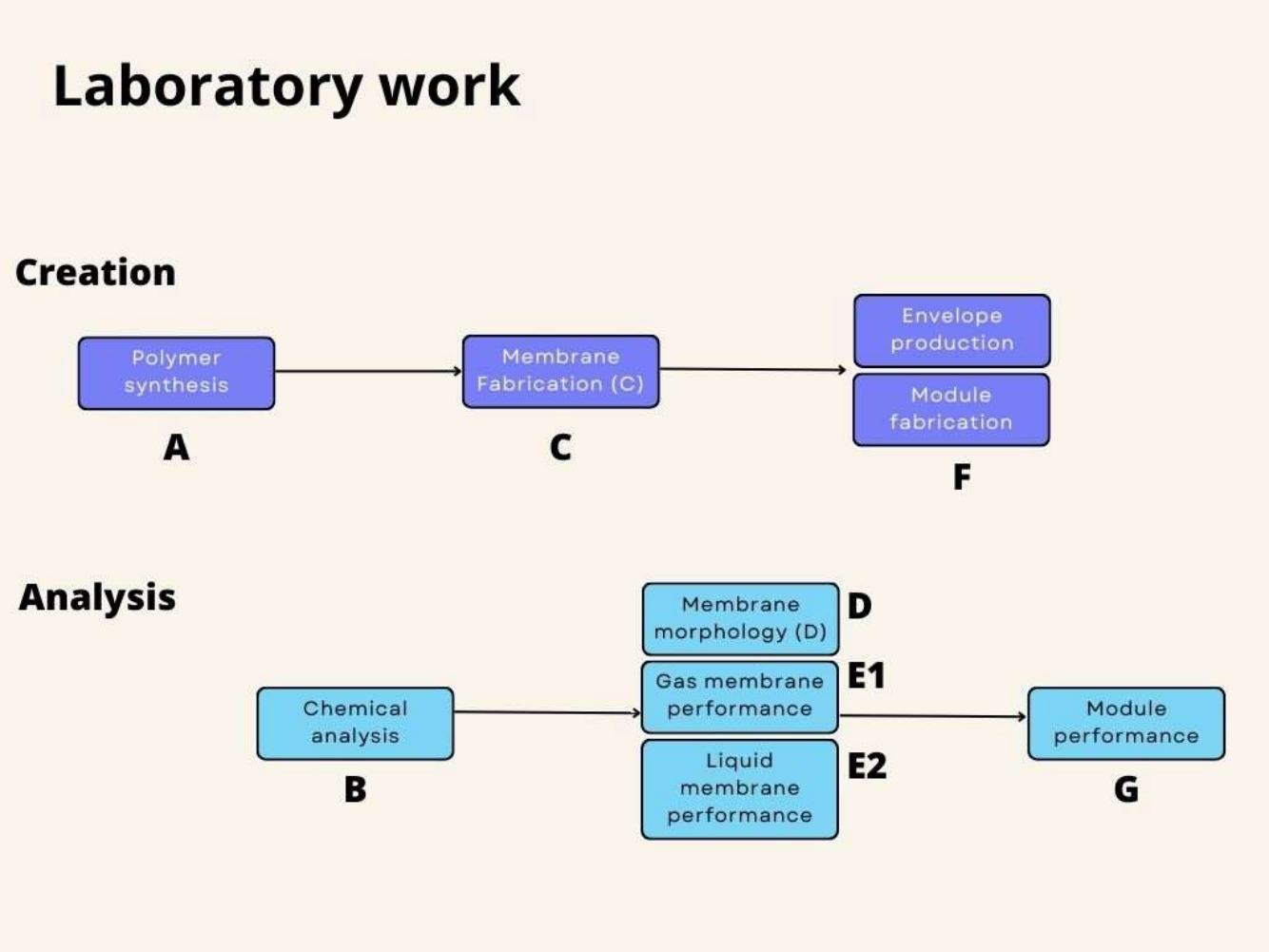


Image by mcmurryjulie on Pixabay

Goals
- Combine knowledge engineering and domain modelling through collaborative work of ontology engineers and domain experts
- Ontology development through use of formal languages and tooling

Domain knowledge



ELNs
- Herbie (in house solution)
- Chemotion (external)

Vocabulary
- domain experts
- ontology reuse

Process modelling
- documentation
- recipes
- data cycle (metadata capture)

Ontology requirements specification

POLYMAT

Use cases:

- Knowledge representation in the fields of polymer and membrane research
- Documentation of scientific work and laboratory processes
- Ontology used in Herbie
- Knowledge representation in modelling laboratory processes

Research application

(X) creation

(Y) analysis

(Z) database, models and ELN

Example competency questions

- ❑ What polymers were used to fabricate a given membrane? **(X)**
- ❑ What equipment was used in a given experiment or measurement? **(X, Y)**
- ❑ Where are the results of a given calculation stored? **(Z)**
- ❑ What computational models were used to perform the experiment? **(X, Y, Z)**
- ❑ What (physical, chemical) calculated characteristics are derived from a measured characteristic? **(Y)**

Ontology requirements specification

POLYLAB

Use cases:

- Reproducibility of a given experiment
- Cross referencing of protocol sub elements in Herbie

Process

(S) polymer synthesis

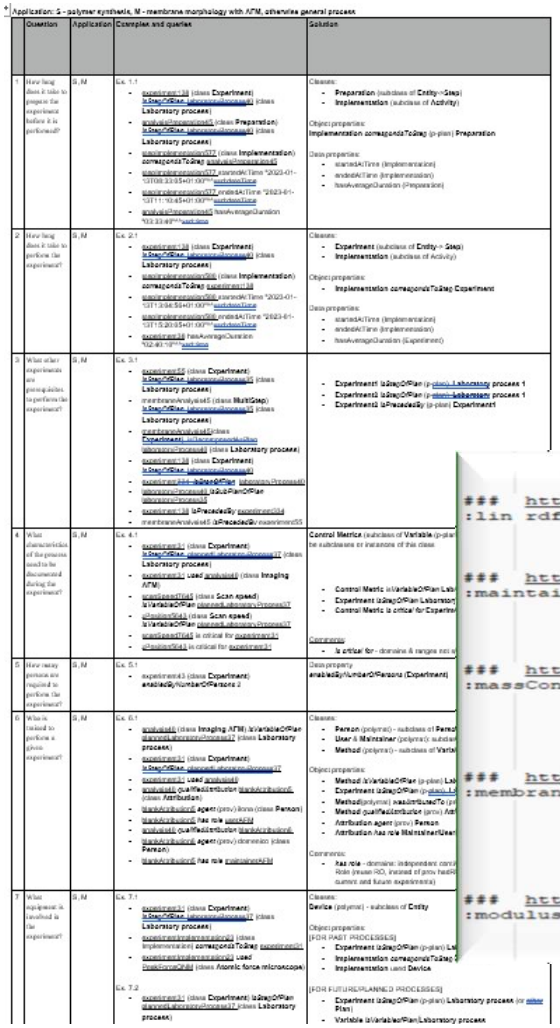
(M) membrane morphology with AFM (atomic force microscope)

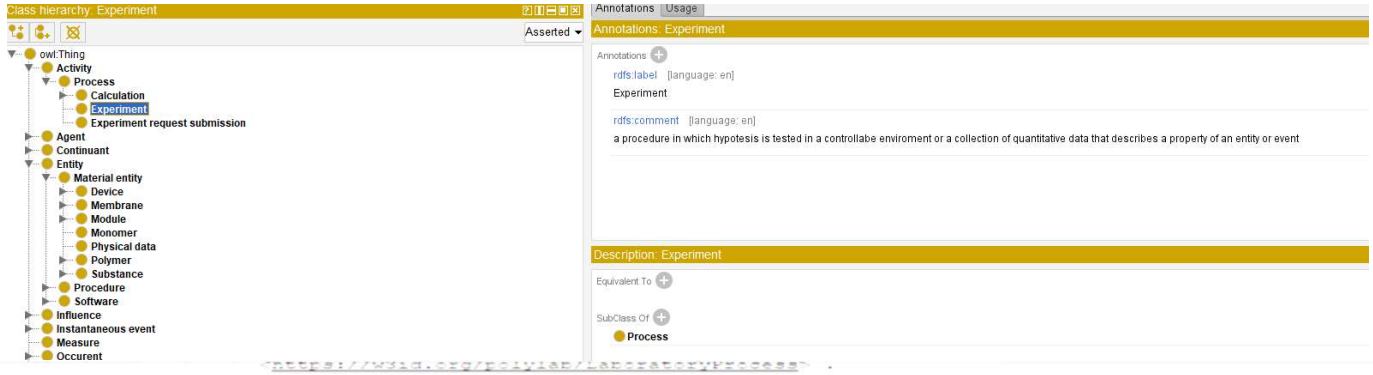
Otherwise general process

Example competency questions

- ❑ Who is trained to perform a given experiment or measurement **(S, M)**
- ❑ What is the order number of a given measurement request?
- ❑ What substances are used in the experiment? **(S)**
- ❑ What properties of the equipment are important for a given measurement? **(M)**

Ontology implementation – conceptualization and encoding






```
<https://w3id.org/polylab/LaboratoryProcess> .

###  http://example.org/polylab_example_data#lin
:lin rdf:type owl:NamedIndividual ,
             <http://www.w3.org/ns/prov#Person> .


###  http://example.org/polylab_example_data#maintainerAFM
:maintainerAFM rdf:type owl:NamedIndividual ,
                        pmat:Maintainer ;
               <http://purl.obolibrary.org/obo/RO_0000087> :maintainerAFM .


###  http://example.org/polylab_example_data#massConcentrationOfIsobutyeleInChloroform
:massConcentrationOfIsobutyeleInChloroform rdf:type owl:NamedIndividual ,
                                                    pmat:MassConcentration ;
                                           <http://www.ontology-of-units-of-measure.org/resource/om-2/hasValue> :_500gL ;
                                           <http://www.w3.org/2004/02/skos/core#related> :chloroform72 .


###  http://example.org/polylab_example_data#membraneAnalysis45
:membraneAnalysis45 rdf:type owl:NamedIndividual ,
                             <http://purl.org/net/p-plan#MultiStep> ;
                    <http://purl.org/net/p-plan#isDecomposedAsPlan> :laboratoryProcess40 ;
                    <http://purl.org/net/p-plan#isPrecededBy> :experiment55 ;
                    <http://purl.org/net/p-plan#isStepOfPlan> :laboratoryProcess35 .


###  http://example.org/polylab_example_data#modulusOfElasticity41
:modulusOfElasticity41 rdf:type owl:NamedIndividual ,
                                <http://www.ontology-of-units-of-measure.org/resource/om-2/ModulusOfElasticity> .
```

Ontology implementation – evaluation

polytetrafluoroethyleneMembrane5 (class: **Membrane for liquids**, **FlatSheetMembrane**) - instance of class **FlatSheetMembrane**
polytetrafluoroethylene1 (class: **Homopolymer**) - instance of class **Homopolymer**

polytetrafluoroethyleneMembrane5 (class: **Membrane for liquids**, **FlatSheetMembrane**) *contains* polytetrafluoroethylene1 (class: **Homopolymer**, **Poly(tetrafluoroethylene)**)
polytetrafluoroethylene1 *has acronym* PTFE
polytetrafluoroethylene1 *has CAS nr* 9002-84-0
polytetrafluoroethylene1 *has IUPAC name* Poly(1,1,2,2-tetrafluoroethylene)
polytetrafluoroethylene1 *has product ID* MHp201203$

CQ1: What polymers were used to fabricate a given membrane?

Q1.1 Return polymers that a specific membrane contains

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX pmat: <https://w3id.org/polymat/>
PREFIX bfo: <http://purl.obolibrary.org/obo/>
PREFIX ex: <http://example.org/polymat_example_data#>

SELECT DISTINCT ?membrane ?polymer
WHERE {

  ?membrane_type rdfs:subClassOf* pmat:Membrane.
  ?membrane rdf:type ?membrane_type .
  ?membrane bfo:RO_0001019 ?polymer .
  ?polymer_type rdfs:subClassOf* pmat:Polymer.
  ?polymer rdf:type ?polymer_type

  FILTER (?membrane=ex:polytetrafluoroethyleneMembrane5>)
}
```

Table | Response | 1 result in 0.056 seconds | Simple view ☐ Ellipse ☑ | Filter query results

	membrane	polymer
1	<http://example.org/polymat_example_data#polytetrafluoroethyleneMembrane5>	<http://example.org/polymat_example_data#polytetrafluoroethylene1>

Showing 1 to 1 of 1 entries

Main takeouts

- Crucial role of people in FAIRification
- Domain specialist and knowledge engineers are one team
- ELNs as a medium to facilitate digitalisation and FAIRification in the labs

F	well described data are findable
A	clear documentation of access
I	shared vocabularies facilitate interoperability
R	reuse, reproduce and avoid repetitions

Contact: Marta.Dembska@dlr.de

DLR Institute of Data Science, Jena