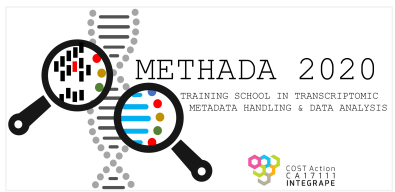
Metadata handling: Part I

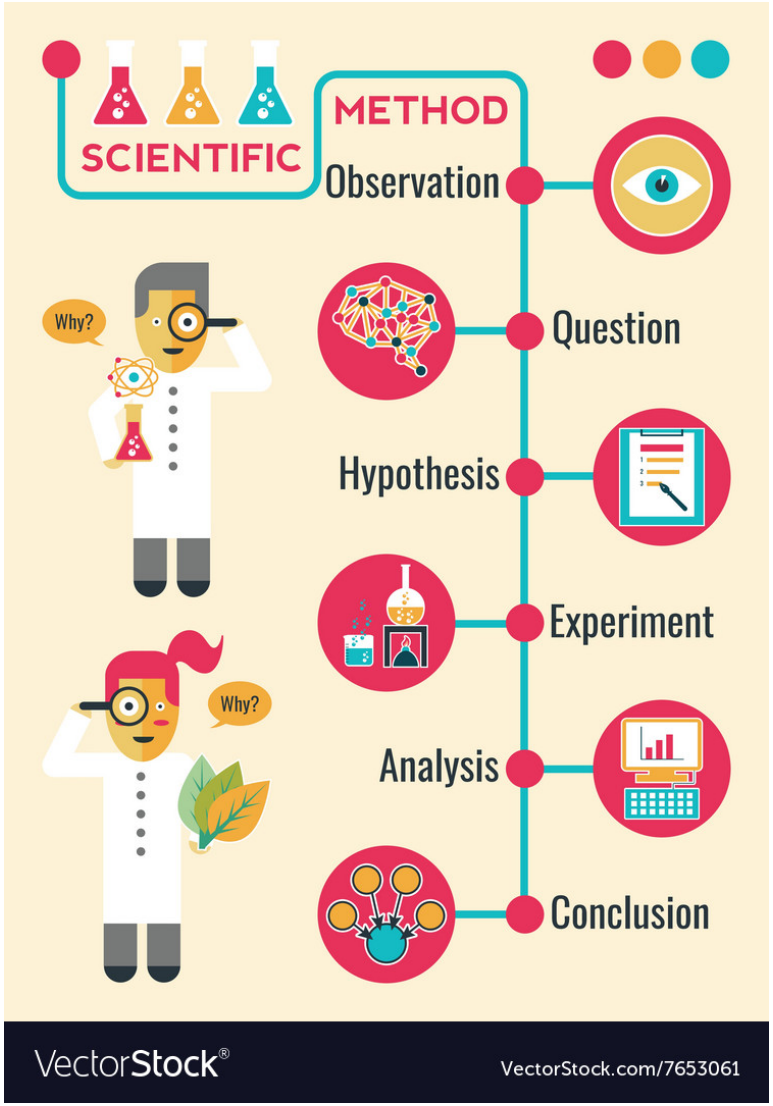
Marco Moretto
marco.moretto@fmach.it

Fondazione Edmund Mach
Research and Innovation Centre
Computational Biology Unit

5th February 2020



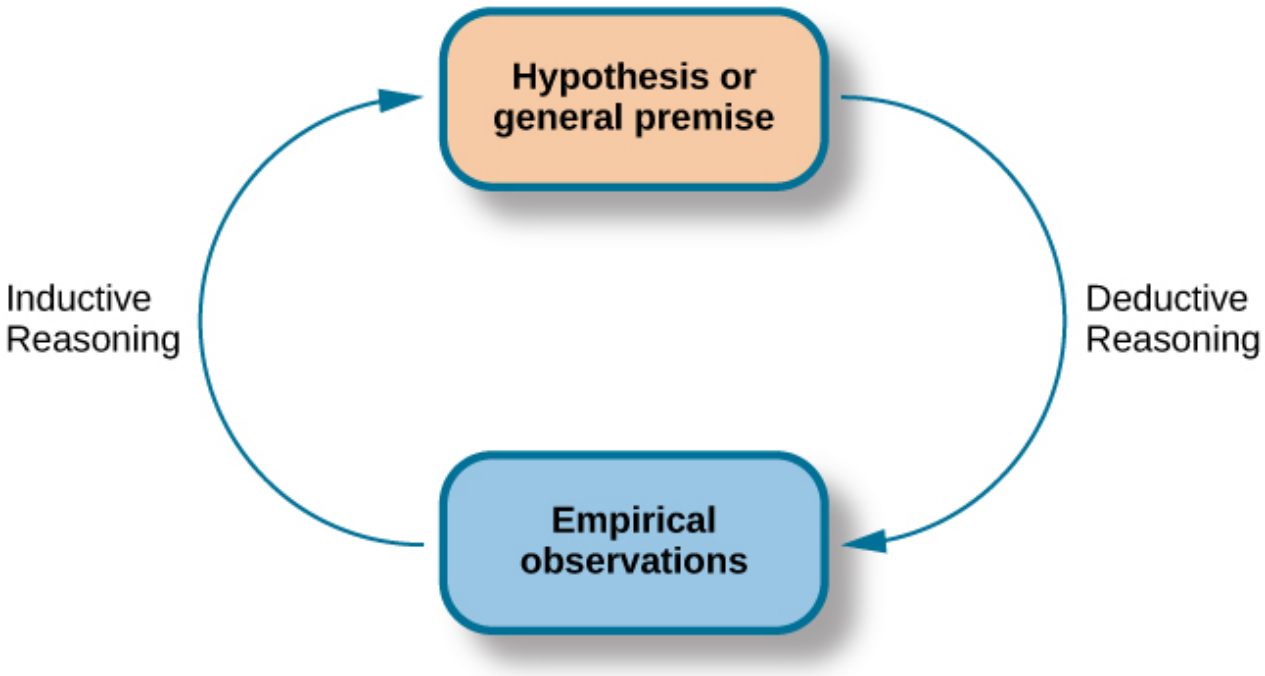
The scientific method



The scientific method

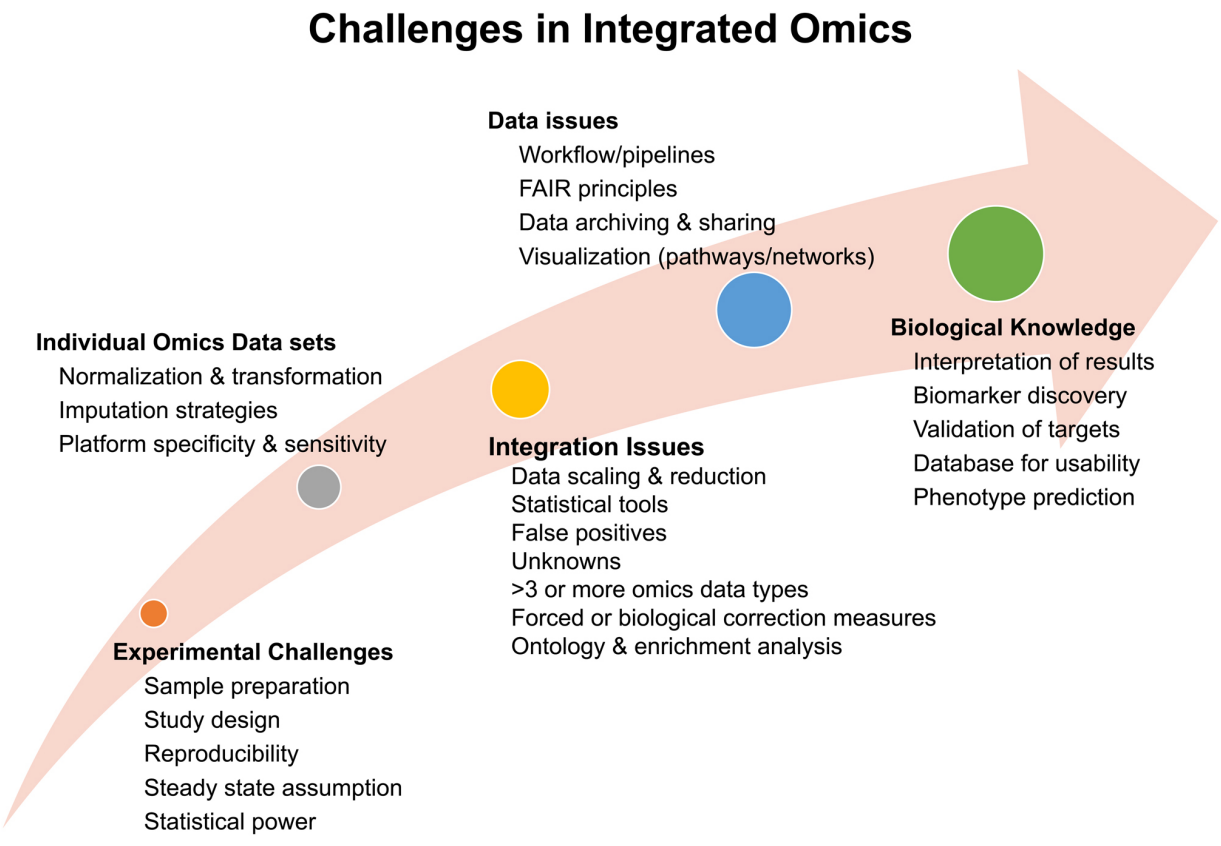
Inductive learning

The scientific method



In order to be valid, a conclusion has to be general (reproducible)

Data integration



To tackle **complex** scientific questions, experimental data sources often need to be **harmonized**

The FAIR principales

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

The FAIR principales

- Findable
  - (meta)data are assigned a globally unique and persistent identifier
  - data are described with rich metadata (defined by R1 below)
  - metadata clearly and explicitly include the identifier of the data it describes
  - (meta)data are registered or indexed in a searchable resource
- Accessible
  - (meta)data are retrievable by their identifier using a standardized communications protocol
  - the protocol is open, free, and universally implementable
  - the protocol allows for an authentication and authorization procedure, where necessary
  - metadata are accessible, even when the data are no longer available
- Interoperable
  - (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - (meta)data use vocabularies that follow FAIR principles
  - (meta)data include qualified references to other (meta)data
- Reusable
  - meta(data) are richly described with a plurality of accurate and relevant attributes
  - (meta)data are released with a clear and accessible data usage license
  - (meta)data are associated with detailed provenance
  - (meta)data meet domain-relevant community standards

The FAIR principales

- Findable
  - (meta)data are assigned a globally unique and persistent identifier
  - **data are described with rich metadata (defined by R1 below)**
  - metadata clearly and explicitly include the identifier of the data it describes
  - (meta)data are registered or indexed in a searchable resource
- Accessible
  - **(meta)data are retrievable by their identifier using a standardized communications protocol**
  - the protocol is open, free, and universally implementable
  - the protocol allows for an authentication and authorization procedure, where necessary
  - metadata are accessible, even when the data are no longer available
- Interoperable
  - **(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.**
  - (meta)data use vocabularies that follow FAIR principles
  - (meta)data include qualified references to other (meta)data
- Reusable
  - **meta(data) are richly described with a plurality of accurate and relevant attributes**
  - (meta)data are released with a clear and accessible data usage license
  - (meta)data are associated with detailed provenance
  - (meta)data meet domain-relevant community standards

Example of FAIRness

Uniprot is a comprehensive resource for protein sequence and annotation data.

- All entries are uniquely identfied by a stable URL
- Records contain rich metadata in different format
  - Text
  - HTML
  - RDF
- RDF formatted response utilizes shared vocabularies and ontologies

Example of FAIRness

Uniprot is a comprehensive resource for protein sequence and annotation data.

- All entries are uniquely identfied by a stable URL
- Records contain riche metadata in different format
  - Text
  - HTML
  - **RDF**
- **RDF** formatted response utilizes shared vocabularies and **ontologies**

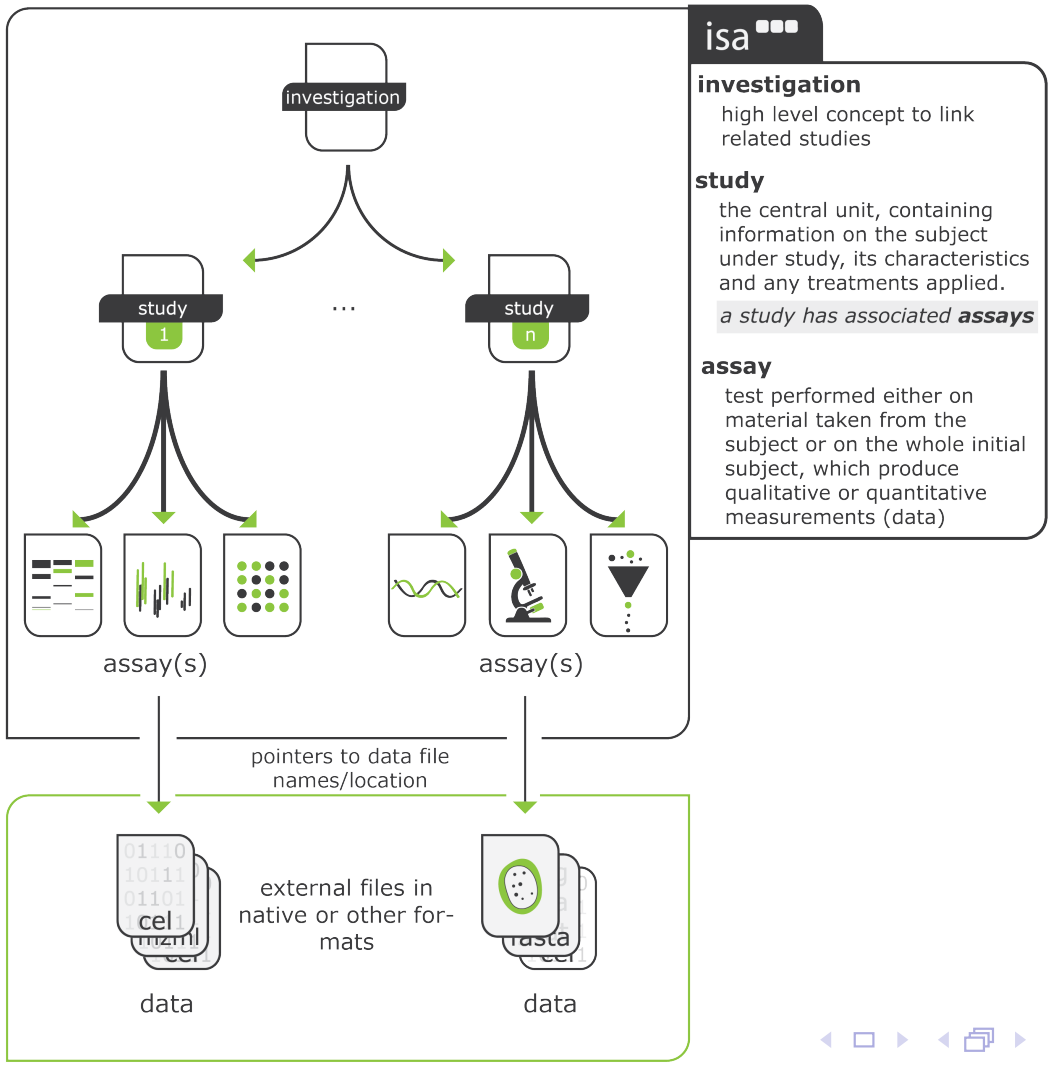
Example of FAIRness

isatools

ISA is a community-driven metadata tracking **framework** to facilitate standards-compliant collection, curation, management and reuse of life sciences datasets. The ISA model has several representation: Tabular, JSON and RDF.

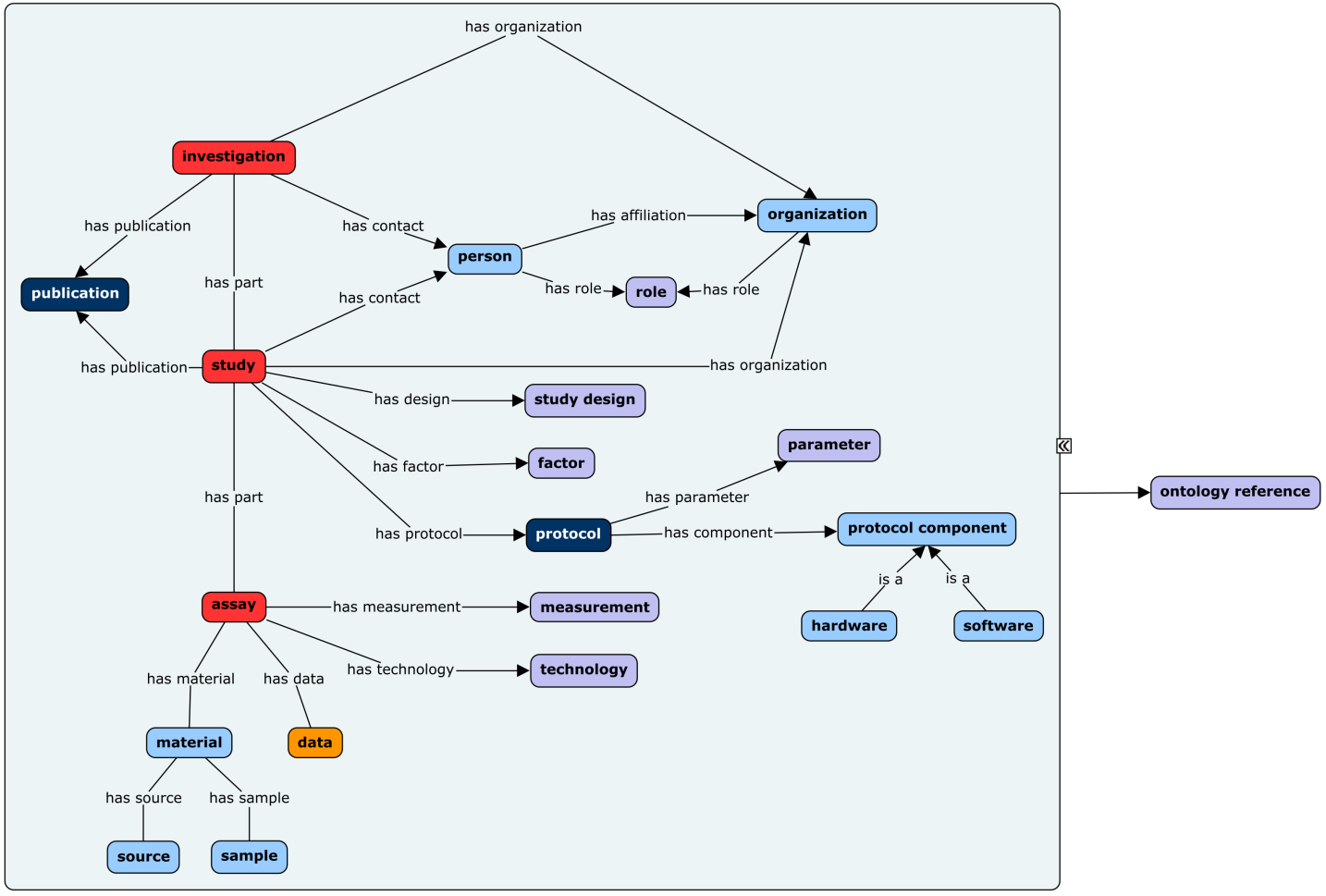
Example of FAIRness

ISA - Investigation Study Assay



Example of FAIRness

ISA Abstract Model



Example of FAIRness

Example of ISA

Example of FAIRness

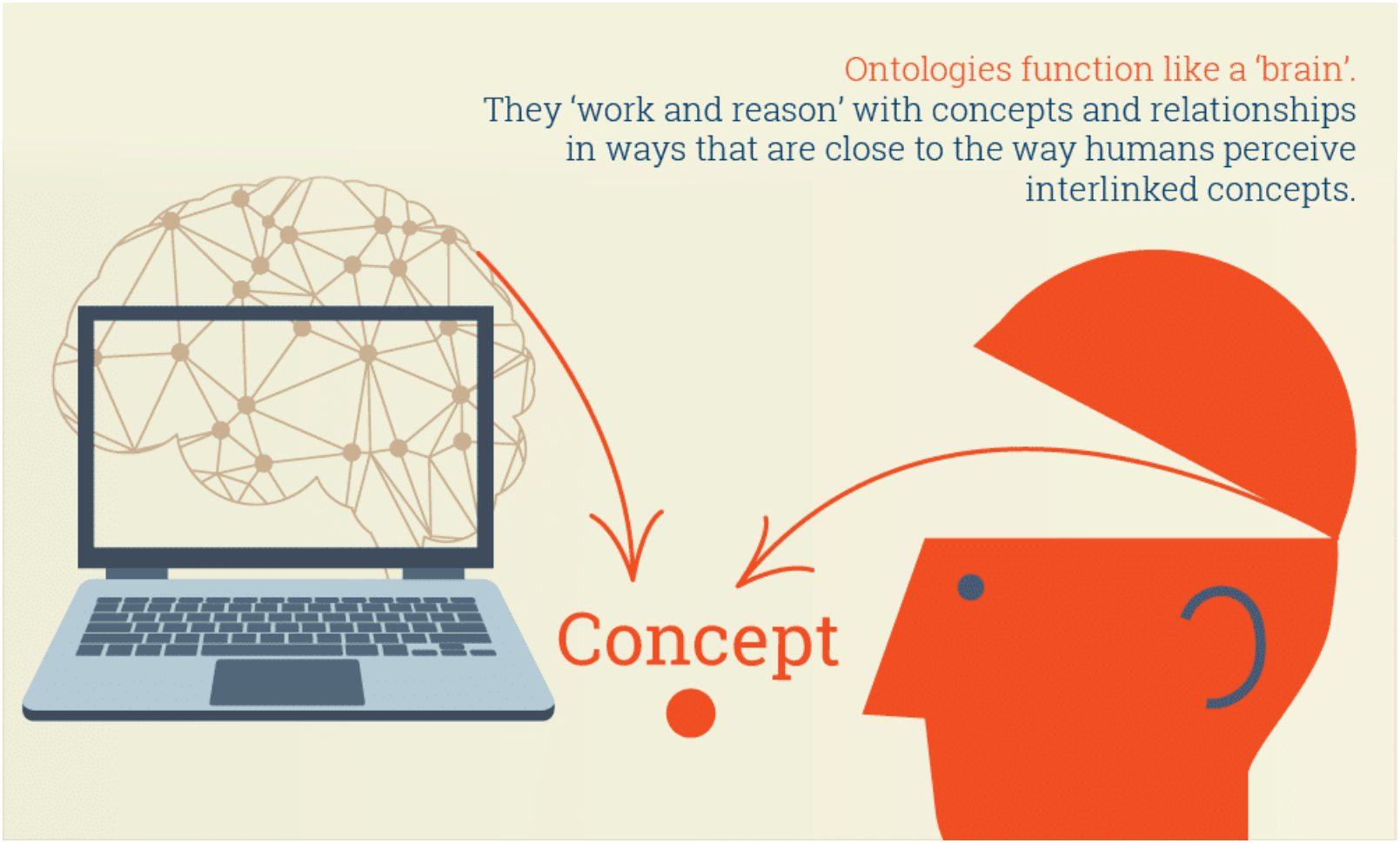
Minimum Information About a Plant Phenotyping Experiment

Example of FAIRness

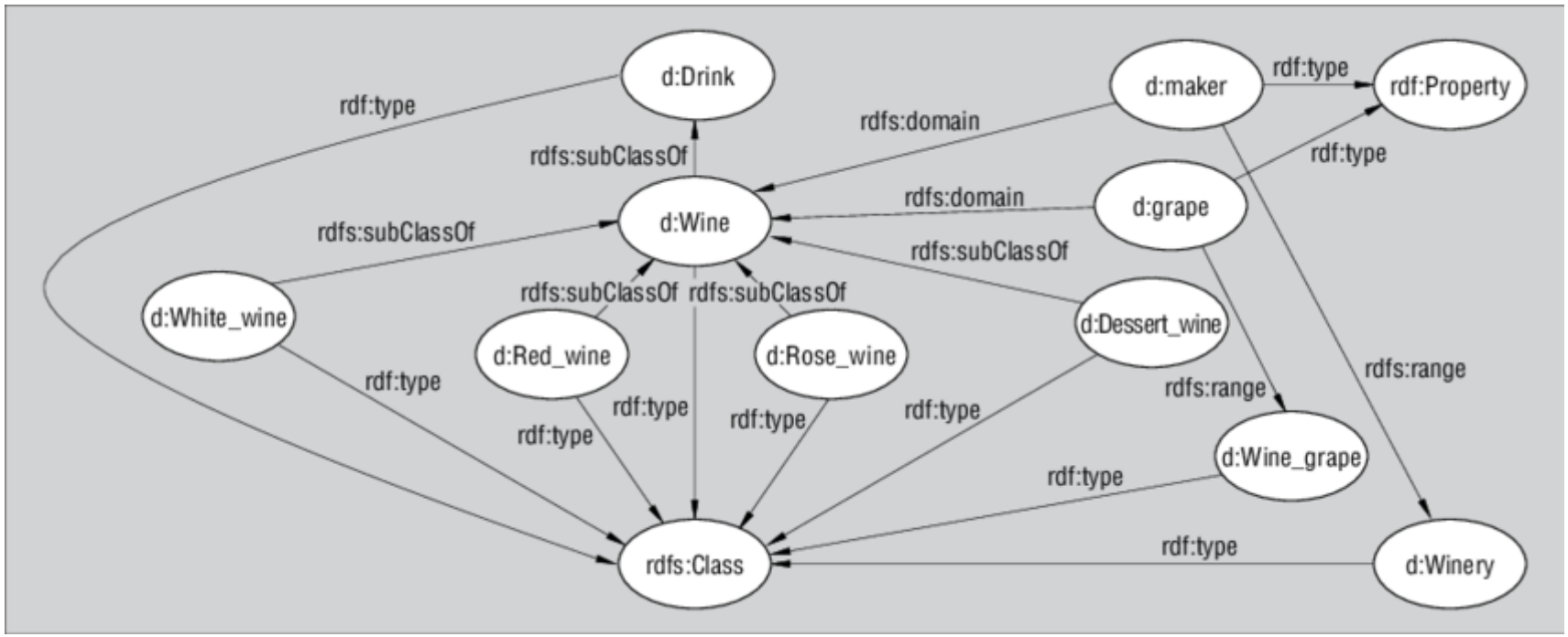
Minimum Information About a Plant Phenotyping Experiment

It defines a **list of attributes** *that might be necessary to fully describe a* **phenotyping** *experiment. An MI document should rather be considered as a* **checklist**, *and consulted by a person describing or depositing the data to ensure the inclusion of all important data characteristics, i.e. what is meaningful for the* **interpretation** *and potential* **replication** *of the research.*

A quick detour: Ontologies

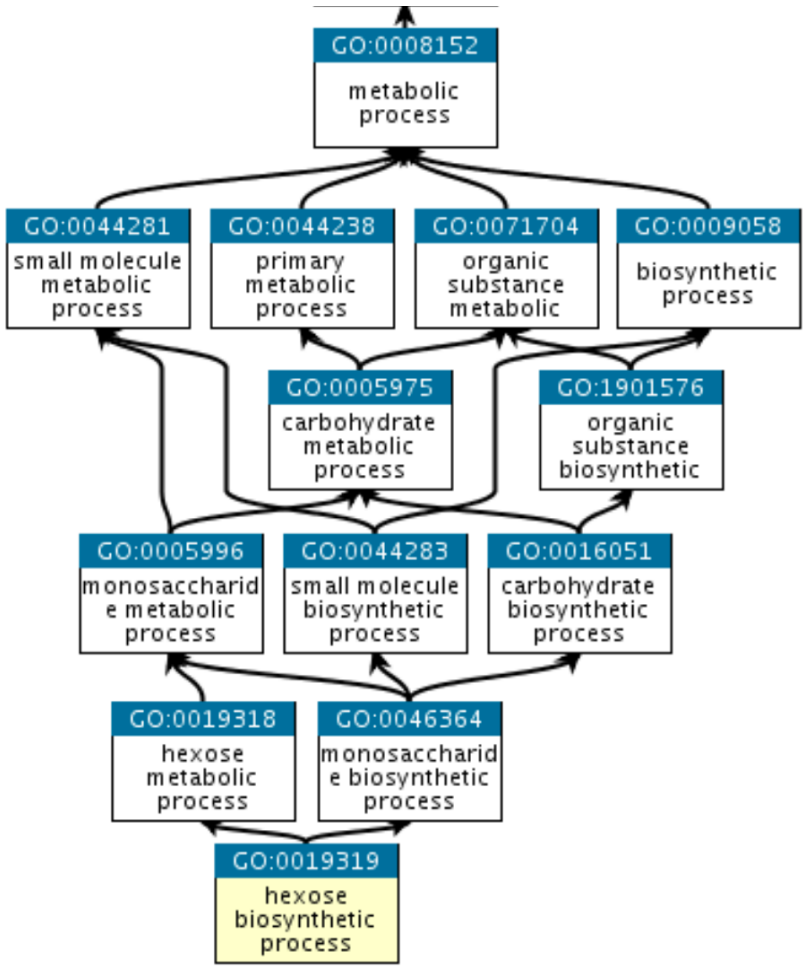


A quick detour: Ontologies



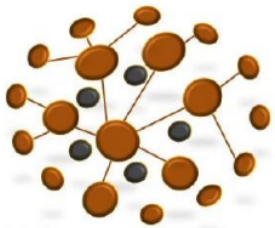
A quick detour: Ontologies

The Gene Ontology

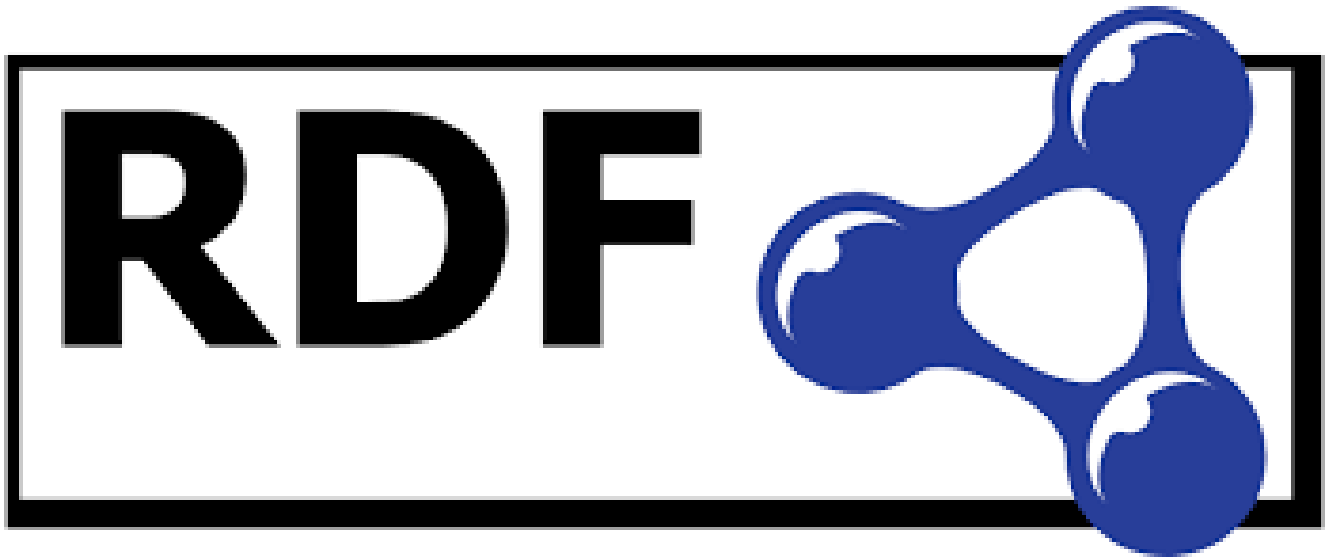


A quick detour: Ontologies

GENEONTOLOGY
Unifying Biology



A quick detour: RDF



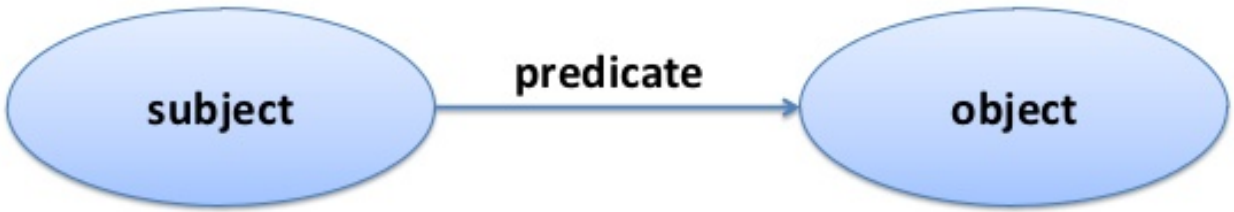
A quick detour: RDF

Resource Description Framework

- World Wide Web Consortium (W3C) standard;
- Metadata data model;
- Based on the idea of making statements about resources;
- Statements are expressions of the form *subject–predicate–object*, known as triples;

A quick detour: RDF

The RDF Triple



Subject - the resource being described

Predicate - a property of that resource

Object - the value of the property

Subject and predicate are defined using URIs.

Object can either be a URI or a 'literal' (text, number, date, etc.)

A quick detour: RDF