Open Science

Publicly Available Resources

Alberto Santos & Yesid Cuesta-Astroz

Instituto Colombiano de Medicina Tropical

The Novo Nordisk Foundation
Center for Biosustainability

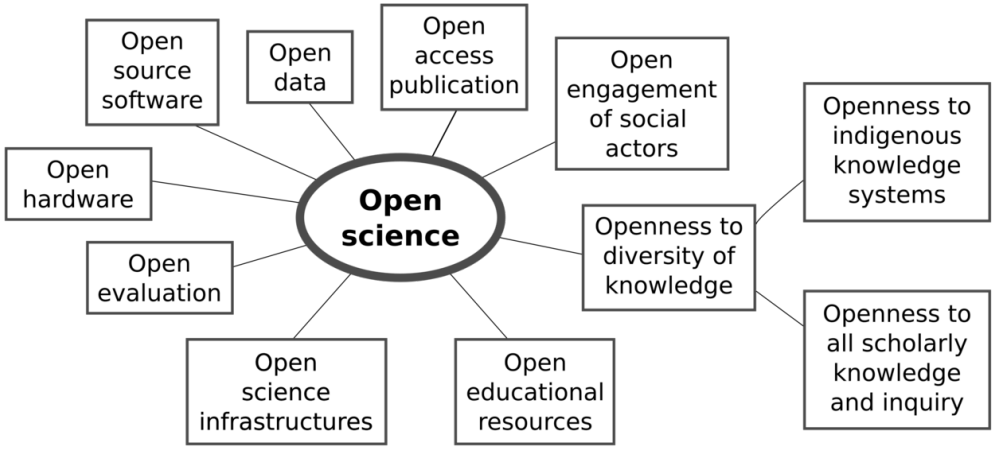
Outline

- What is Open Science?
- FAIR Data and Software
- Challenges sharing and reusing data
- Standardisation and Ontologies
- Publicly available resources

What is Open Science

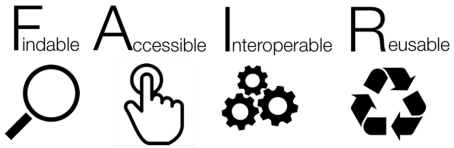
Impact, Contribution, Trust

- Make scientific research **accessible** to all levels of society:
  - Publications
  - Samples
  - Methods
  - **Software**
  - **Data**
- Advantages:
  - **Reproducibility** and **replicability**
  - Societal **responsibility** — publicly funded, publicly available
  - **Multi-purpose** of research outputs
- Disadvantages: concerns of data **misuse**



https://en.wikipedia.org/wiki/Open_science

FAIR Data and Software



- **F**indable and **A**ccessible
  - Add enough **metadata** — data about your data
  - Deposit your data in **public repositories** or make them available in **databases**
- **I**nteroperable:
  - Use **standard** and **open formats**
  - Provide **all data needed** to reproduce your analysis
- **R**eusable:
  - **Describe** your data well, e.g., good metadata but also
  - Attach a **license**

Minimum Information for Biological and Biomedical Investigations

Zenodo
Figshare
Pride
Metabolights
GEO
GitHub

Provide README files describing the data
Use descriptive column headers for the data tables

Challenges Sharing and Reusing

The marshmallow test — delayed gratification

- **Open does not mean FAIR**
- Requires an **effort**
- **Metadata** becomes **the most important data**
- In many cases there are **no standards or multiple ones**
- **Most** of the **data** out there **not FAIR**

https://imgflip.com/memegenerator

https://en.wikipedia.org/wiki/Stanford_marshmallow_experiment

Standardisation and Ontologies

- Data **standardisation** requires defining **terminologies** and **vocabularies** that:
  - Assign **unique identifiers** to entities/concepts such as proteins, genes, diseases
  - **Describe** those entities/concepts and **provide meaning**
  - **Relate** those concepts to other terms
  - Classify those entities/concepts into **categories**
- **Solution** —> **Ontologies**
- **Ontology**:

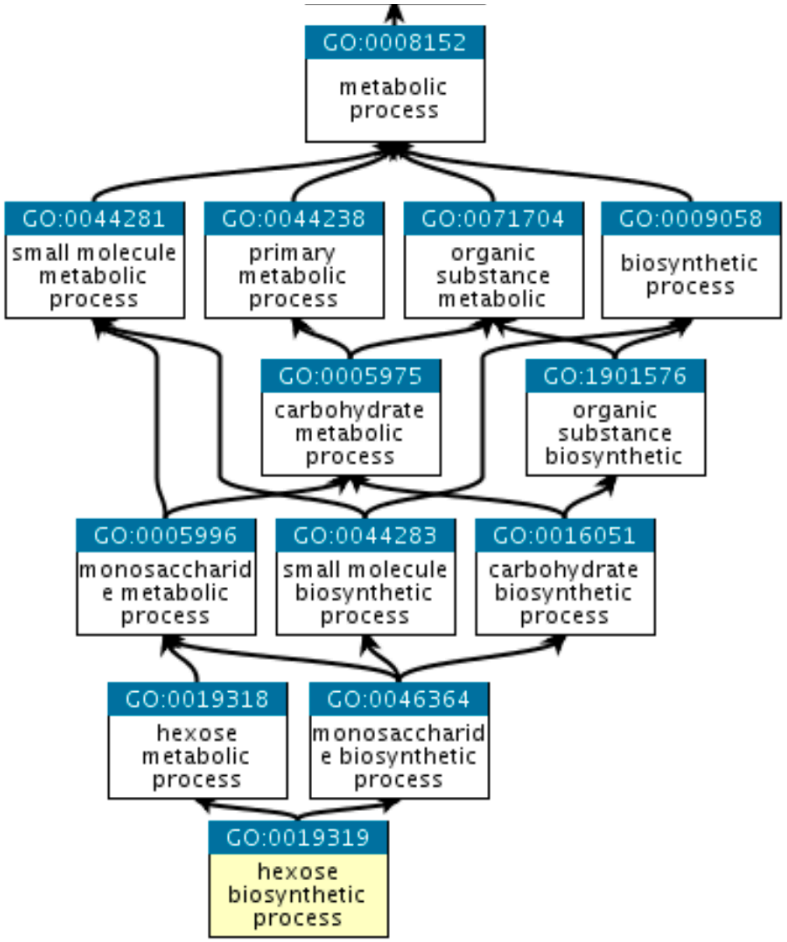
  formal way of representing knowledge in which concepts are described both by their meaning and their relationship to each other

  A collection of terms and their definitions for a specific domain

https://www.nature.com/articles/nrg1295

Ontologies

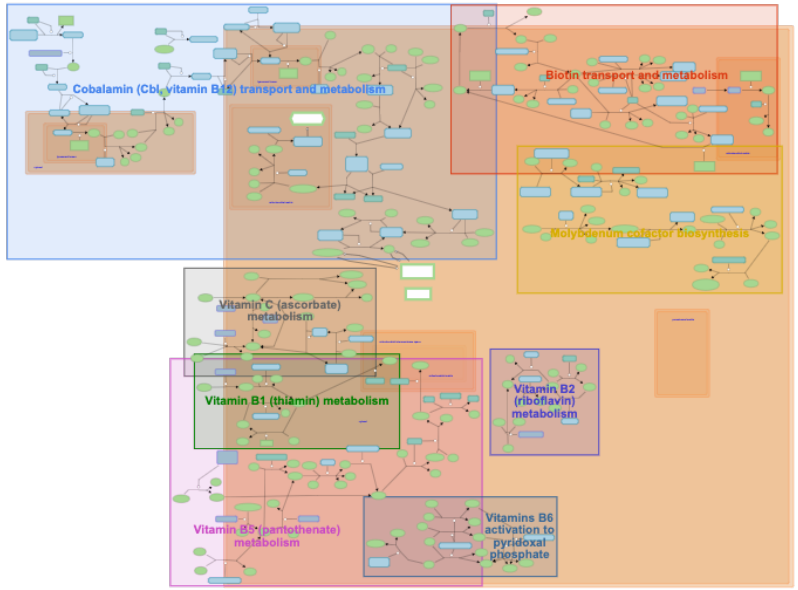
Gene Ontology



Disease Ontology

- disease
  - disease by infectious agent
    - bacterial infectious disease
    - fungal infectious disease
    - infective endocarditis
    - ornithine translocase deficiency
    - osteomyelitis
    - parasitic infectious disease
    - pelvic inflammatory disease
    - viral infectious disease
  - disease of anatomical entity
    - cardiovascular system disease
    - endocrine system disease
    - gastrointestinal system disease
    - hematopoietic system disease
    - immune system disease
    - integumentary system disease
    - musculoskeletal system disease
    - nervous system disease
    - reproductive system disease
    - respiratory system disease
    - thoracic disease
    - urinary system disease
  - disease of cellular proliferation
  - disease of mental health
  - disease of metabolism
  - genetic disease
  - physical disorder
  - syndrome

REACTOME Pathways



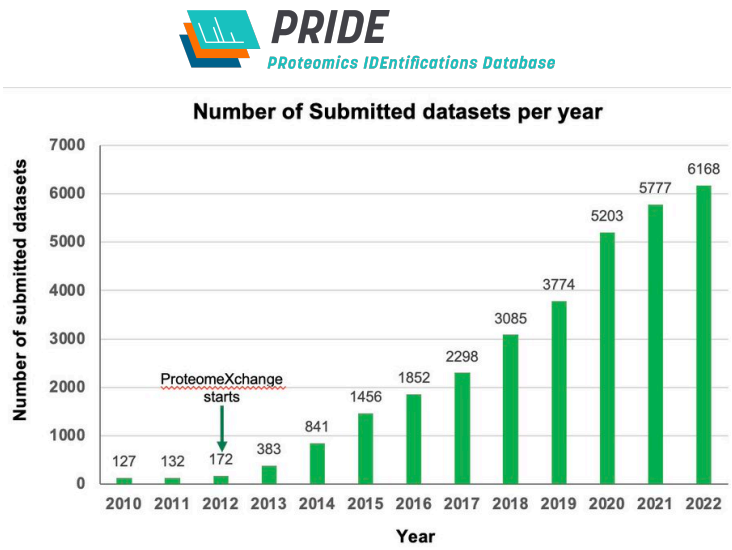
https://www.ebi.ac.uk/ols/ontologies

https://reactome.org/

http://geneontology.org/

Publicly Available Resources

Be a Data Parasite



- **Do not reinvent the wheel**
- **Extend** the **life** and **purpose** of publicly available **data**
- Build **in-silico hypotheses** before jumping into experiments (cheaper, higher success rate)
- **Download — Use — Test — Transform — Upload**
- **Growing number** of **resources** and **datasets** available

https://www.ebi.ac.uk/pride/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6258825/

Examples of Microbes Resources

ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation https://aledb.org/

MiMeDB: the Human Microbial Metabolome Database https://mimedb.org/

Web of microbes (WoM): a curated microbial exometabolomics database for linking chemistry and microbes https://metatlas.nersc.gov/wom/project-begin.view

mBodyMap: a curated database for microbes across human body and their associations with health and diseases https://mbodymap.microbiome.cloud/

MicroPhenoDB Associates Metagenomic Data with Pathogenic Microbes, Microbial Core Genes, and Human Disease Phenotypes http://www.liwzlab.cn/microphenodb

BacDive in 2022: the knowledge base for standardized bacterial and archaeal data https://bacdive.dsmz.de/

MASI: microbiota—active substance interactions database http://www.aiddlab.com/MASI/

iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning https://imodulondb.org/index.html

MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters https://mibig.secondarymetabolites.org/