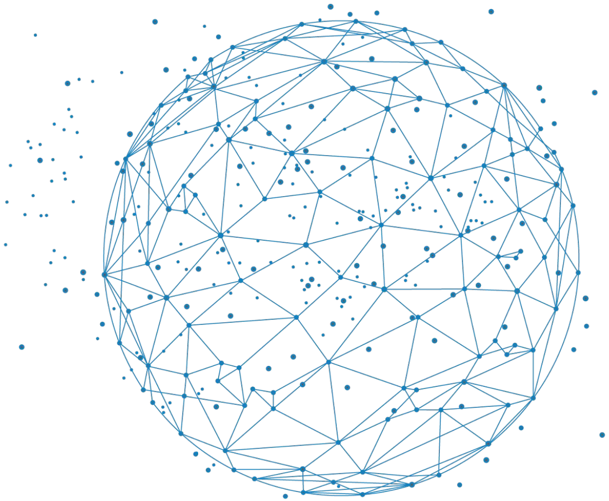


biotic

Metadata standards and ontologies

Claire Rioualen

IFB-core - Marseille

Biotic virtual lunch – February 24th, 2025

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

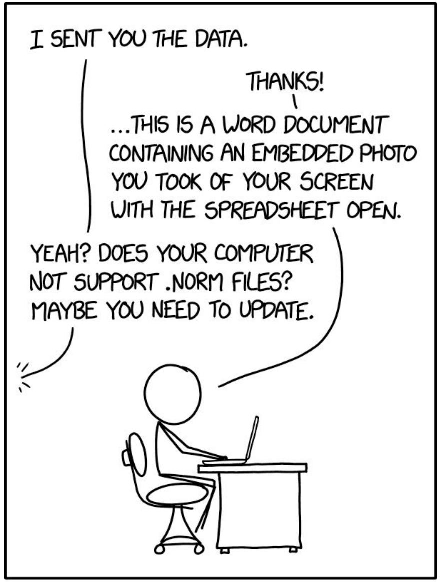
elixir FRANCE

cnrs INRAE Inserm cea FRANCE 2030

Metadata issues and misuse: a never-ending struggle

- Ever tried downloading and analysing data from GEO or ArrayExpress?
  - Search for synonyms (ChIP-seq vs ChIP-sequencing)
  - Extract metadata with distinct field names (source, condition…)
  - Different terms for similar information (Escherichia coli K-12, MG1655…)

https://xkcd.com/2116/

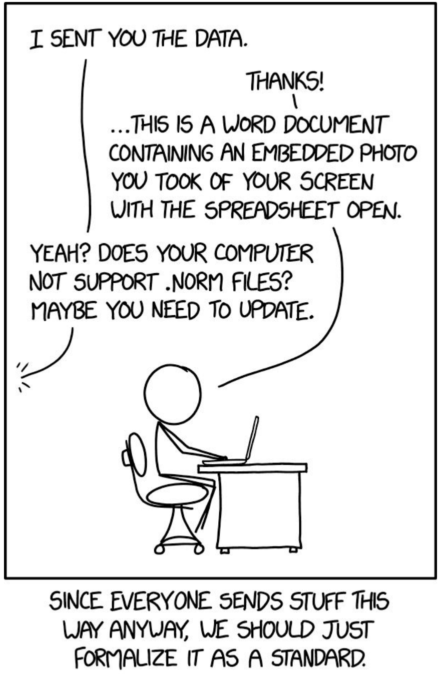

I SENT YOU THE DATA.
THANKS!
...THIS IS A WORD DOCUMENT CONTAINING AN EMBEDDED PHOTO YOU TOOK OF YOUR SCREEN WITH THE SPREADSHEET OPEN.
YEAH? DOES YOUR COMPUTER NOT SUPPORT .NORM FILES? MAYBE YOU NEED TO UPDATE.

SINCE EVERYONE SENDS STUFF THIS WAY ANYWAY, WE SHOULD JUST FORMALIZE IT AS A STANDARD.

Metadata issues and misuse: a never-ending struggle

- Ever tried downloading and analysing data from GEO or ArrayExpress?
  - Search for synonyms (ChIP-seq vs ChIP-sequencing)
  - Extract metadata with distinct field names (source, condition…)
  - Different terms for similar information (Escherichia coli K-12, MG1655…)
- Ever tried integrating and re-analysing several public datasets?
  - Different gene or protein names
  - Metadata and actual data being inconsistent
  - Detailed methods or code lacking from publications
  - Authors not reachable, lab tech working elsewhere…

https://xkcd.com/2116/

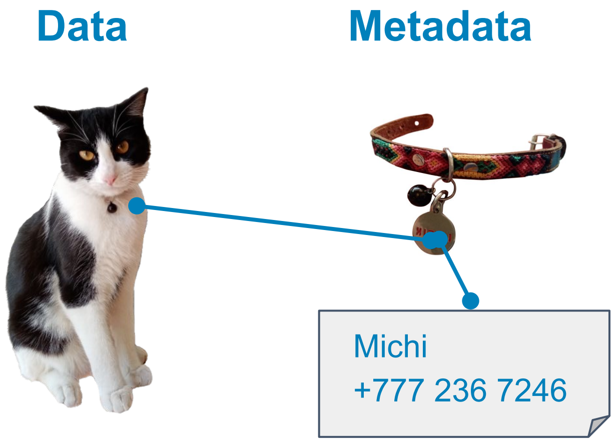

I SENT YOU THE DATA.
THANKS!
...THIS IS A WORD DOCUMENT CONTAINING AN EMBEDDED PHOTO YOU TOOK OF YOUR SCREEN WITH THE SPREADSHEET OPEN.
YEAH? DOES YOUR COMPUTER NOT SUPPORT .NORM FILES? MAYBE YOU NEED TO UPDATE.
SINCE EVERYONE SENDS STUFF THIS WAY ANYWAY, WE SHOULD JUST FORMALIZE IT AS A STANDARD.

> *“Metadata, you see, is really a love note - it might be to yourself, but in fact it’s a love note to the person after you, or the machine after you, where you’ve saved someone that amount of time to find something by telling them what this thing is”*

Jason Scott
http://ascii.textfiles.com/archives/3181

Data and metadata for FAIR science

- Big data poses many challenges owing to its amount, complexity and heterogeneity
- Metadata is essential to annotate its **contents, origin and meaning**
- Metadata alone is not a guarantee that the data and associated results are **FAIR**



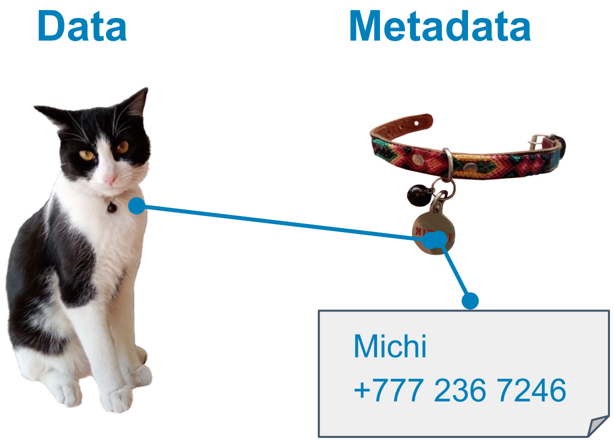
Data

Metadata

Michi
+777 236 7246

Data and metadata for FAIR science

- Big data poses many challenges owing to its amount, complexity and heterogeneity
- Metadata is essential to annotate its **contents, origin and meaning**
- Metadata alone is not a guarantee that the data and associated results are **FAIR**
- **Standard metadata** improves the findability and reusability of the data for future users
- **Semantic metadata** improves data interoperability by using **machine-readable metadata schemes**



Data

Metadata

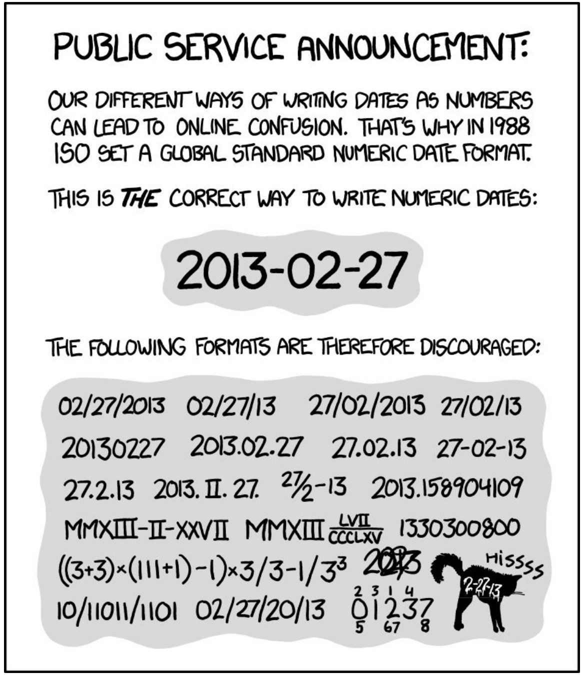
Michi
+777 236 7246

Findable Accessible Interoperable Reusable

Standard metadata requires a standard vocabulary

- **Controlled vocabulary**, thesaurus, subject headings
  - Preferred unique terminology
- **Taxonomy**
  - Terminology
  - Hierarchy of terms
- **Ontology**
  - Terminology
  - Alternative terms
  - Definition
  - Properties associated with terms
  - Hierarchical and semantic relations

https://xkcd.com/1179/

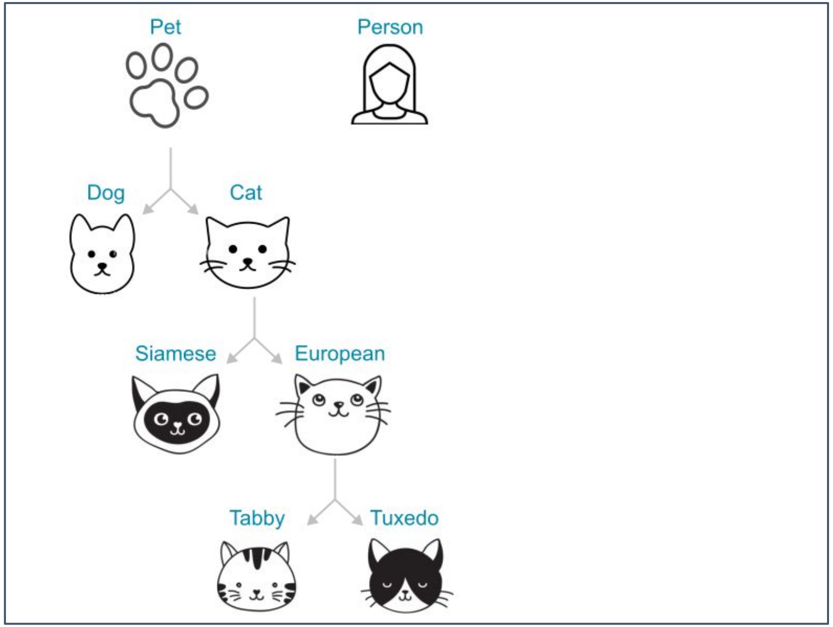

PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
10/11011/1101 02/27/20/13 01237
HISSSS
2-27-13

What is an ontology?

Structured set of concepts related to a given **field of knowledge**, their **definitions**, **relations**, unique and **permanent identifiers**, and other related **properties**

- **Class**: category of being

Pet ontology

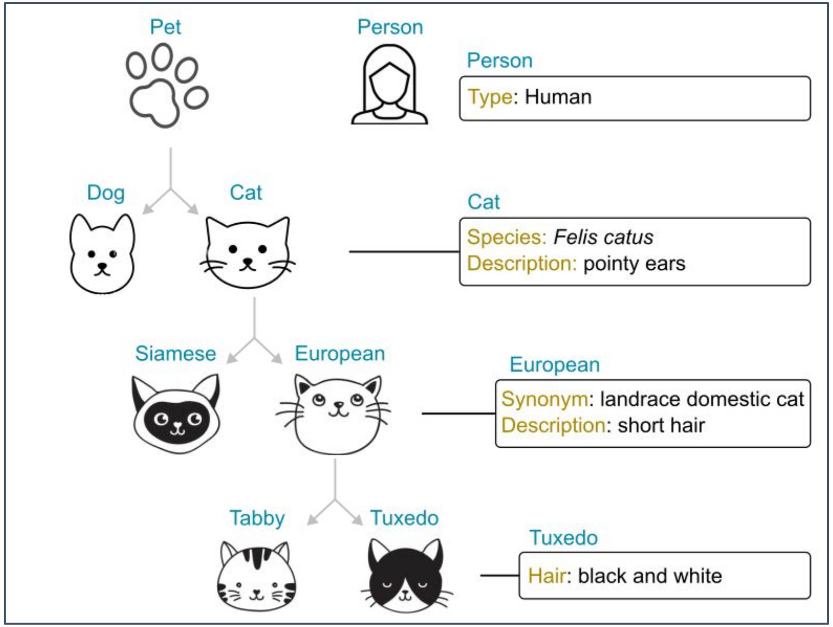


What is an ontology?

Structured set of concepts related to a given **field of knowledge**, their definitions, **relations**, unique and **permanent identifiers**, and other related **properties**

- **Class**: category of being
- **Attributes**: properties associated with a class

Pet ontology



Person
Type: Human

Cat
Species: *Felis catus*
Description: pointy ears

European
Synonym: landrace domestic cat
Description: short hair

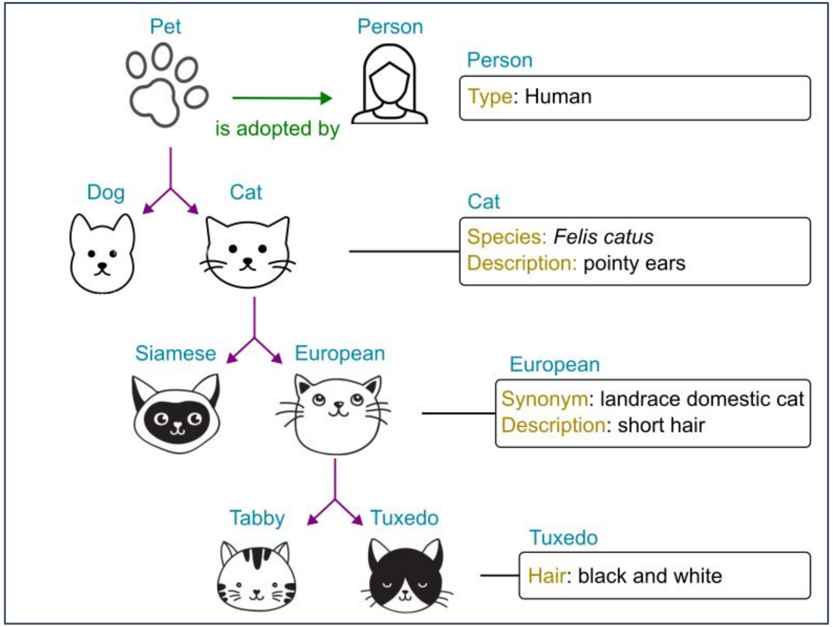
Tuxedo
Hair: black and white

What is an ontology?

Structured set of concepts related to a given **field of knowledge**, their definitions, **relations**, unique and **permanent identifiers**, and other related **properties**

- **Class**: category of being
- **Attributes**: properties associated with a class
- Relations: **hierarchical** or **semantic**

Pet ontology



What is an ontology?

Structured set of concepts related to a given **field of knowledge**, their definitions, **relations**, unique and **permanent identifiers**, and other related **properties**

- **Class**: category of being
- **Attributes**: properties associated with a class
- Relations: **hierarchical** or **semantic**
- **Individual**: instance of a class

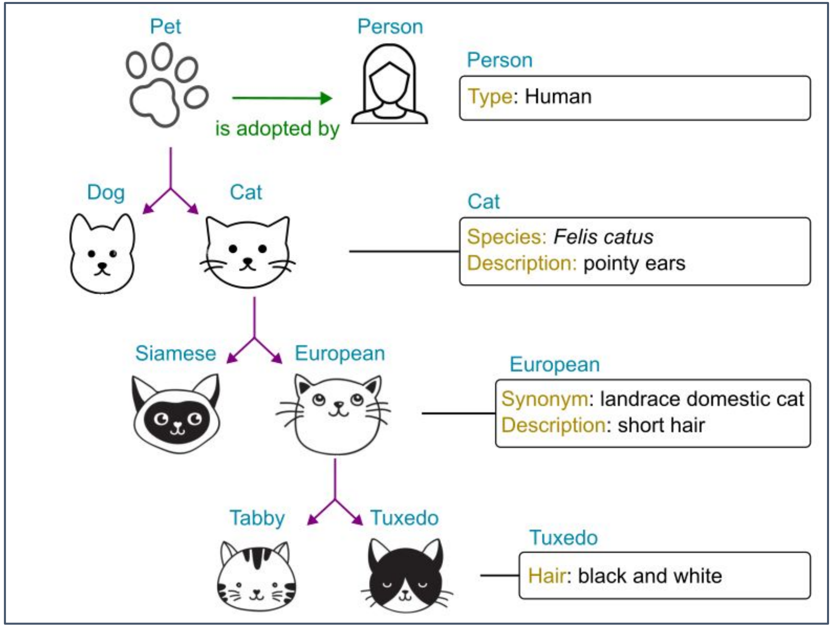
Individual: Michi
Class: Tuxedo

- Subclass of Pet, Cat, European
- Black and white, short hair, pointy ears

Individual: Claire
Class: Person

- Adopted by an individual of class Person

Pet ontology



Pet
Person
is adopted by
Dog
Cat
Siamese
European
Tabby
Tuxedo

Person
Type: Human

Cat
Species: *Felis catus*
Description: pointy ears

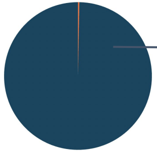
European
Synonym: landrace domestic cat
Description: short hair

Tuxedo
Hair: black and white

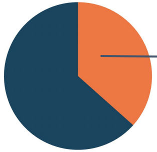
Metadata schemes: the example of schema.org

schema.org: a terminology and **metadata scheme** for *Things*

- Used for web page annotations, it improves indexation by search engines



0.3% of websites use schema.org annotations in their metadata



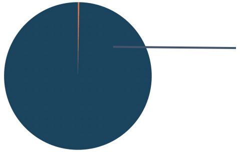
36.6% of Google search results include websites with schema.org metadata

Source: Beard et al. 2016

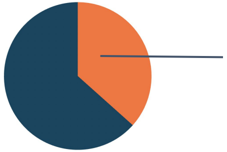
Metadata schemes: the example of schema.org

schema.org: a terminology and **metadata scheme** for *Things*

- Used for web page annotations, it improves indexation by search engines

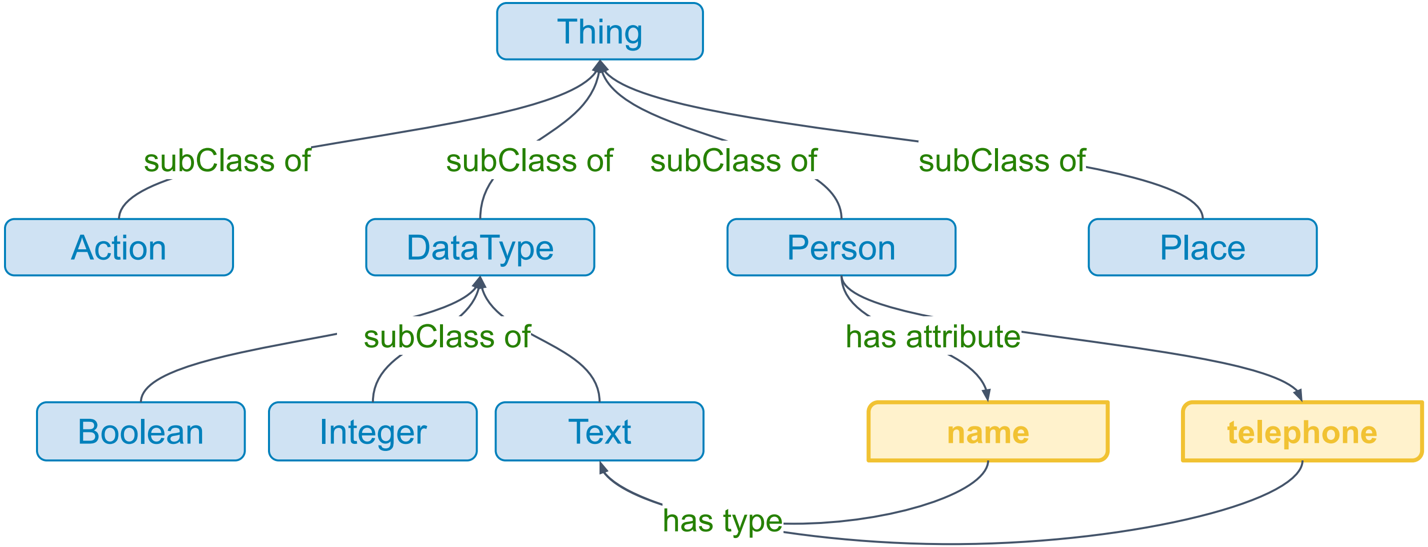


0.3% of websites use schema.org annotations in their metadata



36.6% of Google search results include websites with schema.org metadata

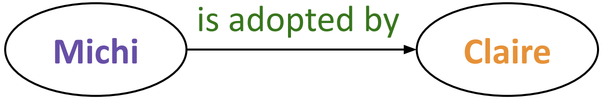
Source: Beard et al. 2016



Ontologies: standards and representations

The **World Wide Web Consortium (W3C)** endorses multiple standards for metadata

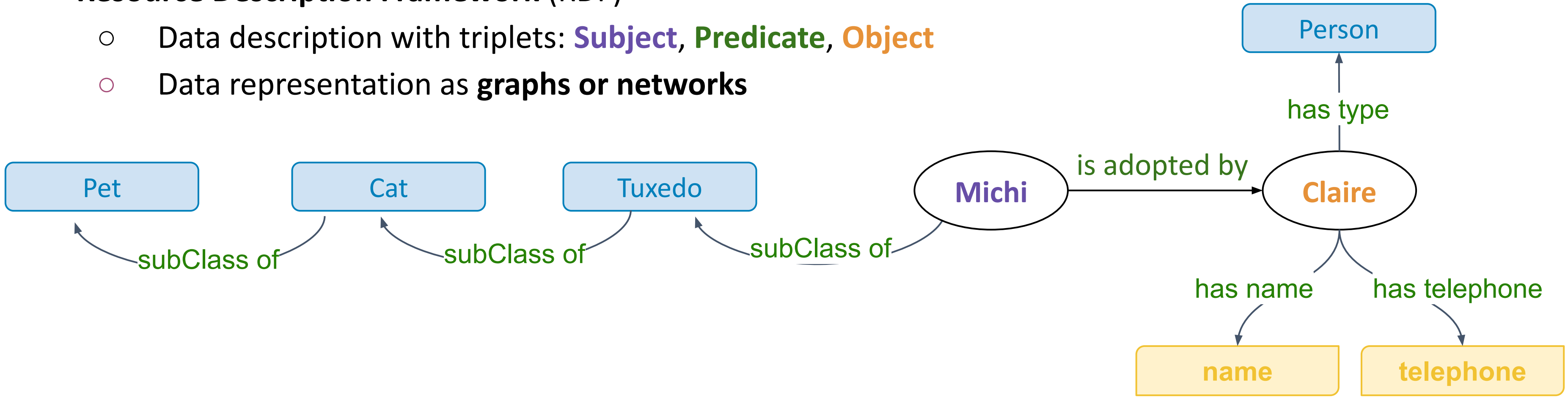
- **Resource Description Framework** (RDF)
    - ○ Data description with triplets: **Subject**, **Predicate**, **Object**



Ontologies: standards and representations

The **World Wide Web Consortium (W3C)** endorses multiple standards for metadata

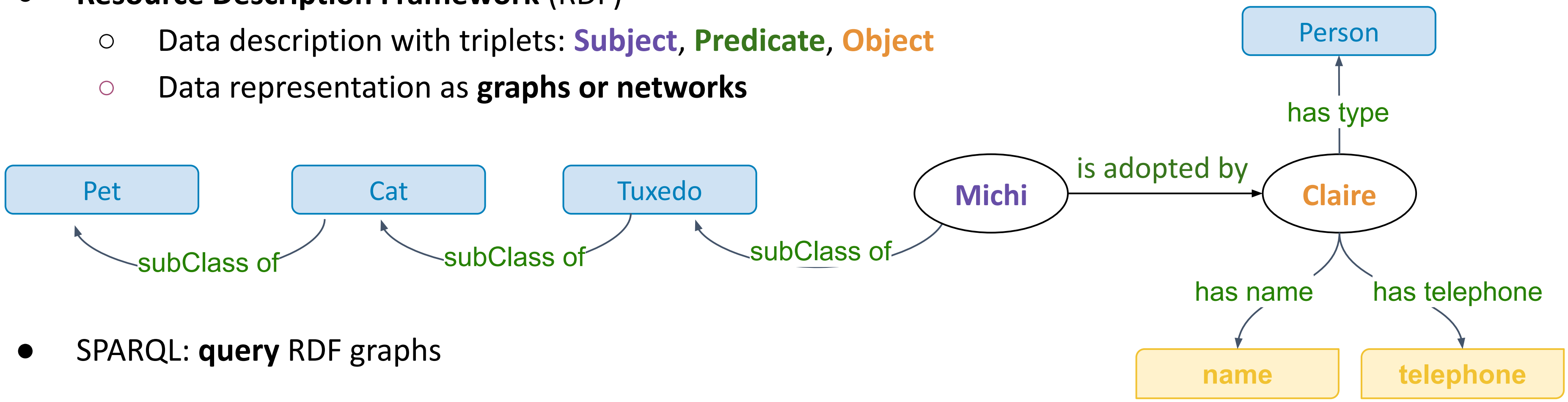
- **Resource Description Framework** (RDF)
  - Data description with triplets: **Subject**, **Predicate**, **Object**
  - Data representation as **graphs or networks**



Ontologies: standards and representations

The **World Wide Web Consortium (W3C)** endorses multiple standards for metadata

- **Resource Description Framework** (RDF)
  - Data description with triplets: **Subject**, **Predicate**, **Object**
  - Data representation as **graphs or networks**



- SPARQL: **query** RDF graphs

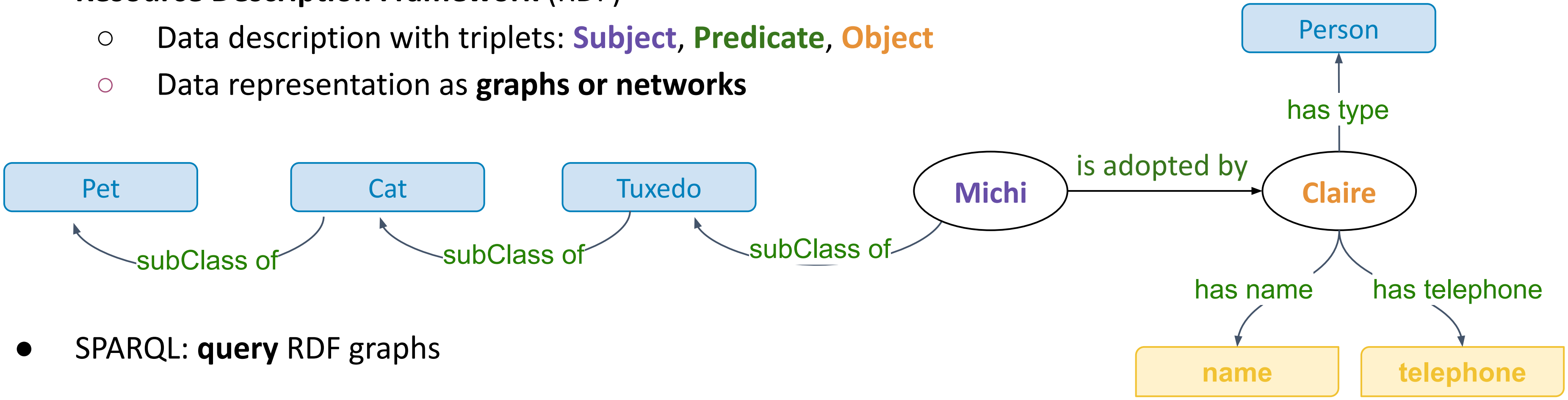
```
PREFIX po: <http://pet_ontolgy.org>
PREFIX sch: <http://schema.org>

SELECT   ?pet ?owner
WHERE { ?pet rdf:subClass_of po:Cat
        ?pet po:is_adopted_by ?owner
        ?owner sch:has_name Claire  }
```

Ontologies: standards and representations

The **World Wide Web Consortium (W3C)** endorses multiple standards for metadata

- **Resource Description Framework** (RDF)
  - Data description with triplets: **Subject**, **Predicate**, **Object**
  - Data representation as **graphs or networks**

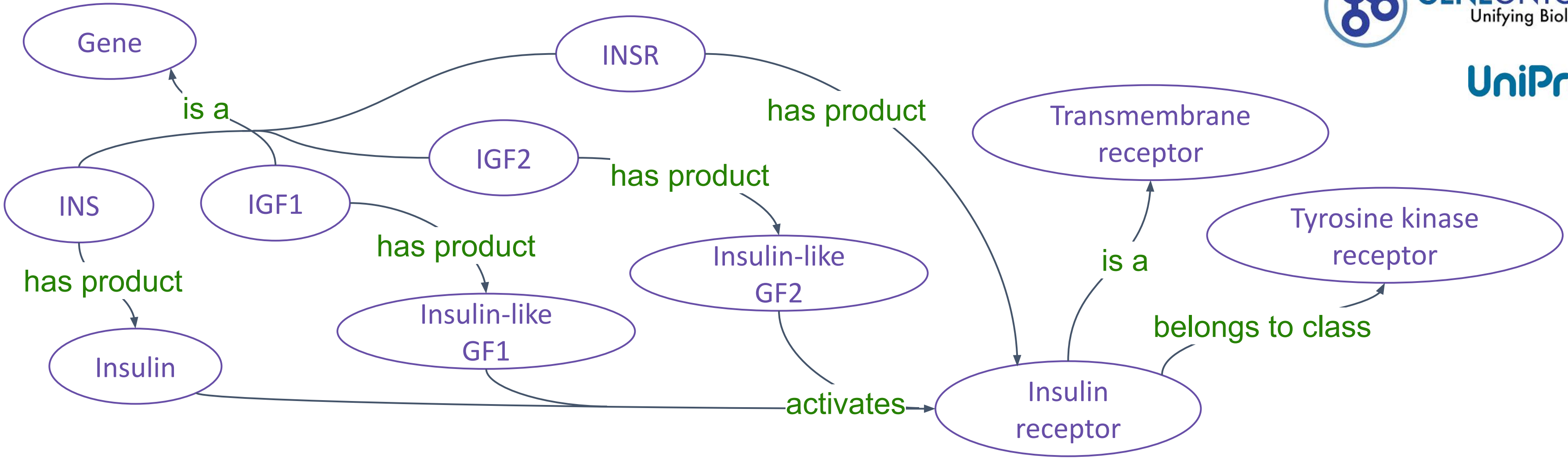


- SPARQL: **query** RDF graphs

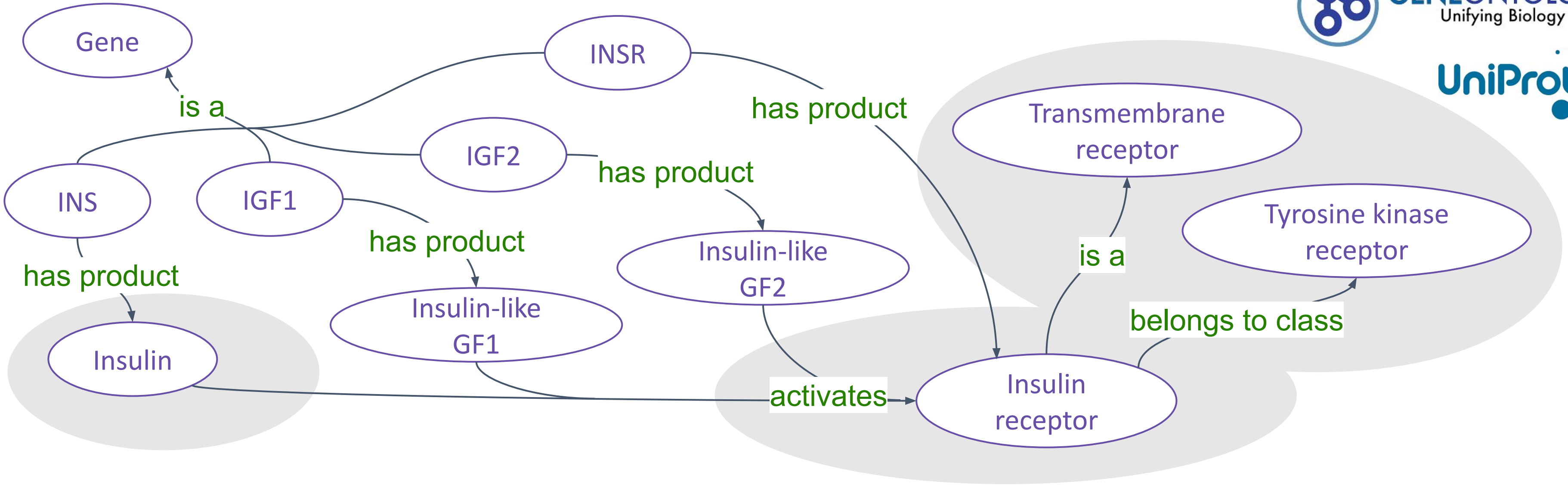
```
PREFIX po: <http://pet_ontolgy.org>
PREFIX sch: <http://schema.org>

SELECT   ?pet ?owner
WHERE { ?pet rdf:subClass_of po:Cat
        ?pet po:is_adopted_by ?owner
        ?owner sch:has_name Claire  }
```

UniprotKB and the Gene Ontology

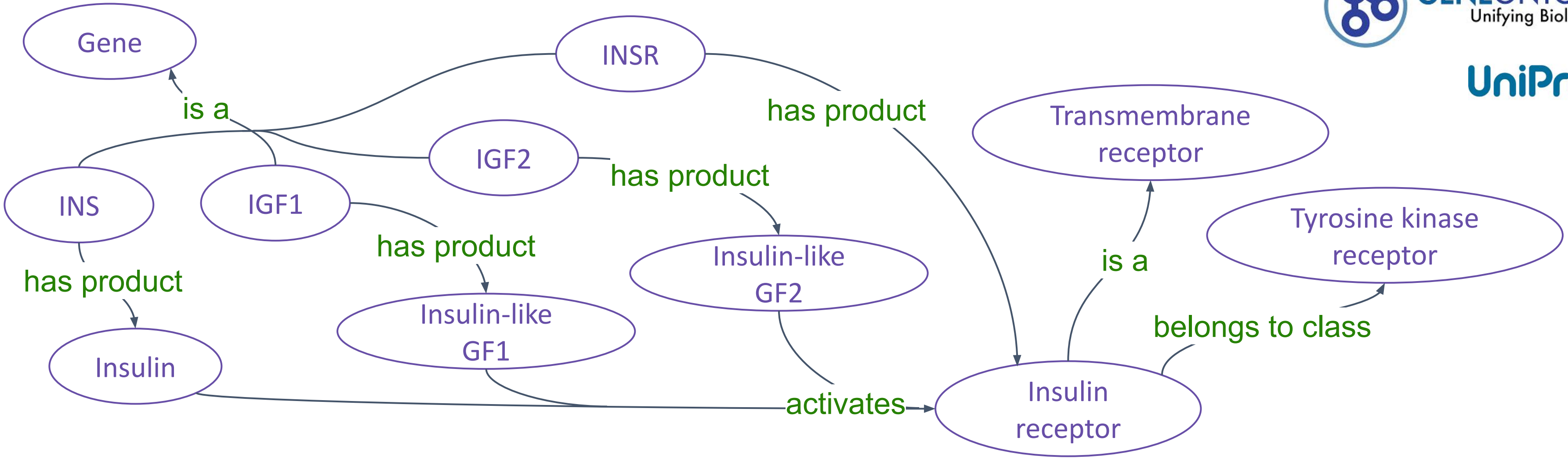


UniprotKB and the Gene Ontology



“The insulin receptor is a transmembrane receptor activated by insulin, IGF-I, IGF-II, and belongs to the class of tyrosine kinase receptors.”

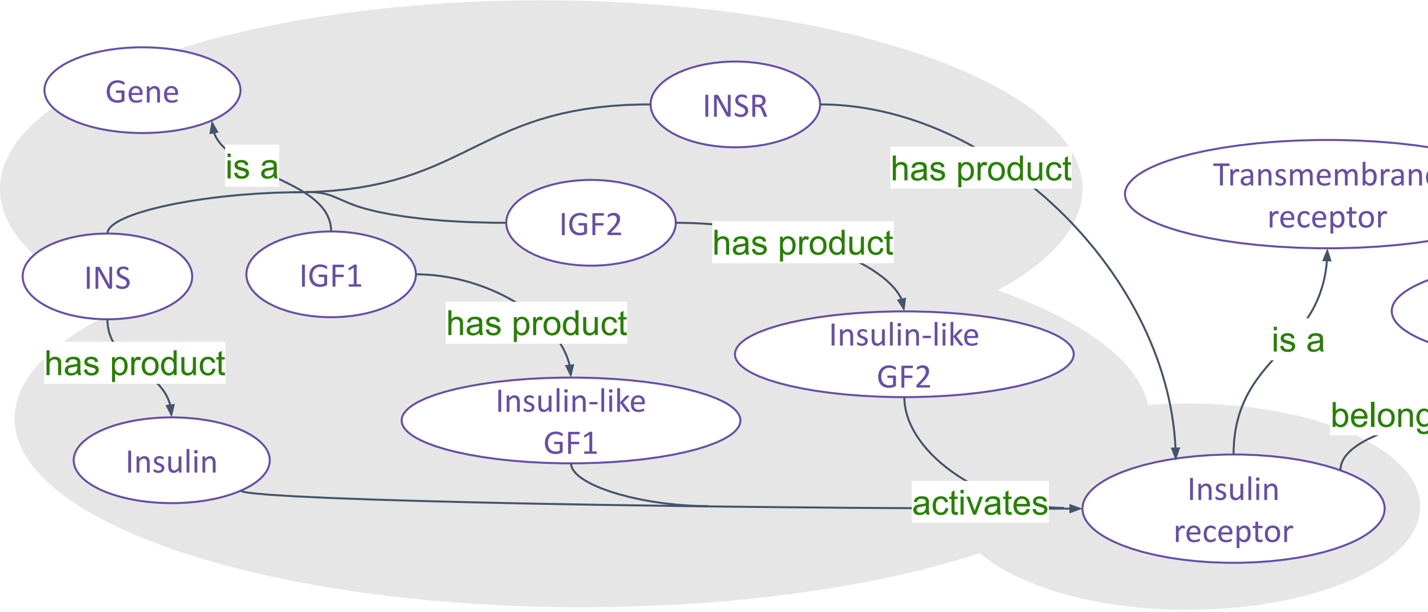
UniprotKB and the Gene Ontology



“The insulin receptor is a transmembrane receptor activated by insulin, IGF-I, IGF-II, and belongs to the class of tyrosine kinase receptors.”

What are the genes coding for the activators of insulin receptors?

UniprotKB and the Gene Ontology



“The insulin receptor is a transmembrane receptor activated by insulin, IGF-I, IGF-II, and belongs to the class of tyrosine kinase receptors.”

What are the genes coding for the activators of insulin receptors?

- INS, IGF1, IGF2

```
SELECT ?gene
WHERE {
  ?gene go:has_product ?protein
  ?protein upr:activates upr:insulin_receptor
}
```

Ontologies for life sciences

Numerous ontologies, from very general to very specific domains

- MeSH - Medical Subject Headings
  - Pubmed indexing and search
- HPO - Human Phenotype Ontology
  - Multi-lingual support
  - Layperson synonyms
- SO - Sequence Ontology
  - Sequence attributes, sequence features…
- EFO - Experimental Factor Ontology
  - Cell types, biological processes, protocols..
- Experimental biology: Cellosaurus, Microbial Conditions Ontology…
- Species specific: fish ontology, potato ontology, banana ontology…

🇬🇧Macroencephaly
🇯🇵大頭
🤷🏾Big head

The EDAM ontology for bioinformatics analyses

EDAM

Concepts about **data management and analysis** for life sciences, their **relations** and **attributes**
→ definition, permanent identifier, synonyms, references…

Organisation in **4 main classes**:

- **Topic**: field of study or technology → *Genomics, Sequencing*

The EDAM ontology for bioinformatics analyses

EDAM

Concepts about **data management and analysis** for life sciences, their **relations** and **attributes**
→ definition, permanent identifier, synonyms, references…

Organisation in **4 main classes**:

- **Topic**: field of study or technology → *Genomics, Sequencing*
- **Operation**: process or function → *DNA mapping, Peak calling*

The EDAM ontology for bioinformatics analyses

Concepts about **data management and analysis** for life sciences, their **relations** and **attributes**
→ definition, permanent identifier, synonyms, references…

Organisation in **4 main classes**:

- **Topic**: field of study or technology → *Genomics, Sequencing*
- **Operation**: process or function → *DNA mapping, Peak calling*
- **Data**: data type → *Gene ID, DNA sequence*

The EDAM ontology for bioinformatics analyses

EDAM

Concepts about **data management and analysis** for life sciences, their **relations** and **attributes**
→ definition, permanent identifier, synonyms, references…

Organisation in **4 main classes**:

- **Topic**: field of study or technology → *Genomics, Sequencing*
- **Operation**: process or function → *DNA mapping, Peak calling*
- **Data**: data type → *Gene ID, DNA sequence*
- **Format**: file format → *FASTQ, BAM, JSON*

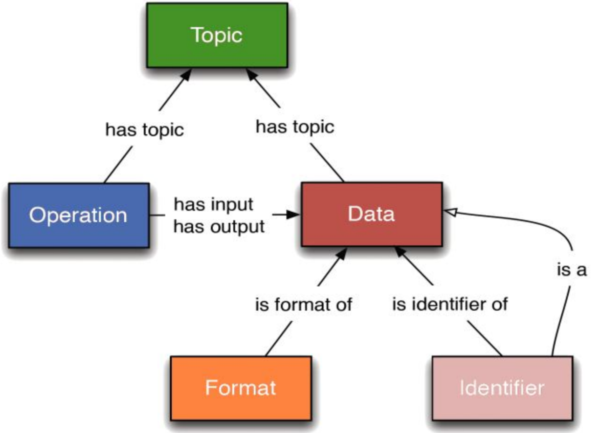
The EDAM ontology for bioinformatics analyses

EDAM

Concepts about **data management and analysis** for life sciences, their **relations** and **attributes**
→ definition, permanent identifier, synonyms, references…

Organisation in **4 main classes**:

- **Topic**: field of study or technology → *Genomics, Sequencing*
- **Operation**: process or function → *DNA mapping, Peak calling*
- **Data**: data type → *Gene ID, DNA sequence*
- **Format**: file format → *FASTQ, BAM, JSON*

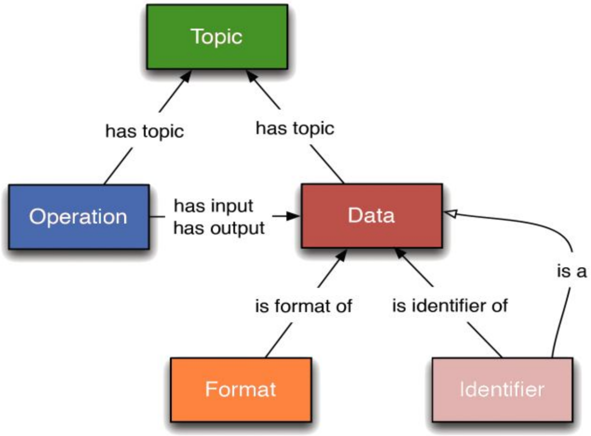


The EDAM ontology for bioinformatics analyses

Concepts about **data management and analysis** for life sciences, their **relations** and **attributes**
→ definition, permanent identifier, synonyms, references…

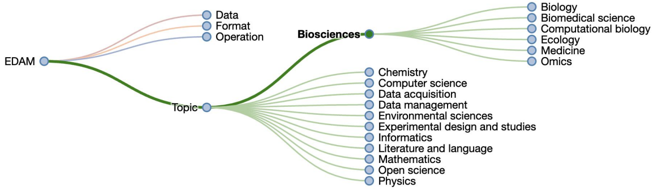
Organisation in **4 main classes**:

- **Topic**: field of study or technology → *Genomics, Sequencing*
- **Operation**: process or function → *DNA mapping, Peak calling*
- **Data**: data type → *Gene ID, DNA sequence*
- **Format**: file format → *FASTQ, BAM, JSON*

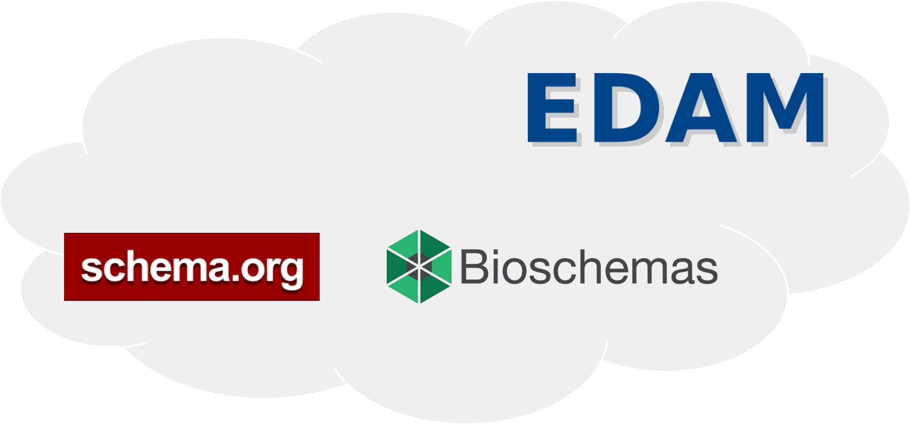


Development

- **Open and collaborative** via Github or the GUI Protégé/WebProtégé
- **Visualisation** and navigation with the EDAM browser
- **Uninterrupted**: knowledge never stops growing!



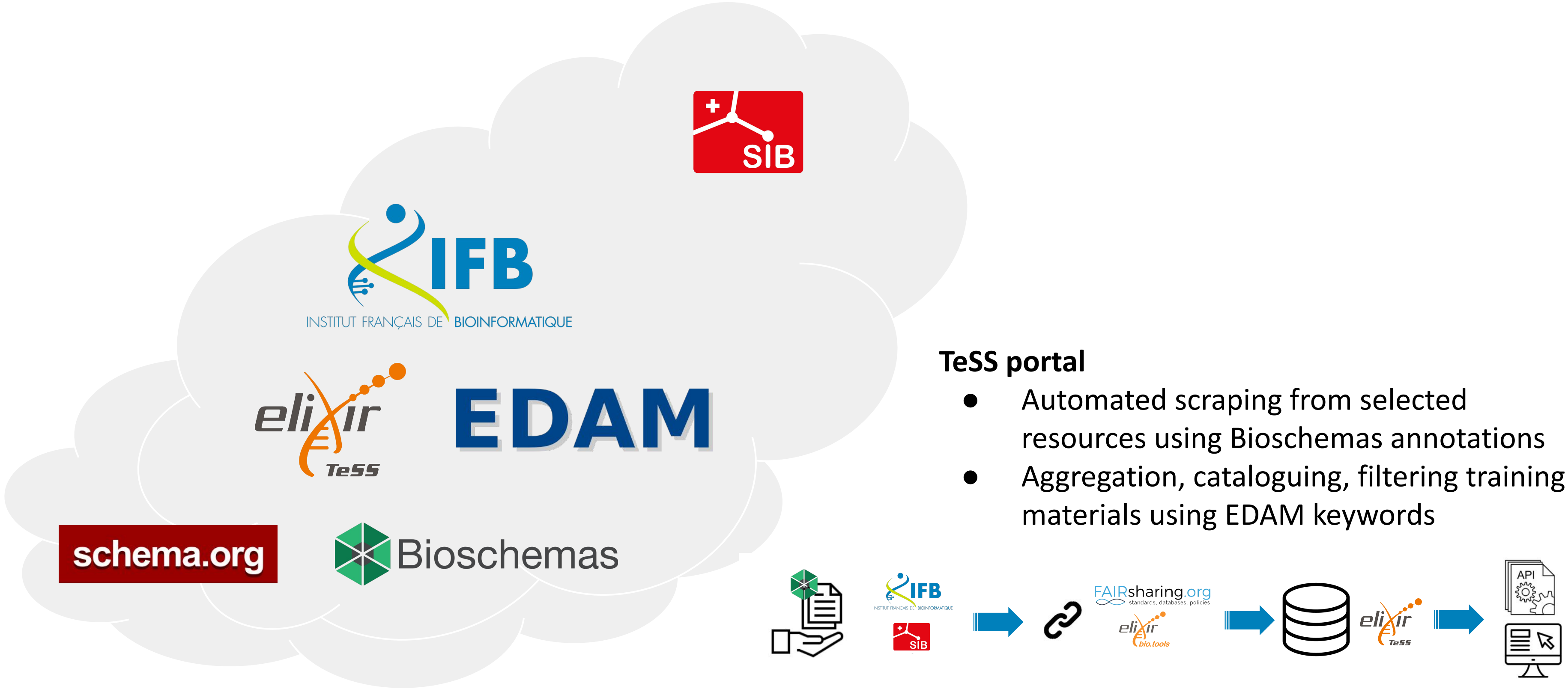
Ontology applications - metadata schemes



Bioschemas

- Metadata schemes using schema.org
- Definition of biology-oriented profiles using EDAM terms

Ontology applications - aggregation and filtering of resources



TeSS portal

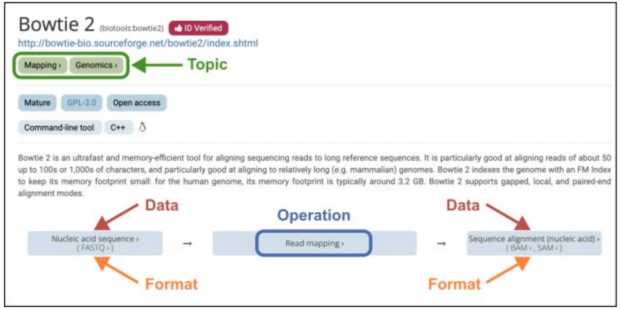
- Automated scraping from selected resources using Bioschemas annotations
- Aggregation, cataloguing, filtering training materials using EDAM keywords

Ontology applications - classification and data extraction

SIB
Expasy
Swiss Bioinformatics Resource Portal
IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE
elixir TeSS
EDAM
schema.org
Bioschemas
elixir bio.tools

Bio.tools catalogue

- 30,000 tools annotated
- Cataloguing, filtering, extracting via an API or a web portal



Bowtie 2 (biotools:bowtie2) ID Verified

http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

Mapping › Genomics ›

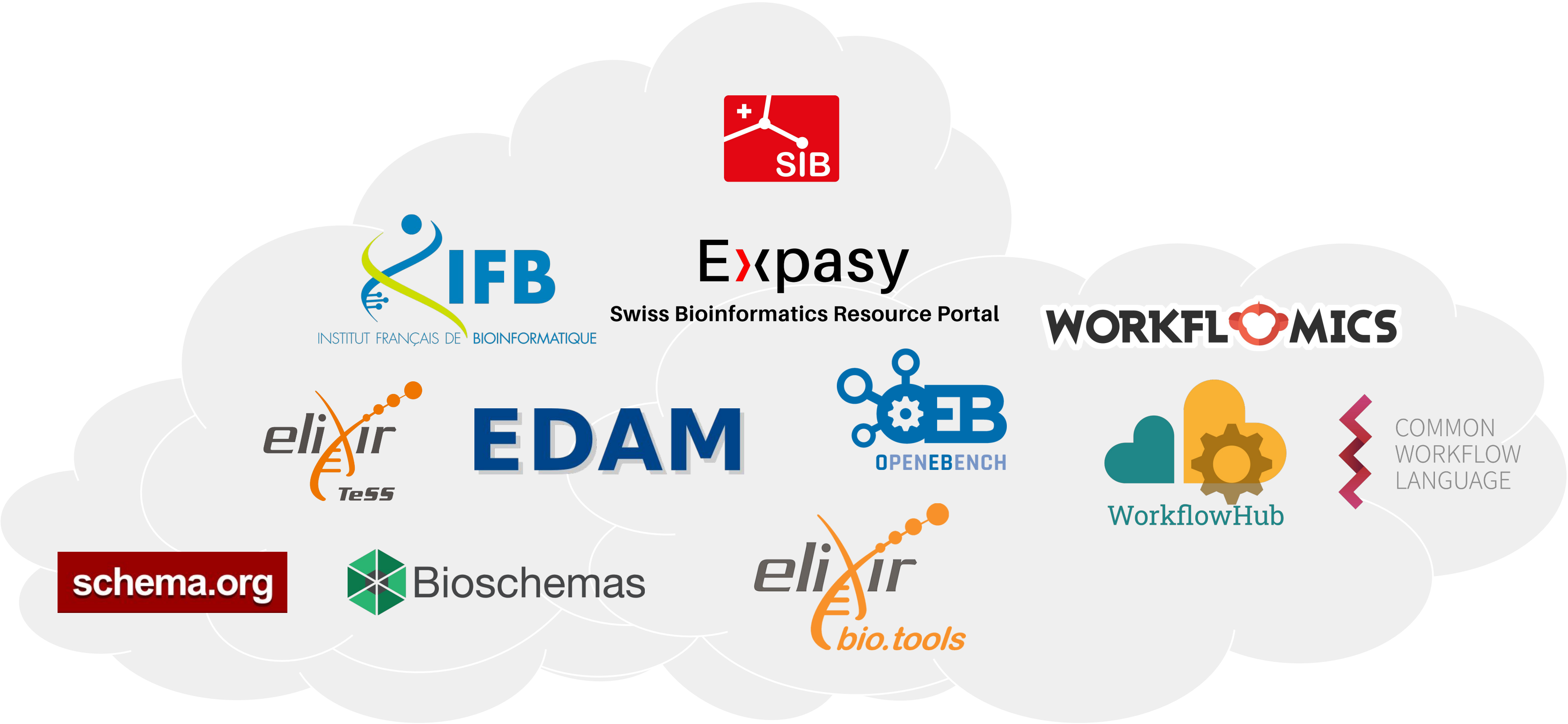
Mature | GPL-3.0 | Open access

Command-line tool | C++

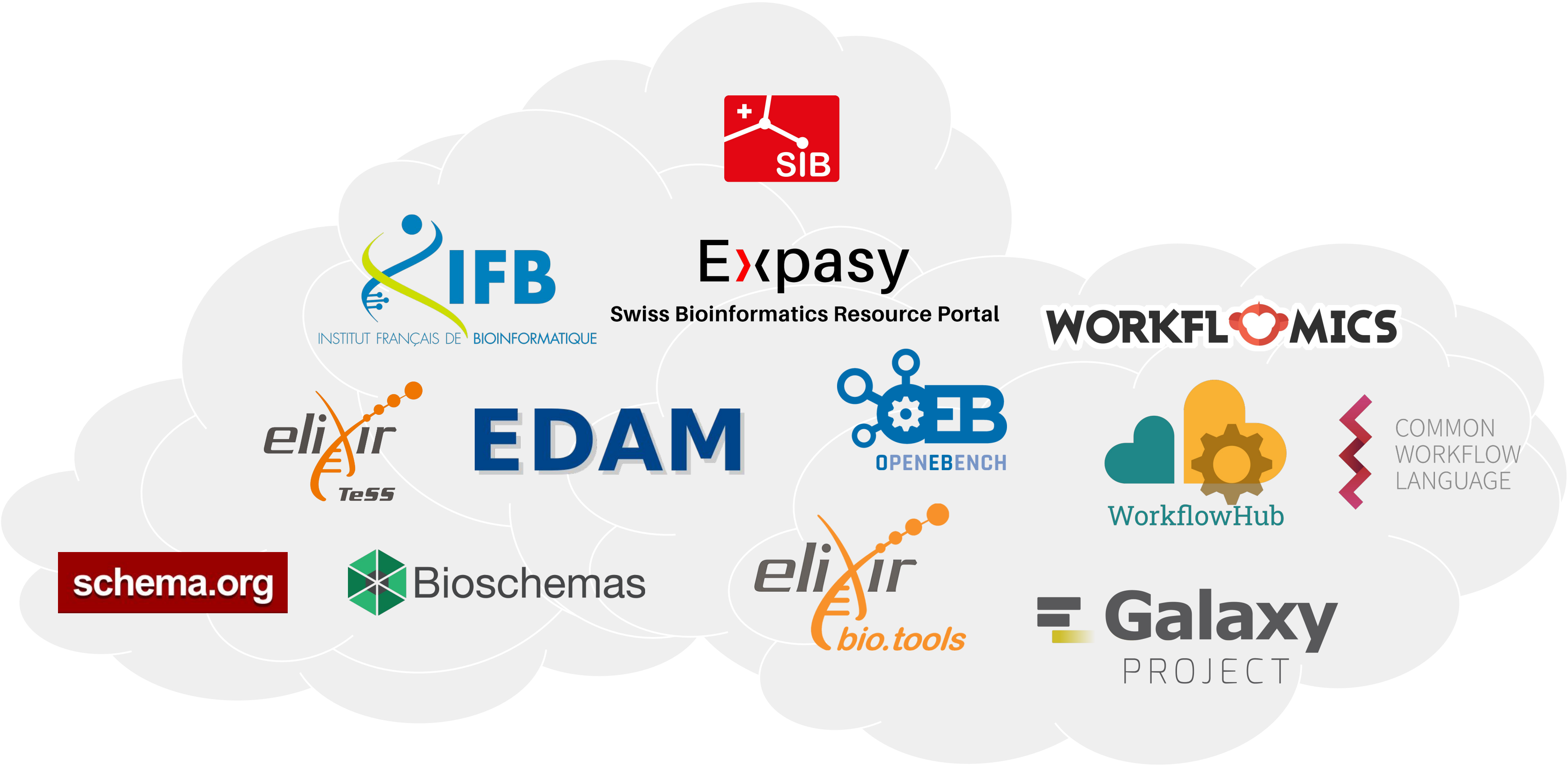
Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

Nucleic acid sequence › (FASTQ ›) → Read mapping › → Sequence alignment (nucleic acid) › (BAM › , SAM ›)

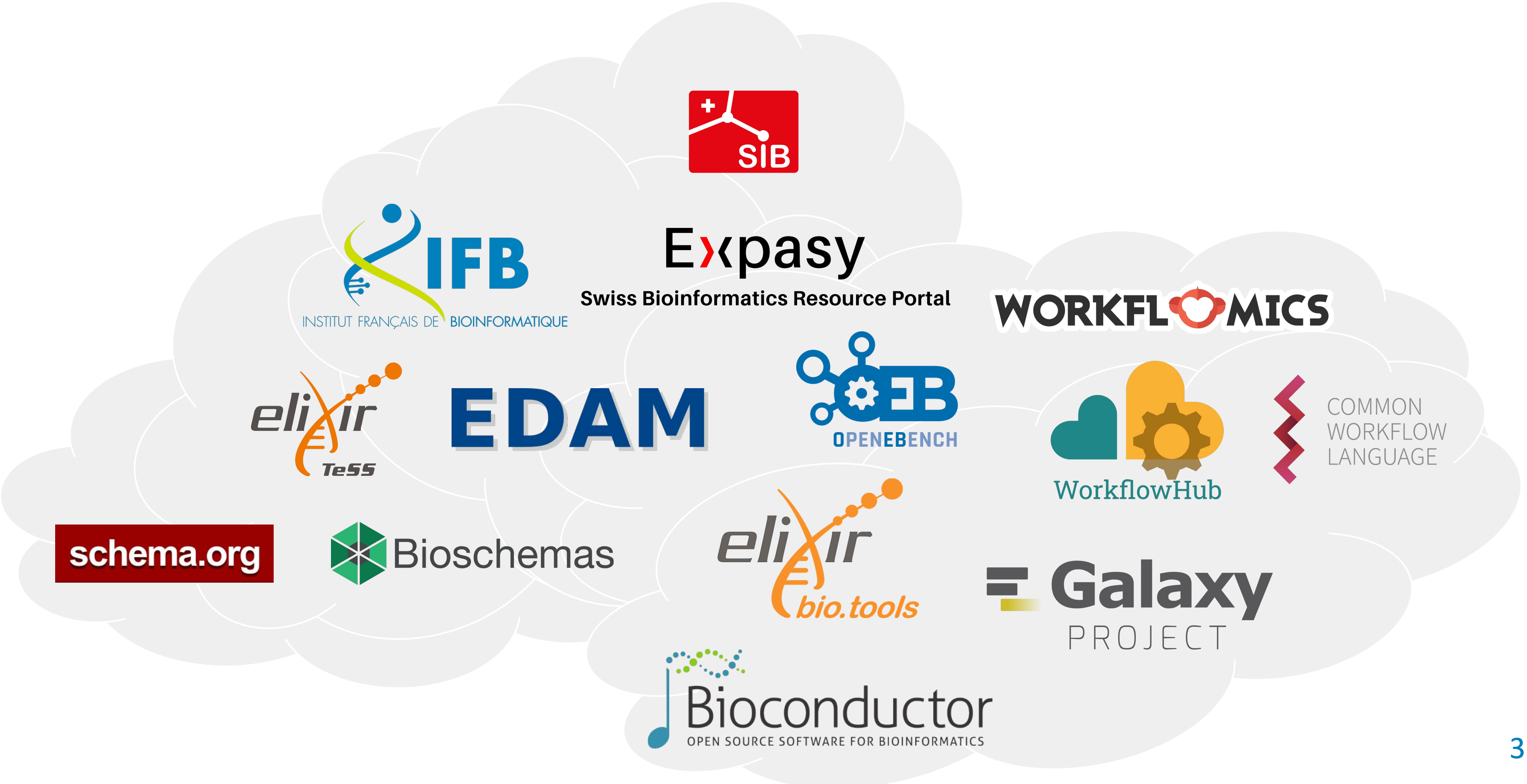
Ontology applications - design and evaluation of workflows



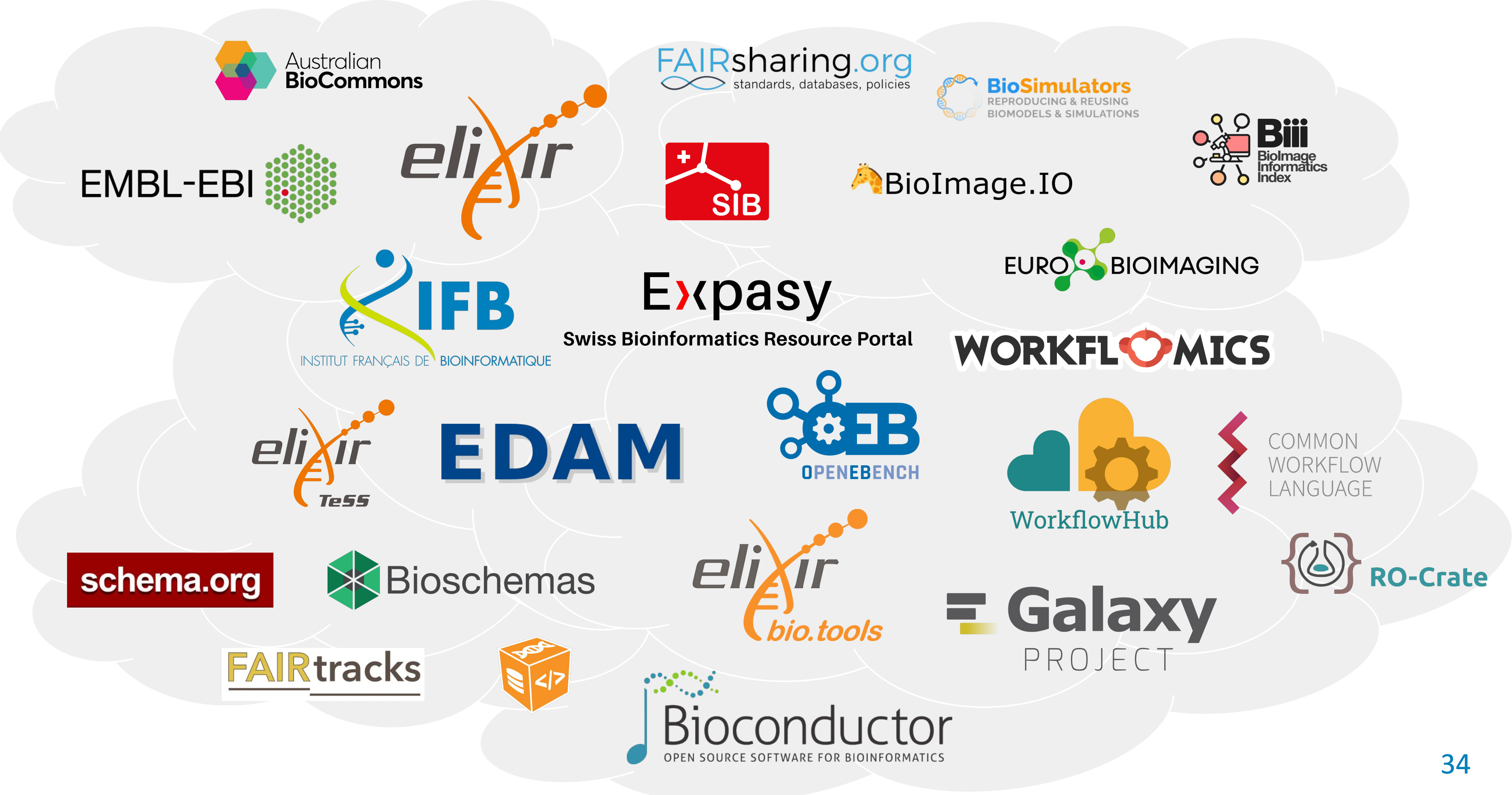
Ontology applications - synchronisation across platforms



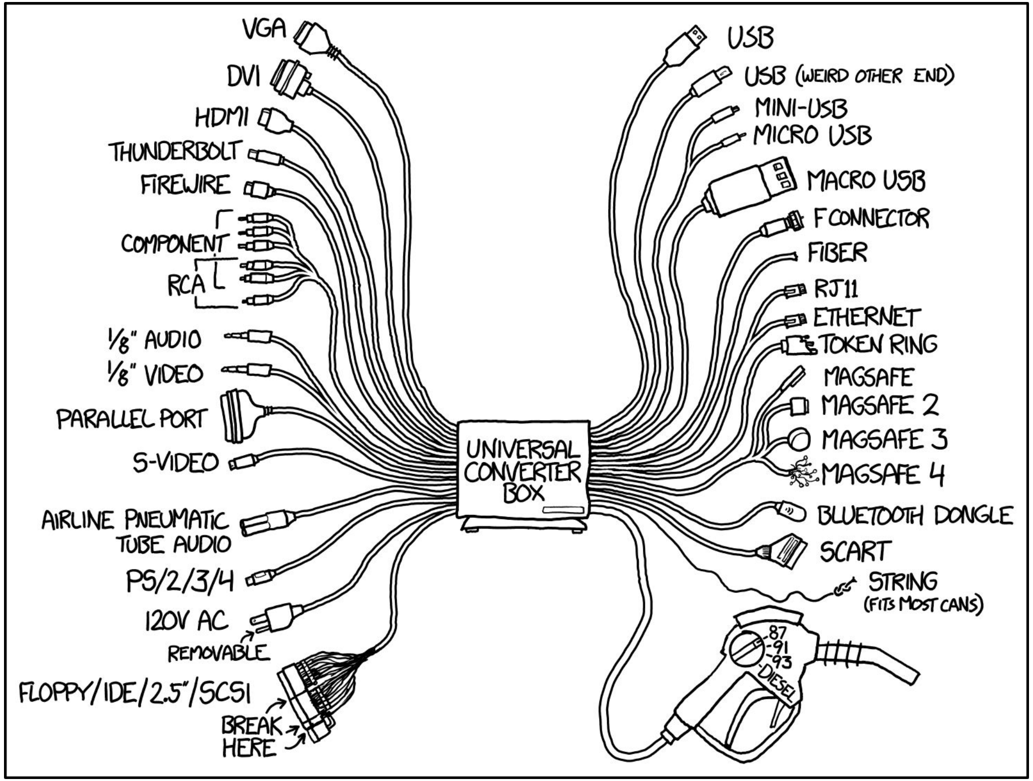
Ontology applications - automated annotation of libraries



Ontology applications - interoperability among a wide ecosystem


Australian BioCommons
FAIRsharing.org standards, databases, policies
BioSimulators REPRODUCING & REUSING BIOMODELS & SIMULATIONS
EMBL-EBI
elixir
SIB
BioImage.IO
Biii BioImage Informatics Index
IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE
Expasy Swiss Bioinformatics Resource Portal
EURO BIOIMAGING
WORKFLOMICS
elixir TeSS
EDAM
OEB OPENEBENCH
WorkflowHub
COMMON WORKFLOW LANGUAGE
schema.org
Bioschemas
elixir bio.tools
RO-Crate
Galaxy PROJECT
FAIRtracks
Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Ontology applications - universal converter box



https://xkcd.com/1406

Ontologies: to infinity and beyond

Using ontology-based metadata has many more applications

- Data **integration** from heterogeneous sources
- Knowledge representation as **graphs or networks**
- Knowledge **discovery**, **predictions**, hypotheses
- **Inferring**, **querying**, **reasoning**

Ontologies: to infinity and beyond

Using ontology-based metadata has many more applications

- Data **integration** from heterogeneous sources
- Knowledge representation as **graphs or networks**
- Knowledge **discovery**, **predictions**, hypotheses
- **Inferring**, **querying**, **reasoning**

Ontologies provide a basis for **semantic web** and technologies

- Defined by Tim Berners-Lee as a “**web of data**” ≠ “web of documents”
- Allows for the *meaning* of data to be **machine-readable**

Ontologies: to infinity and beyond

Using ontology-based metadata has many more applications

- Data **integration** from heterogeneous sources
- Knowledge representation as **graphs or networks**
- Knowledge **discovery**, **predictions**, hypotheses
- **Inferring**, **querying**, **reasoning**

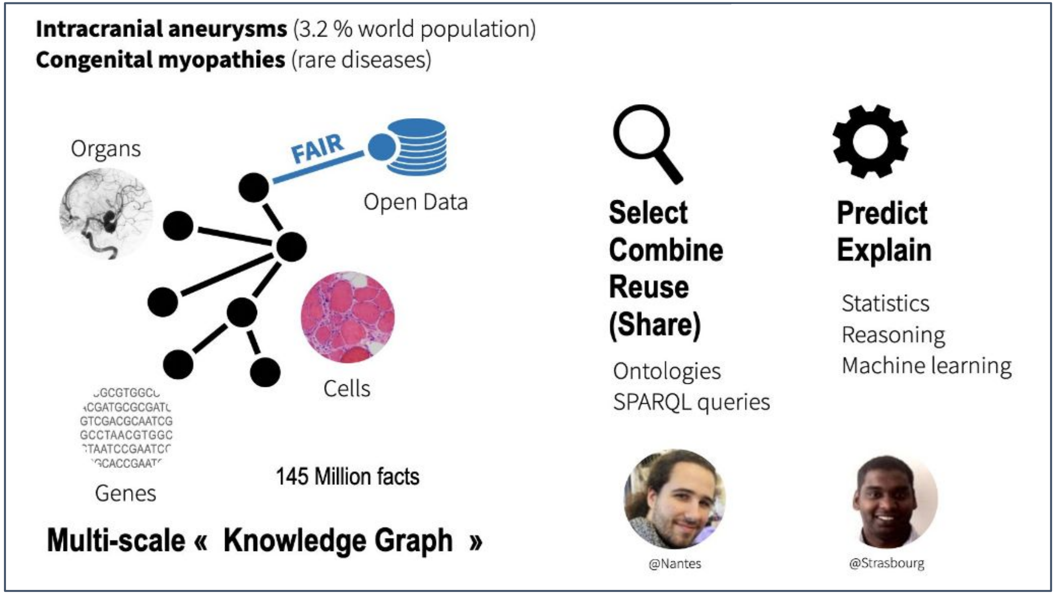
Ontologies provide a basis for **semantic web** and technologies

- Defined by Tim Berners-Lee as a “**web of data**” ≠ “web of documents”
- Allows for the *meaning* of data to be **machine-readable**
- Semantic technologies:
    - **Natural Language Processing** (NLP): speech recognition, biocuration
    - **Large Language Models** (LLMs): generative AI tools such as ChatGPT, Gemini
    - **Image analysis**: mass spec, spatial omics, microscopy…

Data integration, image analysis and prognosis

INEX-MED: bridging imaging-omics-clinical data for the study of intracranial aneurysms

- ICAN cohort: 3,400 subjects, 3,000 MRIs, 800 whole genomes
- Association between imaging phenotypes and omics signatures?
- Patients with higher/lower risks of aneurysm rupturing?



Intracranial aneurysms (3.2 % world population)
Congenital myopathies (rare diseases)

145 Million facts

Multi-scale « Knowledge Graph »

Select
Combine
Reuse
(Share)

Ontologies
SPARQL queries

Predict
Explain

Statistics
Reasoning
Machine learning

@Nantes

@Strasbourg

Source: course on semantic web by A. Gaignard, P. Larmande and O. Dameron - ETBII 2024

Take-away and resources

Ontologies and semantic metadata

- Annotation of diverse resources: data, articles, software tools, training materials and events…
- Improve data interoperability and accessibility
- Integration and analysis of heterogeneous data

Take-away and resources

Ontologies and semantic metadata

- Annotation of diverse resources: data, articles, software tools, training materials and events…
- Improve data interoperability and accessibility
- Integration and analysis of heterogeneous data

Useful resources

- **OBO foundry**: community development of interoperable ontologies for biological sciences
- **Ontology Lookup Service (OLS)** and **BioPortal**: access and search biomedical ontologies
- **Ontotext** learning resources: what are ontologies and semantic web?
- FAIR-Checker: improve the FAIRness of your web resources (Gaignard et al., 2023)

Links links links

elixir

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

SIB

EMBL-EBI

Australian BioCommons

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

EDAM

Bioschemas

schema.org

FAIRtracks

elixir TeSS

FAIRsharing.org
standards, databases, policies

elixir bio.tools

Expasy
Swiss Bioinformatics Resource Portal

COMMON WORKFLOW LANGUAGE

WorkflowHub

WORKFLOMICS

Galaxy PROJECT

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

OEB
OPENEBENCH

RO-Crate

BioImage.IO

BioSimulators
REPRODUCING & REUSING BIOMODELS & SIMULATIONS

EURO BIOIMAGING

Biii
BioImage Informatics Index

The Sequence Ontology

GENEONTOLOGY
Unifying Biology

human phenotype ontology

MeSH

efo

UniProt

W3C®

OLS

BioPortal

ontotext

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

Thank you!

biotic	Biotic organisation committee
EDAM	Hervé Ménager, Alban Gaignard, Matúš Kalaš
IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE	Open science & interoperability Software tools
elixir	Tools platform (WP2, WP3) Interoperability platform
Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS	Maria Doyle, Vincent Carey

cnrs INRAE Inserm cea FRANCE 2030