Topic 9. Data integration

Why, how, promise and pitfalls. Standards, ontologies, ID mapping and metadata

The Genotype to Phenotype challenge

Topic 5 – Genes & Genomes

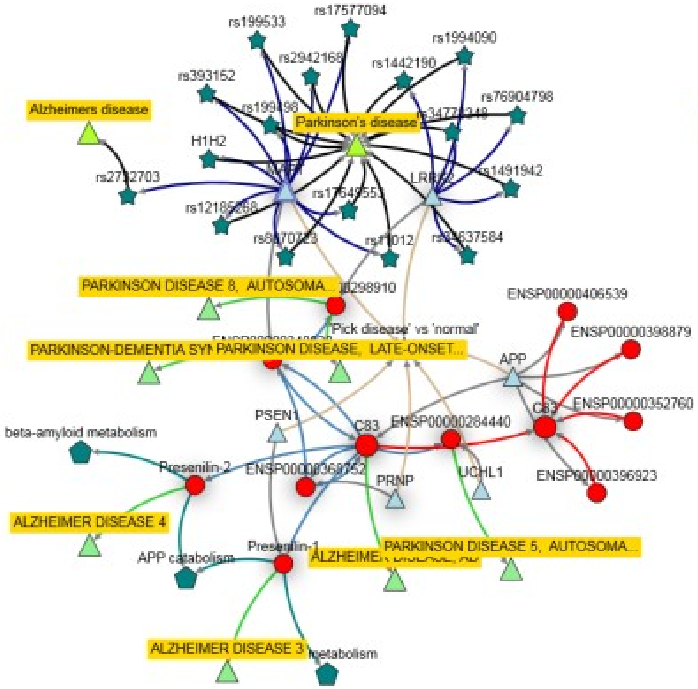
Topic 6 – Functional genomics

Topic 7 – Networks & Pathways

Topic 8 – Phenotypes & Diseases

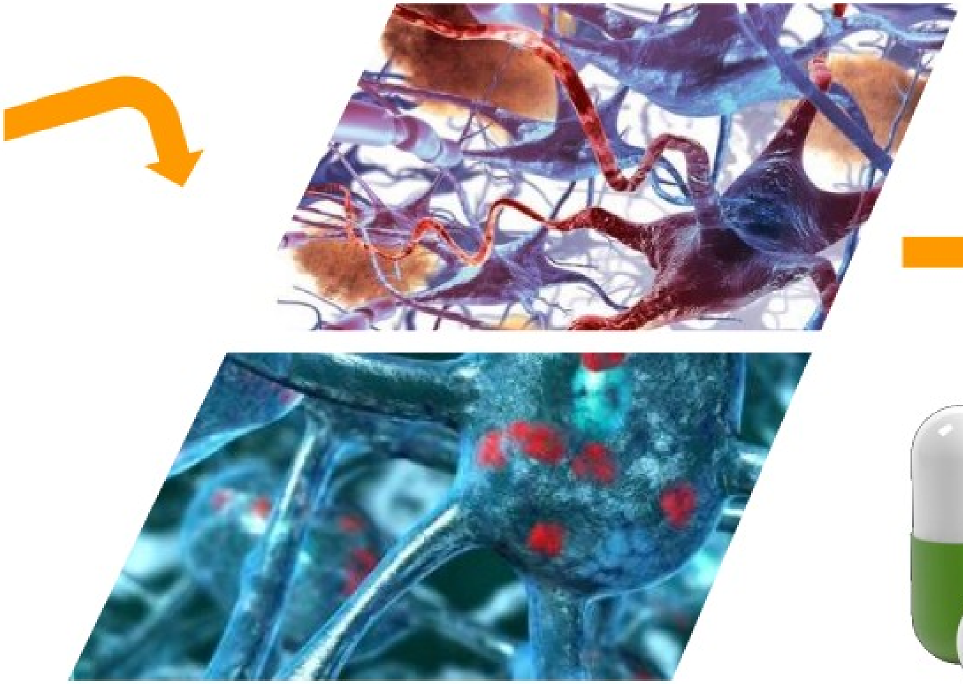
Genotype

GWAS, QTL



Biological Knowledge Discovery

Data selection, processing, transformation, integration, interpretation



Phenotype

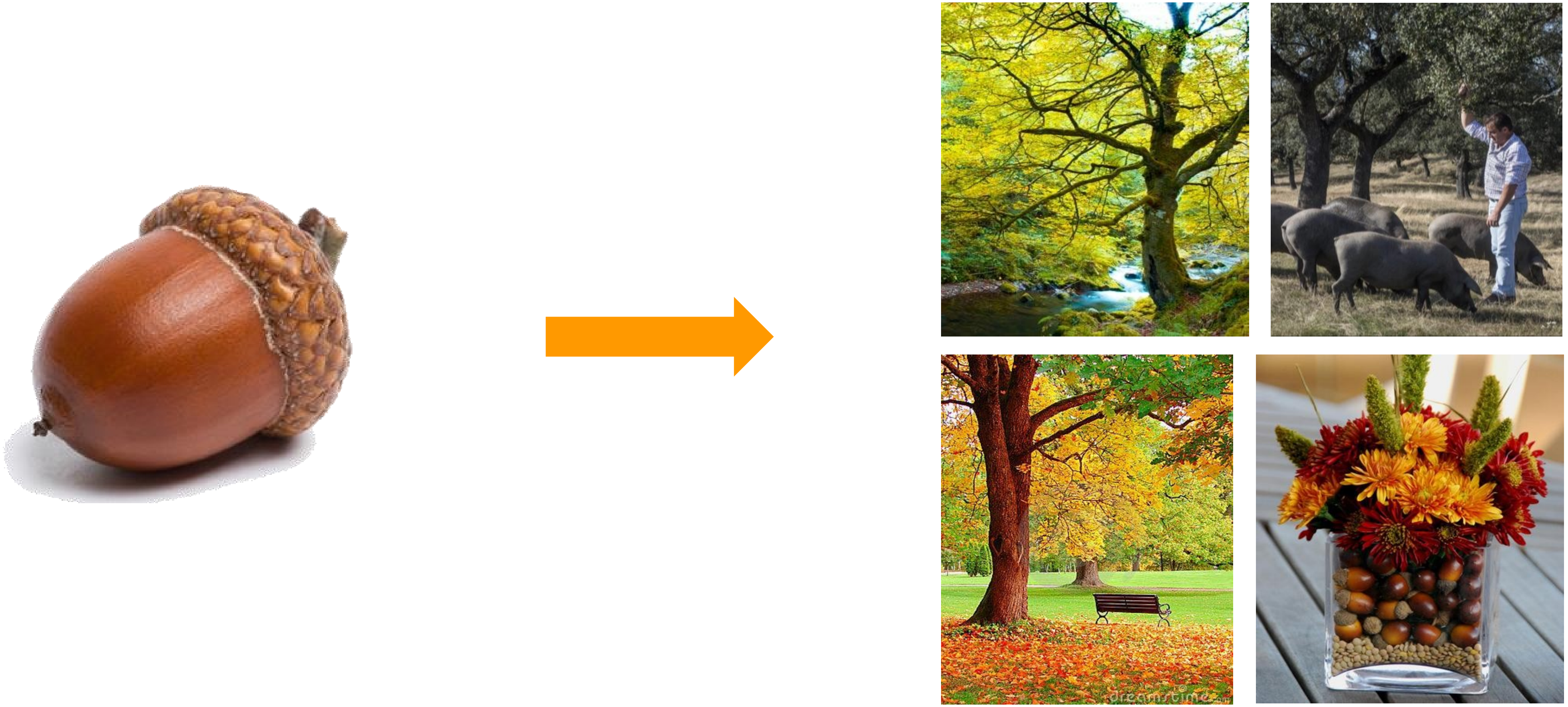
Alzheimer, Parkinson

Drugs

Precision medicine

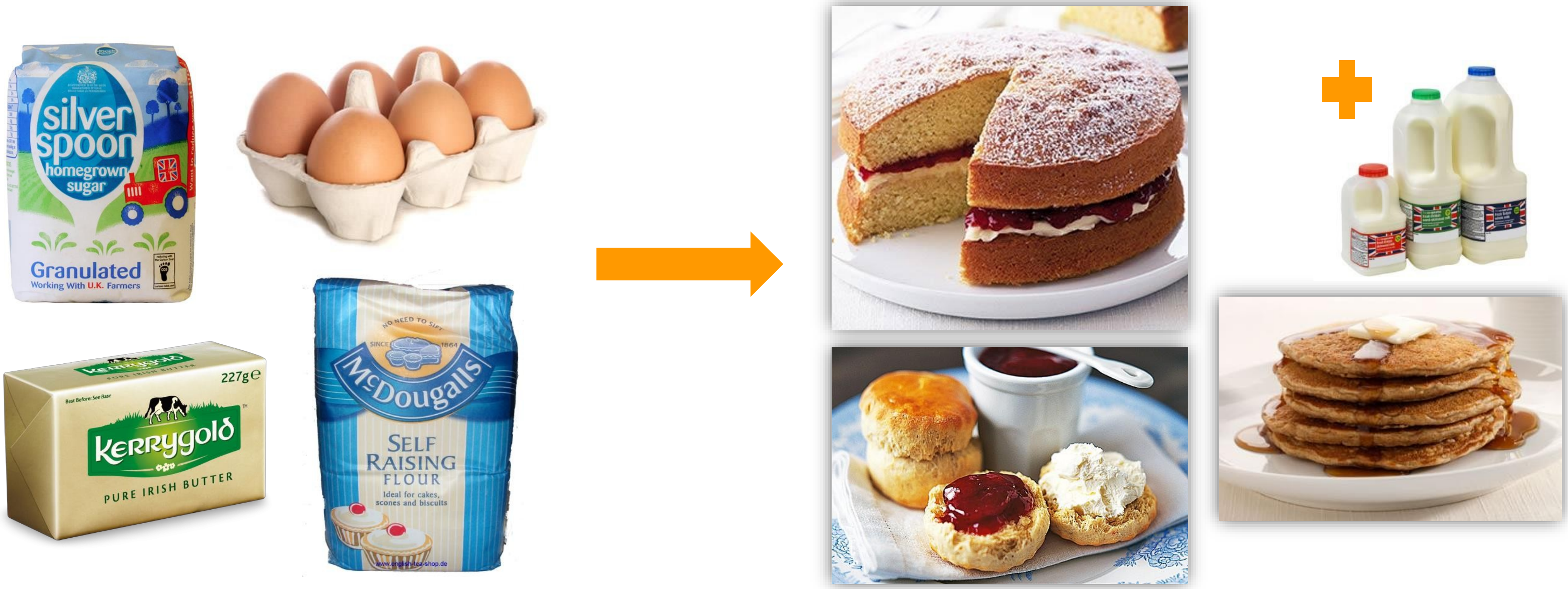
Data integration: some important considerations

1. You need to understand the limitations/context of the data you start with, to be able to get something useful from a larger integrated dataset



Data integration: some important considerations

2. **Even when you have the same data to work with, integrating it in different ways can give completely different views of the bigger biological picture**

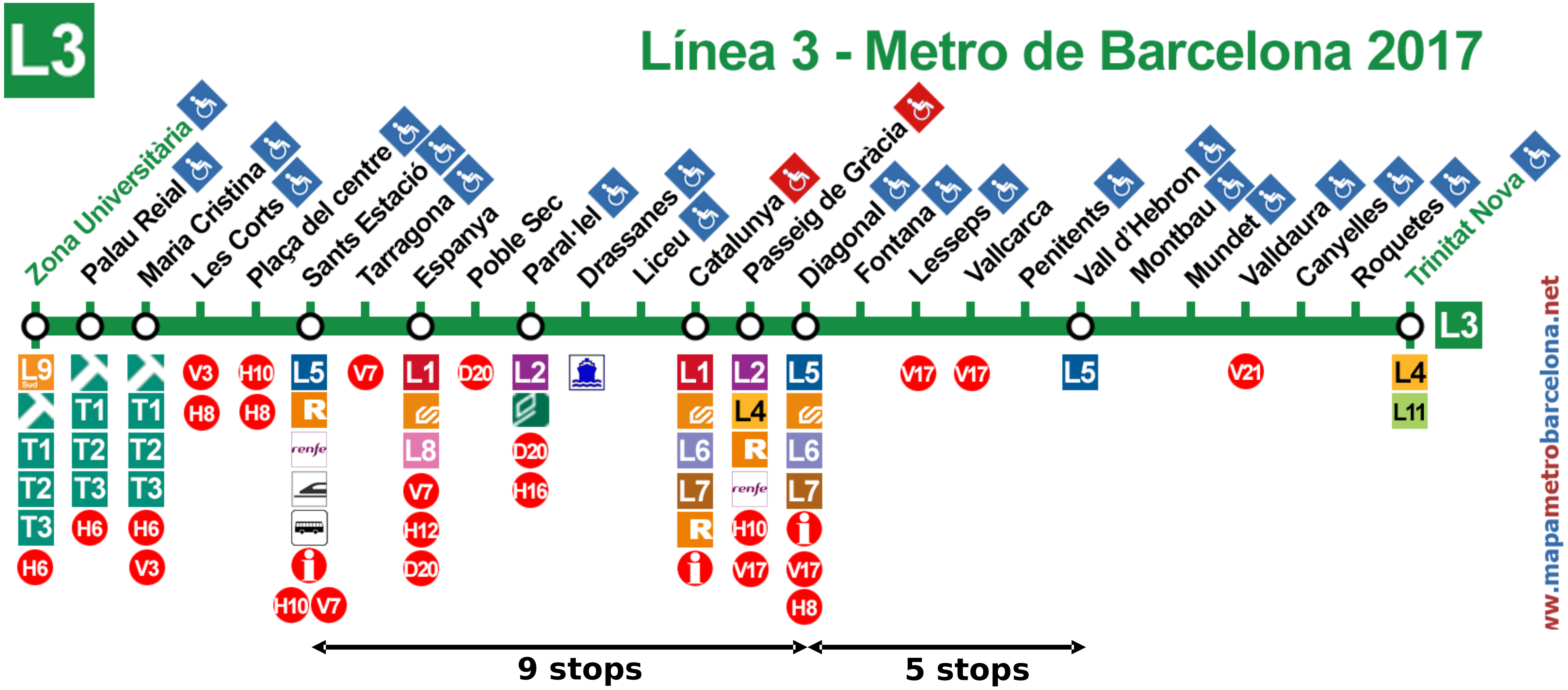


Data integration: some important considerations

3. **Even if you understand the data and have a good knowledge of the biology, integrating data without the right skills can give a poor result or false impression of reality**

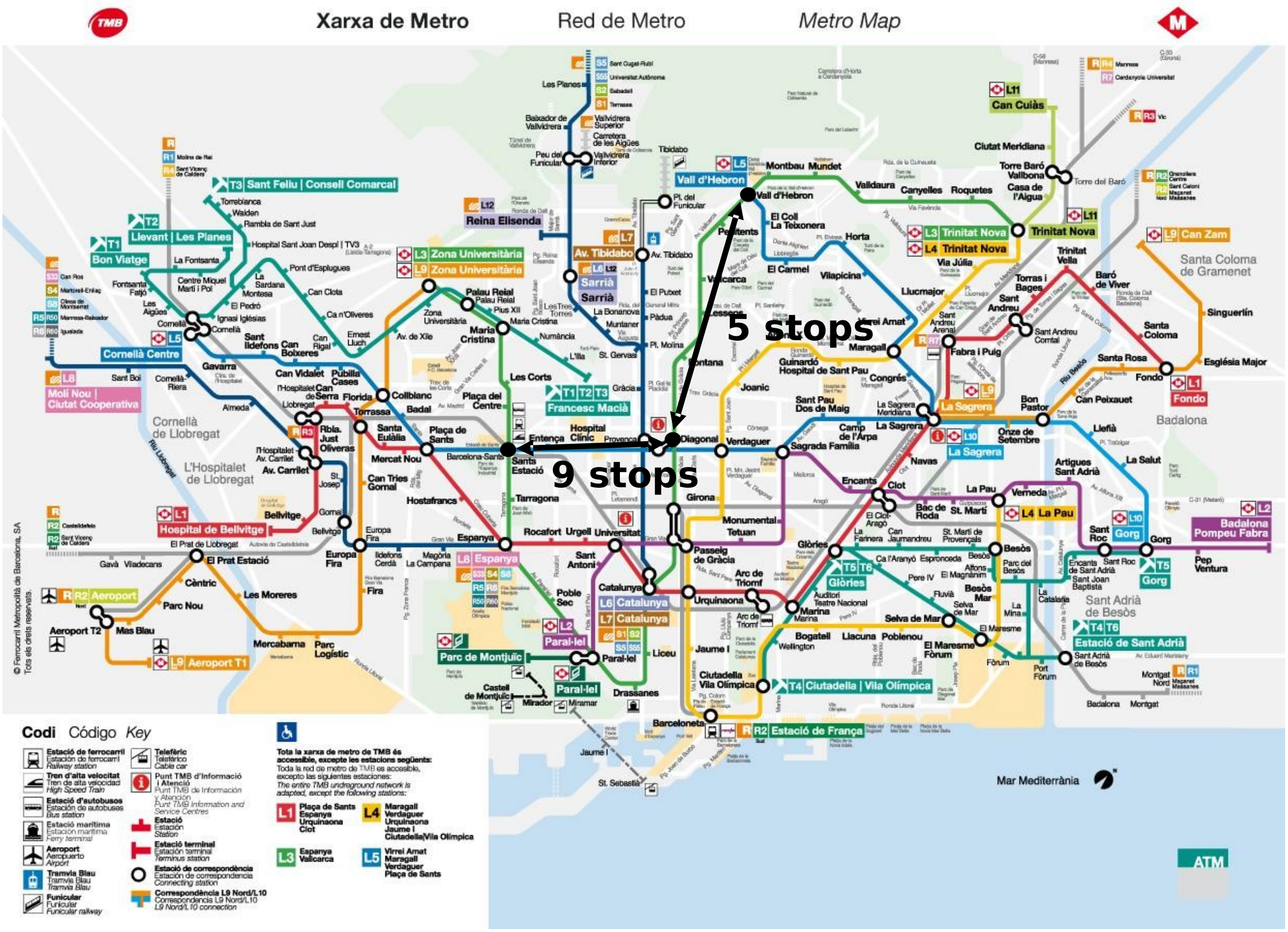
Data integration: some important considerations

4. How (and which) data is represented has a significant effect on our understanding - aiding us in some areas, but can mis-represent the truth in others



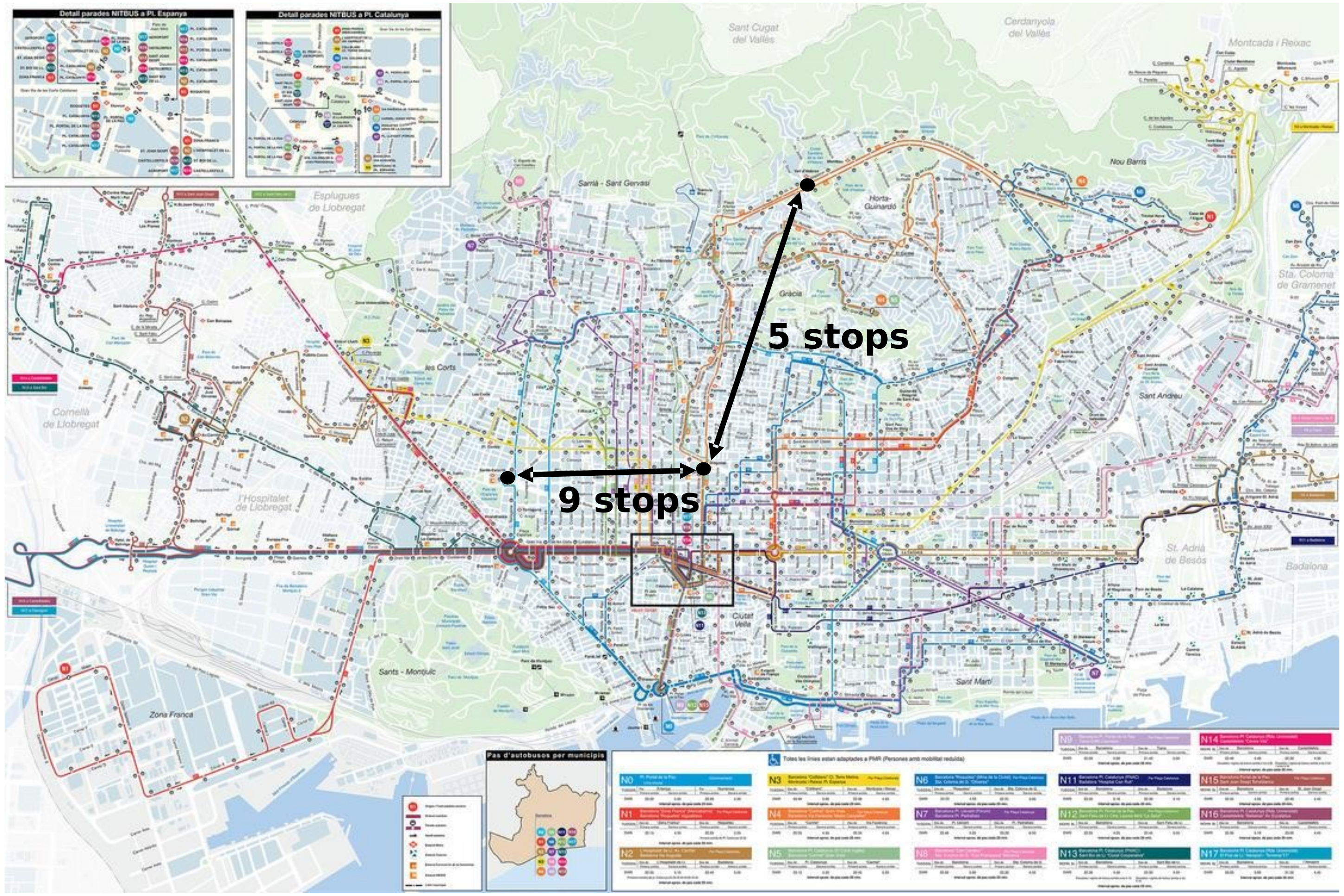
Data integration: some important considerations

4. **How (and which) data is represented has a significant effect on our understanding - aiding us in some areas, but can mis-represent the truth in others**



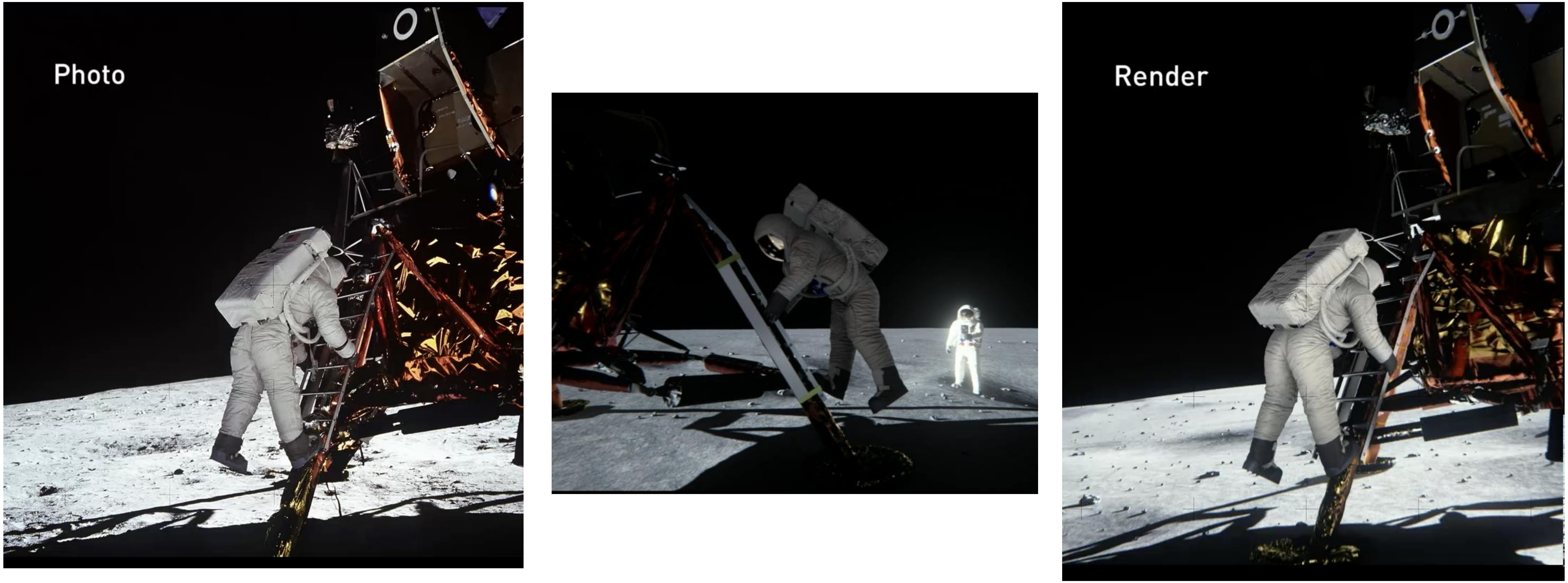
Data integration: some important considerations

4. **How (and which) data is represented has a significant effect on our understanding - aiding us in some areas, but can mis-represent the truth in others**



Data integration: some important considerations

5. **We do not (usually) have all the data - unless we recognize this we can be led to false conclusions.**



Debunking Lunar Landing Conspiracies with Maxwell and VXGI
https://www.youtube.com/watch?v=O9y_AVYMEUs

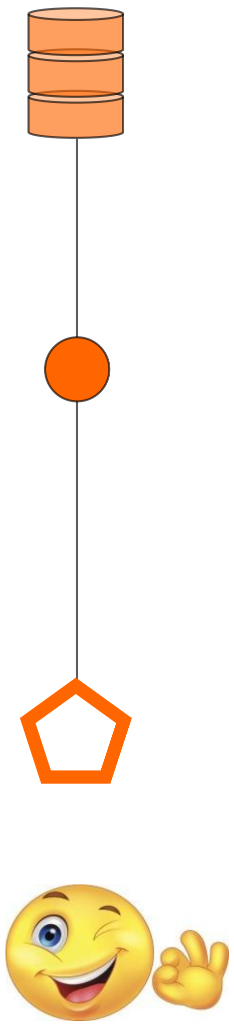
Data integration: some important considerations

Big data, bioinformatics, and particularly data integration have the potential to help us understand biology better.

However, it never provides a definitive answer - it can be used to guide experimentation; and ideas developed from *in silico* approaches should be validated by laboratory (and preferably orthogonal) methods.

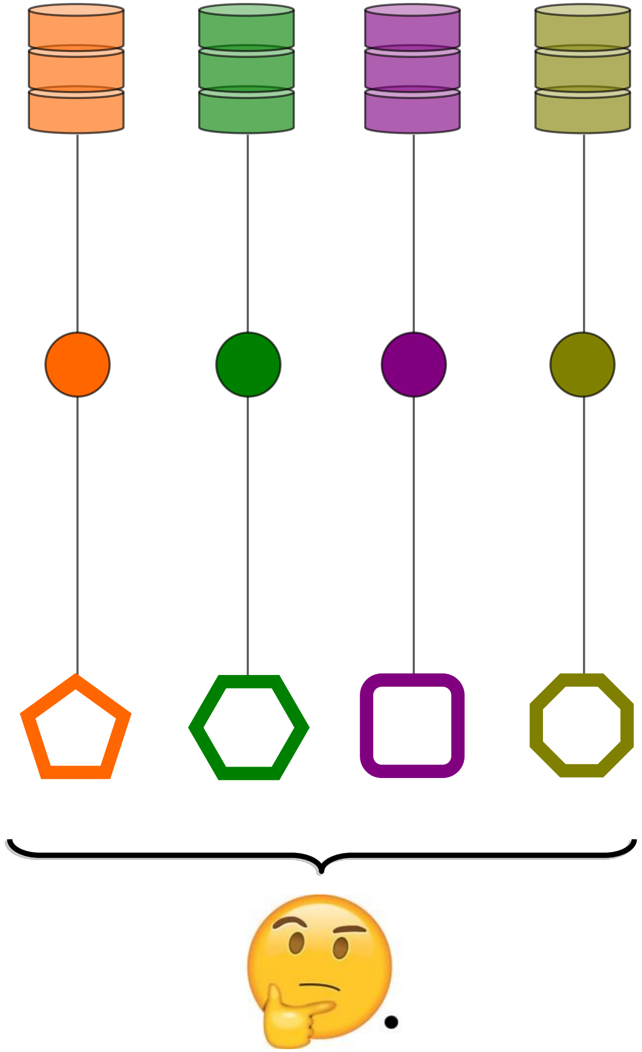
Data integration: state of the art

Ideally



Data integration: state of the art

Reality



Many data resources

- Many to maintain
- New appearing
- Few have a sustained future
- Not easy to find them

Different query interfaces

Variable results

- Formats
- Schemas
- Data content

Data integration

- Redundancy
- Inconsistency

FAIRsharing.org standards, databases, policies

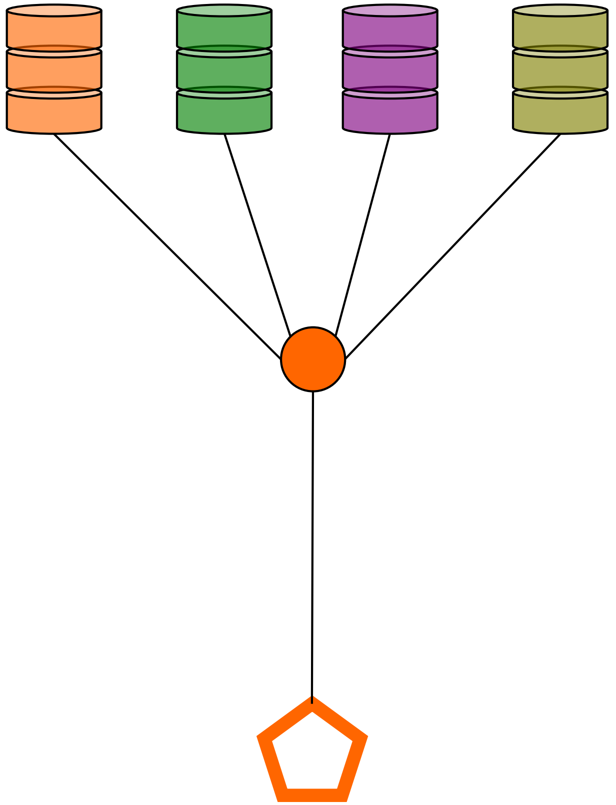
A catalogue of databases, described according to the BioDBcore guidelines, along with the standards used within them; partly compiled with the support of Oxford University Press (NAR Database Issue and DATABASE Journal).

Compiles 2062 databases, 1710 standards & 168 policies*

*20/11/2023

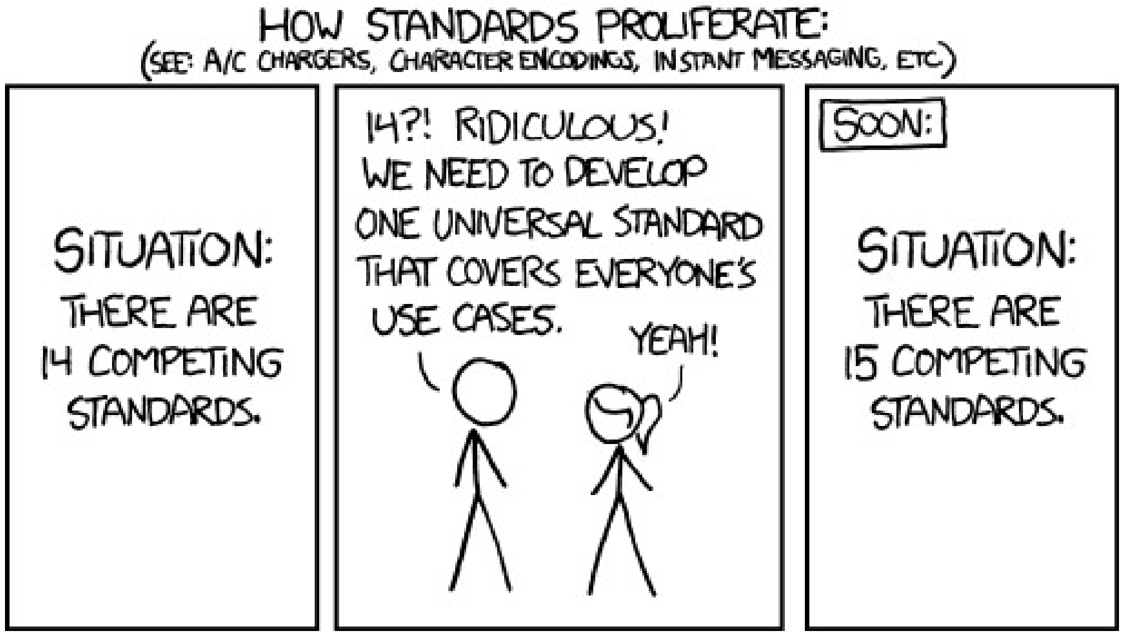
Data integration: state of the art

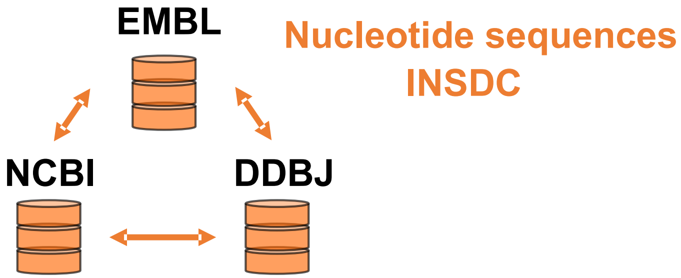
Compromise



Standards & Ontologies

Standards


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.



Nucleotide sequences
INSDC

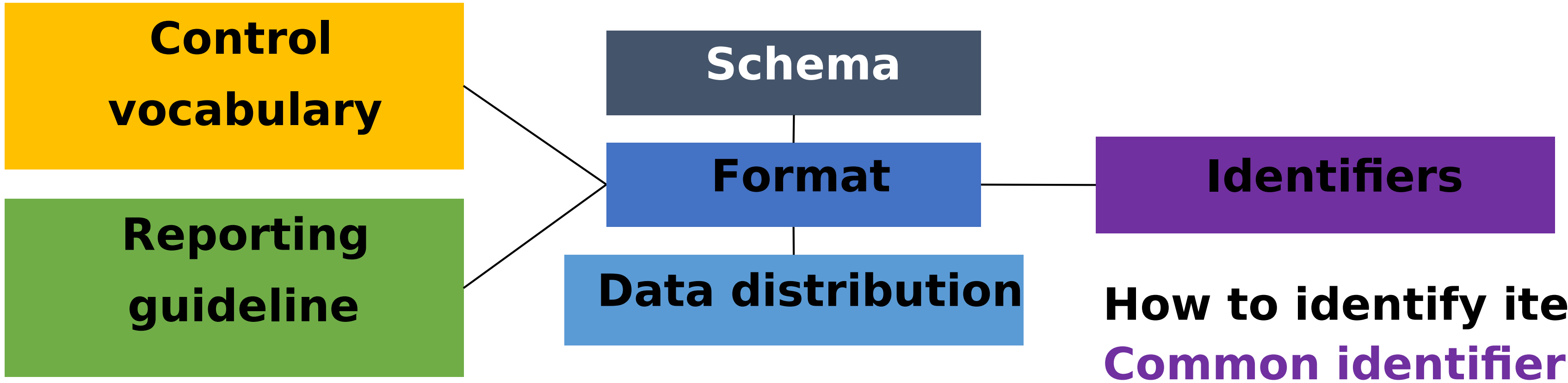
Collaboration among data providers

- **More data coverage**
- **Less redundancy**
- **Less inconsistency**
- **Better data management: access, exchange, sharing, portability, interoperability, annotation, comparison, verification, representation, integration, reusability**

Standards

How to annotate it?
Controlled vocabularies (ontologies)

How/where to store it?
Database, common schemas



Control vocabulary

Reporting guideline

Schema

Format

Data distribution

Identifiers

How to identify items?
Common identifiers

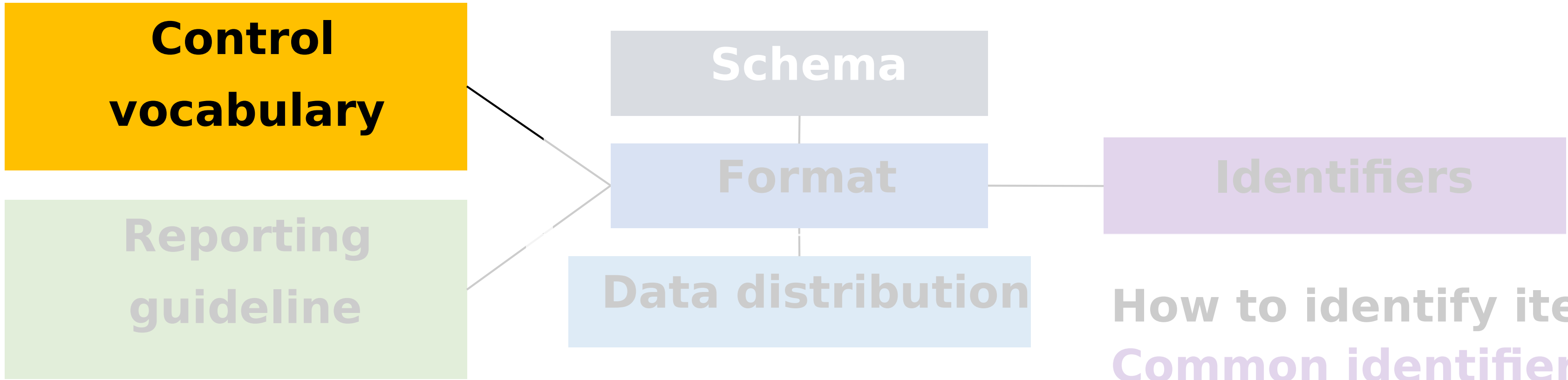
What to store? How to put the data in context?
Minimum information guidelines, Metadata

How to make it available?
Common formats, common query interfaces

Standards

How to annotate it?
Controlled vocabularies (ontologies)

How/where to store it?
Database, common schemas



How to identify items?
Common identifiers

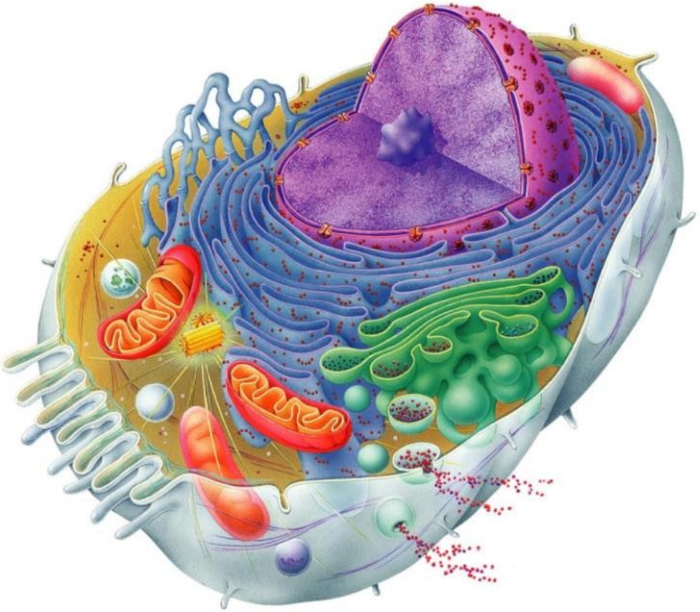
What to store? How to put the data in context?
Minimum information guidelines, Metadata

How to make it available?
Common formats, common query interfaces

Controlled vocabularies (ontologies)

- **Inconsistency in natural language: same name, different concepts**

CELL



Controlled vocabularies (ontologies)

- Controlled vocabularies provide a way to organize knowledge for subsequent retrieval
- Mandate the use of defined preselected terms

- The Gene Ontology (GO)
- A way to capture biological knowledge for individual gene products in a written and computable form
- A set of concepts and their relationships to each other arranged as a hierarchy

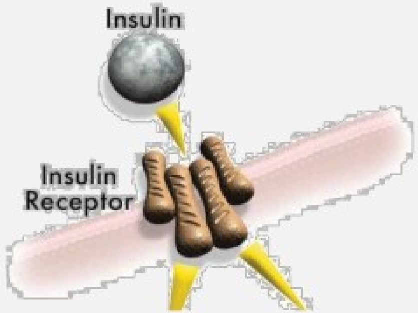
GENEONTOLOGY
Unifying Biology

http://www.geneontology.org/

Controlled vocabularies: The Gene Ontology (GO)

1. Molecular Function

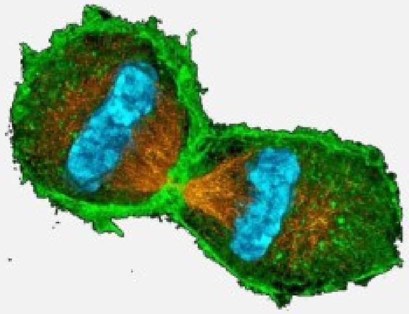
An elemental activity or task or job



- protein kinase activity
- insulin receptor activity

2. Biological Process

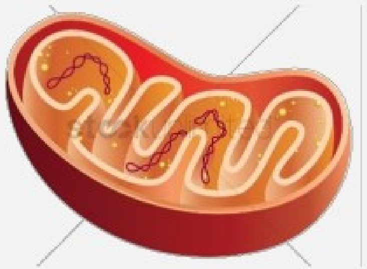
A commonly recognized series of events



- cell division

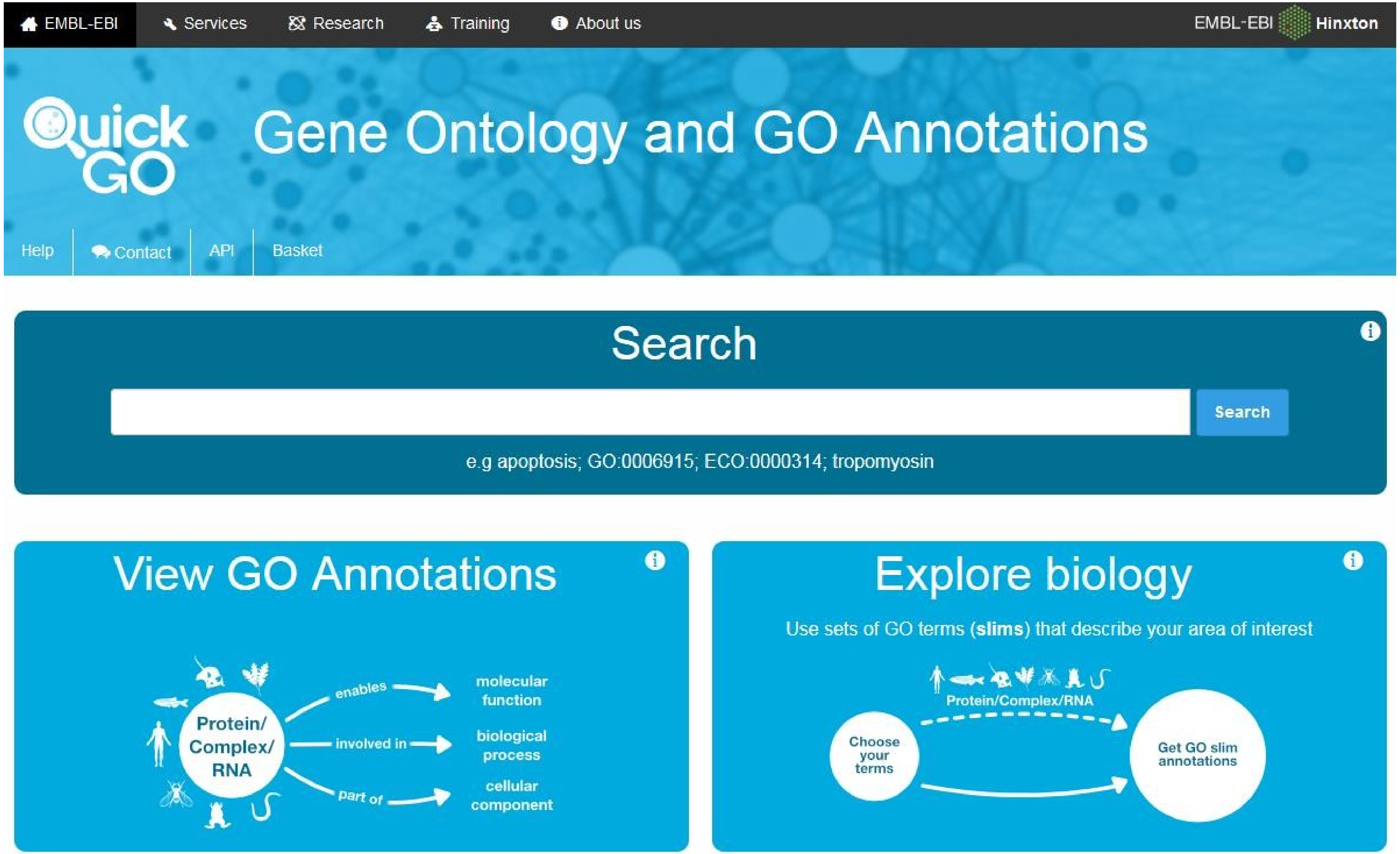
3. Cellular Component

Where a gene product is located



- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

Controlled vocabularies: The Gene Ontology (GO)



Gene Ontology and GO Annotations

Search

e.g apoptosis; GO:0006915; ECO:0000314; tropomyosin

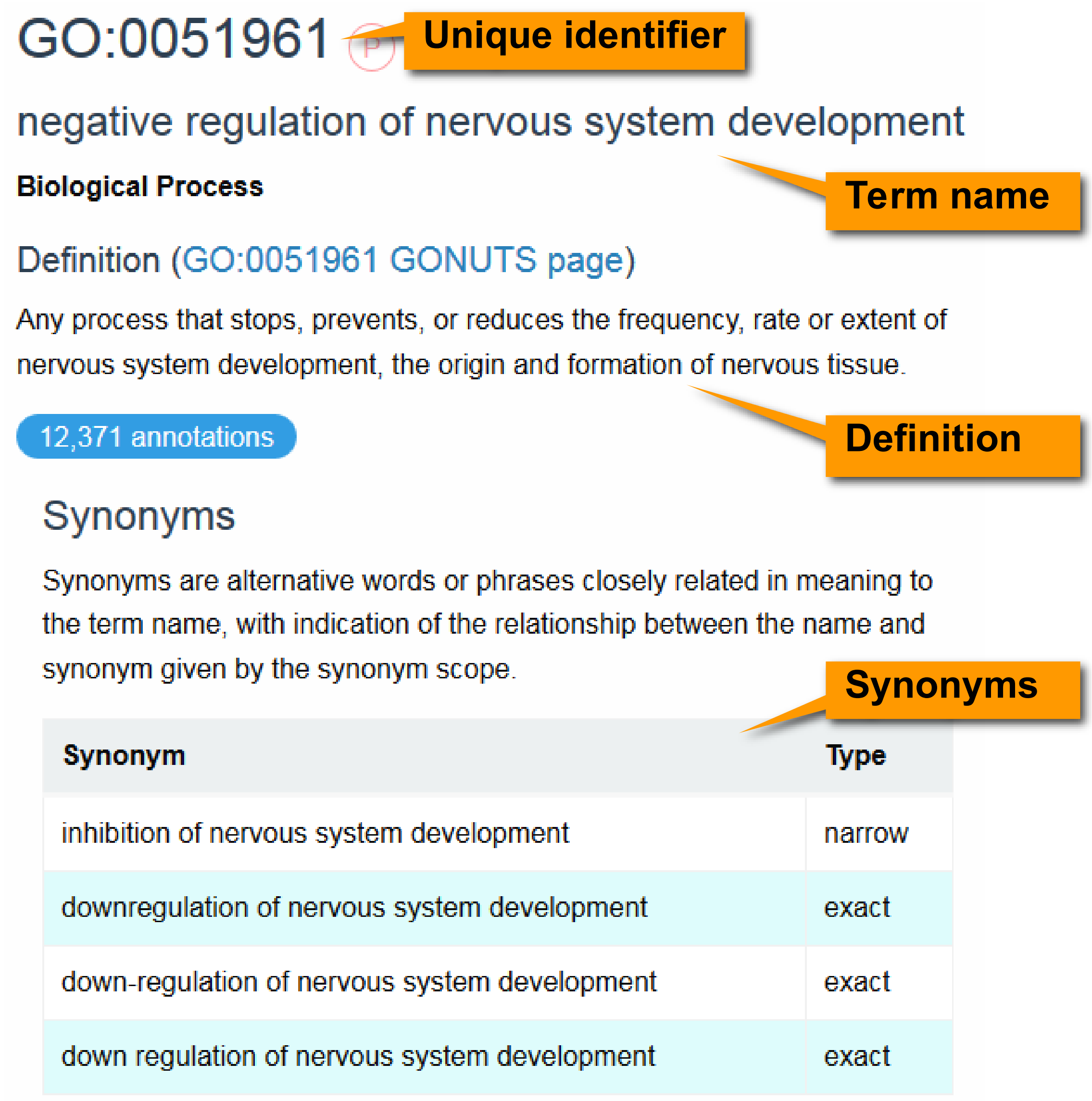
View GO Annotations

Explore biology

Use sets of GO terms (**slims**) that describe your area of interest

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

Controlled vocabularies: The Gene Ontology (GO)



GO:0051961

Unique identifier

negative regulation of nervous system development

Term name

Biological Process

Definition (GO:0051961 GONUTS page)

Any process that stops, prevents, or reduces the frequency, rate or extent of nervous system development, the origin and formation of nervous tissue.

Definition

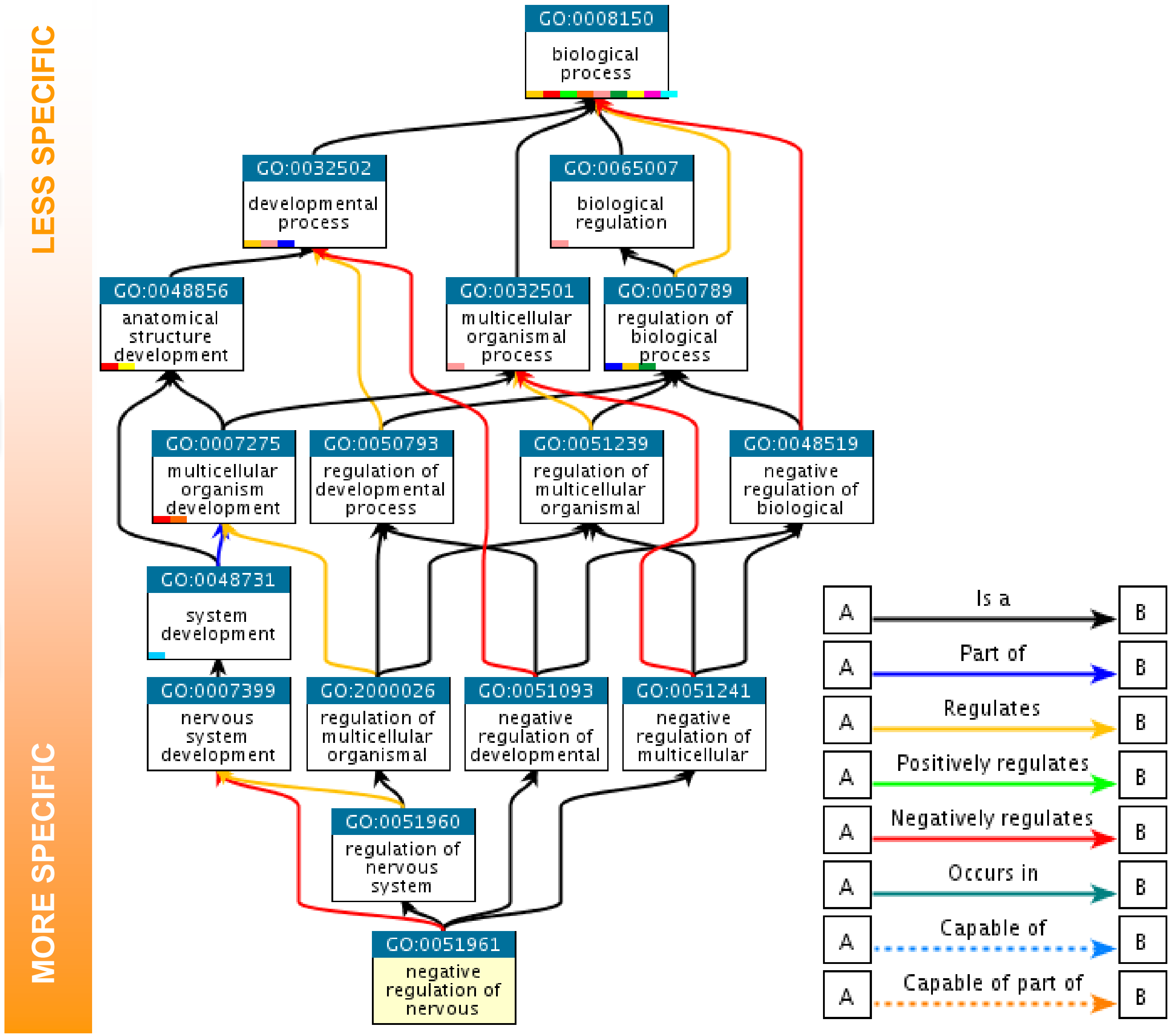
12,371 annotations

Synonyms

Synonyms are alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope.

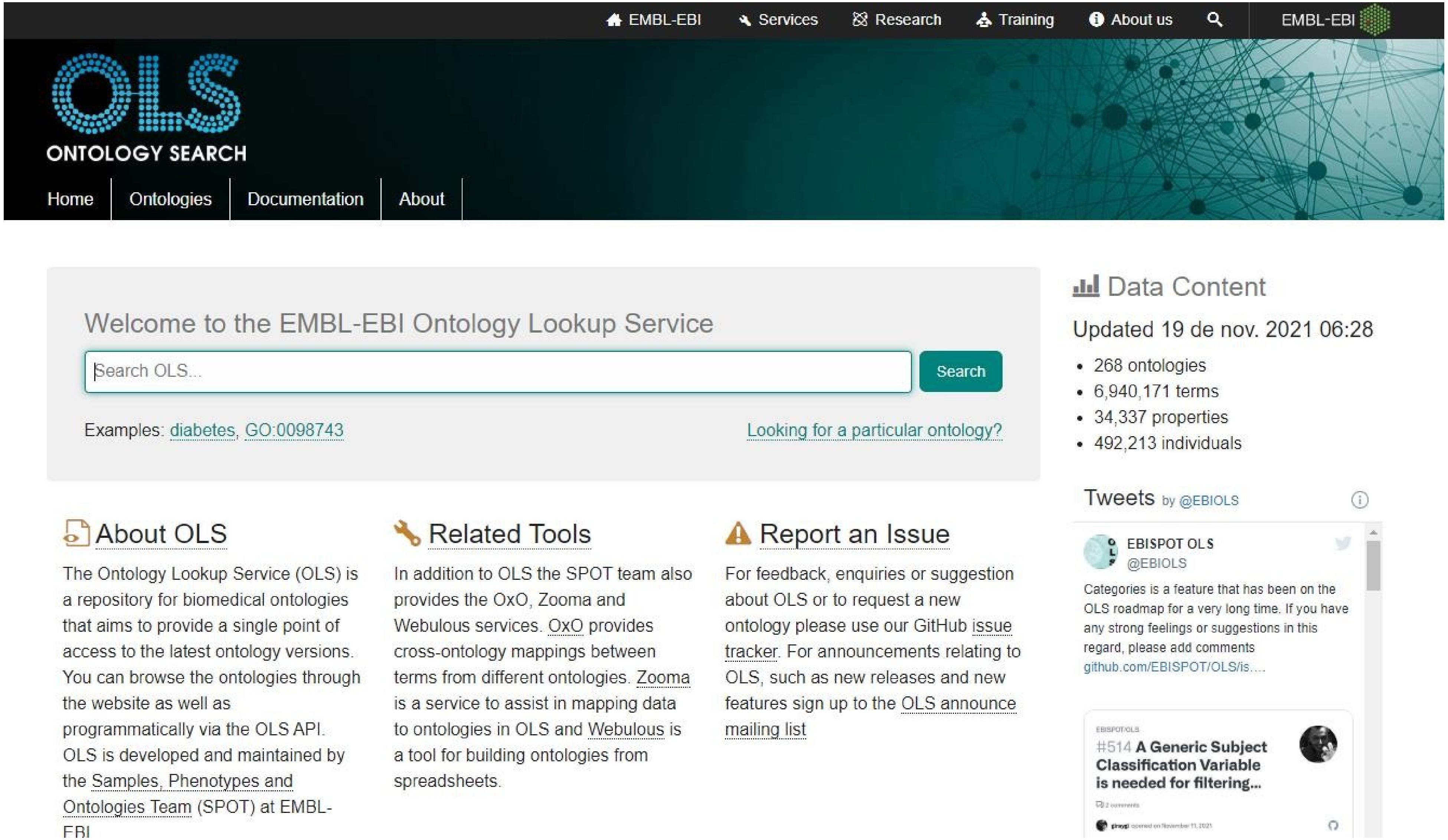
Synonyms

Synonym	Type
inhibition of nervous system development	narrow
downregulation of nervous system development	exact
down-regulation of nervous system development	exact
down regulation of nervous system development	exact


LESS SPECIFIC
MORE SPECIFIC
GO:0008150 biological process
GO:0032502 developmental process
GO:0065007 biological regulation
GO:0048856 anatomical structure development
GO:0032501 multicellular organismal process
GO:0050789 regulation of biological process
GO:0007275 multicellular organism development
GO:0050793 regulation of developmental process
GO:0051239 regulation of multicellular organismal
GO:0048519 negative regulation of biological
GO:0048731 system development
GO:0007399 nervous system development
GO:2000026 regulation of multicellular organismal
GO:0051093 negative regulation of developmental
GO:0051241 negative regulation of multicellular
GO:0051960 regulation of nervous system
GO:0051961 negative regulation of nervous
A Is a B
A Part of B
A Regulates B
A Positively regulates B
A Negatively regulates B
A Occurs in B
A Capable of B
A Capable of part of B

https://www.ebi.ac.uk/QuickGO/term/GO:0051961

Controlled vocabularies: The Ontology Lookup Servi



Welcome to the EMBL-EBI Ontology Lookup Service

Search OLS... [Search]

Examples: diabetes, GO:0098743

Looking for a particular ontology?

Data Content

Updated 19 de nov. 2021 06:28

- 268 ontologies
- 6,940,171 terms
- 34,337 properties
- 492,213 individuals

Tweets by @EBIOLS

EBISPOT OLS
@EBIOLS

Categories is a feature that has been on the OLS roadmap for a very long time. If you have any strong feelings or suggestions in this regard, please add comments github.com/EBISPOT/OLS/is....

EBISPOT/OLS

#514 **A Generic Subject Classification Variable is needed for filtering...**

2 comments

opened on November 11, 2021

About OLS

The Ontology Lookup Service (OLS) is a repository for biomedical ontologies that aims to provide a single point of access to the latest ontology versions. You can browse the ontologies through the website as well as programmatically via the OLS API. OLS is developed and maintained by the Samples, Phenotypes and Ontologies Team (SPOT) at EMBL-EBI

Related Tools

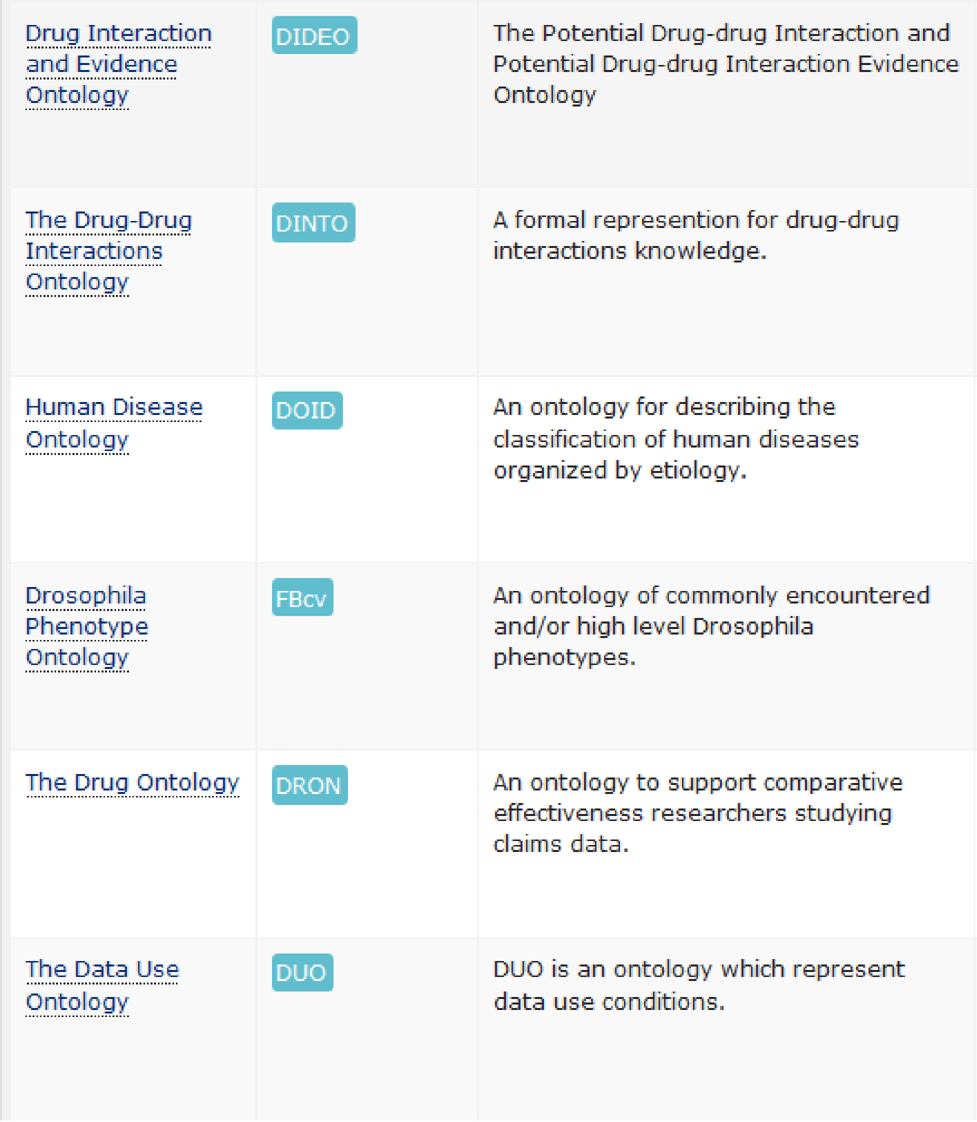
In addition to OLS the SPOT team also provides the OxO, Zooma and Webulous services. OxO provides cross-ontology mappings between terms from different ontologies. Zooma is a service to assist in mapping data to ontologies in OLS and Webulous is a tool for building ontologies from spreadsheets.

Report an Issue

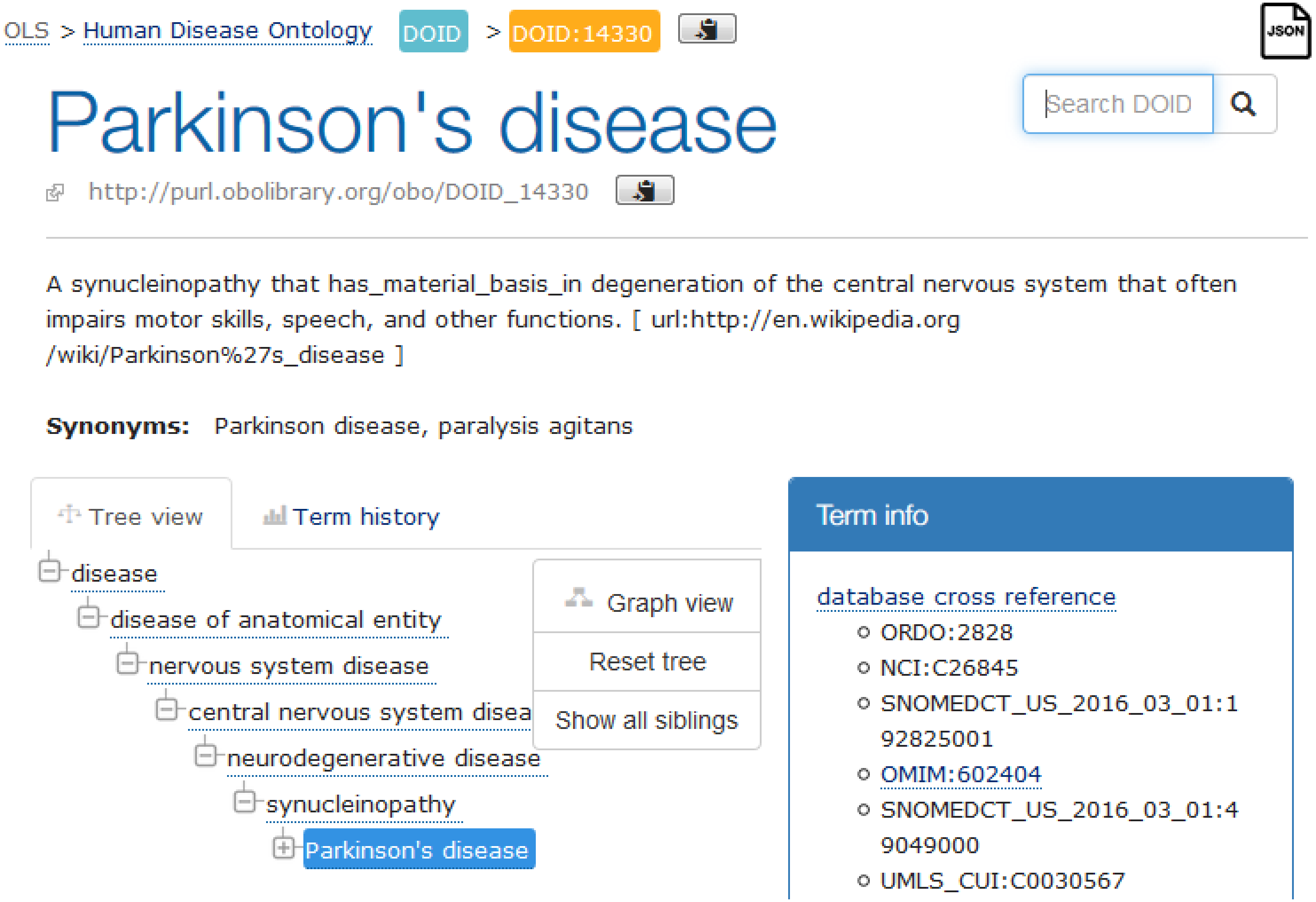
For feedback, enquiries or suggestion about OLS or to request a new ontology please use our GitHub issue tracker. For announcements relating to OLS, such as new releases and new features sign up to the OLS announce mailing list

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

Controlled vocabularies: The Ontology Lookup Service (OLS)



Drug Interaction and Evidence Ontology	DIDEO	The Potential Drug-drug Interaction and Potential Drug-drug Interaction Evidence Ontology
The Drug-Drug Interactions Ontology	DINTO	A formal represention for drug-drug interactions knowledge.
Human Disease Ontology	DOID	An ontology for describing the classification of human diseases organized by etiology.
Drosophila Phenotype Ontology	FBcv	An ontology of commonly encountered and/or high level Drosophila phenotypes.
The Drug Ontology	DRON	An ontology to support comparative effectiveness researchers studying claims data.
The Data Use Ontology	DUO	DUO is an ontology which represent data use conditions.



OLS > Human Disease Ontology DOID > DOID:14330

Parkinson's disease

http://purl.obolibrary.org/obo/DOID_14330

A synucleinopathy that has_material_basis_in degeneration of the central nervous system that often impairs motor skills, speech, and other functions. [url:http://en.wikipedia.org /wiki/Parkinson%27s_disease]

Synonyms: Parkinson disease, paralysis agitans

```
disease
  disease of anatomical entity
    nervous system disease
      central nervous system disea
        neurodegenerative disease
          synucleinopathy
            Parkinson's disease
```

Term info

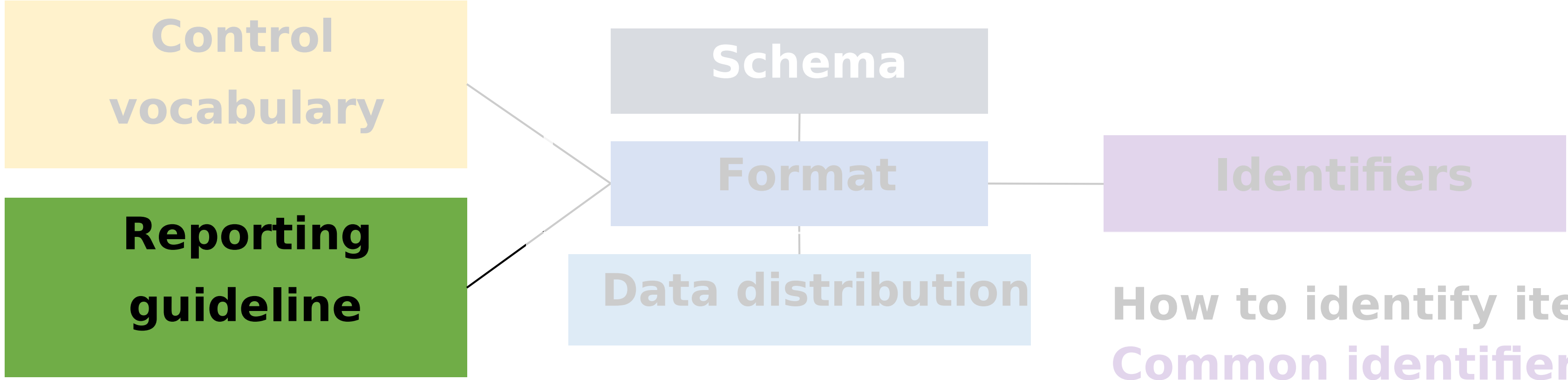
database cross reference

- ORDO:2828
- NCI:C26845
- SNOMEDCT_US_2016_03_01:192825001
- OMIM:602404
- SNOMEDCT_US_2016_03_01:49049000
- UMLS_CUI:C0030567

Standards

How to annotate it?
Controlled vocabularies (ontologies)

How/where to store it?
Database, common schemas



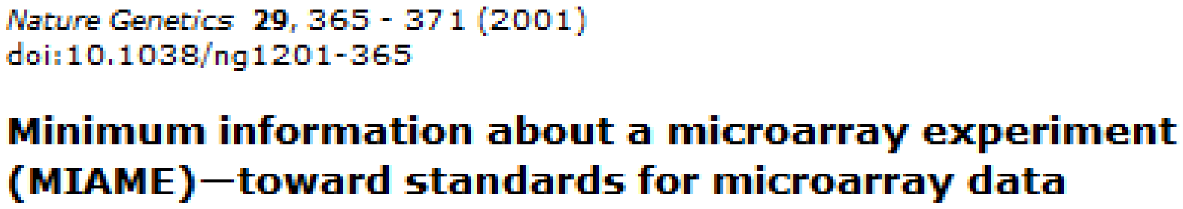
How to identify items?
Common identifiers

What to store? How to put the data in context?
Minimum information guidelines, Metadata

How to make it available?
Common formats, common query interfaces

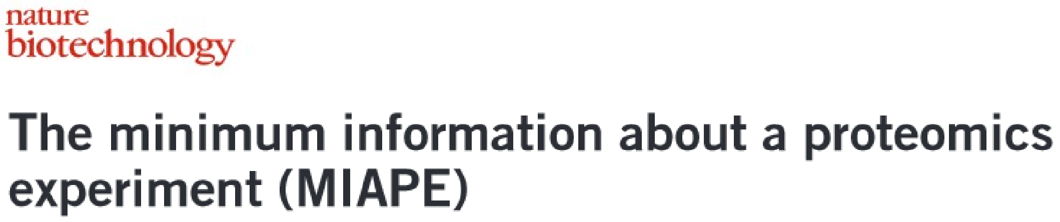
Minimum information guidelines

- The reader can understand, and potentially reproduce, the authors experiments
- The data can be moved into a database by a manual curator
- The data can be captured, in some form, by text-mining



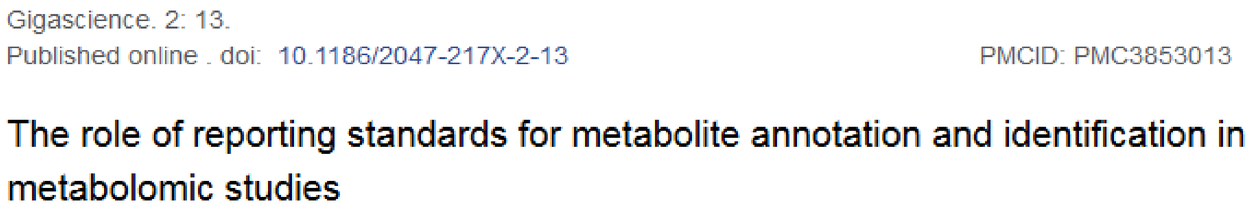
Nature Genetics **29**, 365 - 371 (2001)
doi:10.1038/ng1201-365

Minimum information about a microarray experiment (MIAME)—toward standards for microarray data



nature biotechnology

The minimum information about a proteomics experiment (MIAPE)



Gigascience. 2: 13.
Published online . doi: 10.1186/2047-217X-2-13 PMCID: PMC3853013

The role of reporting standards for metabolite annotation and identification in metabolomic studies

Metadata

- **Metadata is data that describes other data**
- **In the case of experimental data, this is a description of experimental conditions; e.g., patient/environmental details, cell lines, instrumentation**

- Common problems: incomplete and inconsistent metadata

37 year old male
600 yr. old male
adult male
bull
castrated male
cm
dioecious male
diploid male
drone
engorged male
fertile male
four males mixed
individual male

initial phase male
m
make
makle
mal e
male
male (7-2872)
male (7-3074)
male (m-a)
male (m-o)
male caucasian
male child
male fertile

male fetus
male plant
male, 8 weeks old
male, castrated
male, pooled
males
man
men
normale male
ram
rooster
s1 male sterile
sex: male

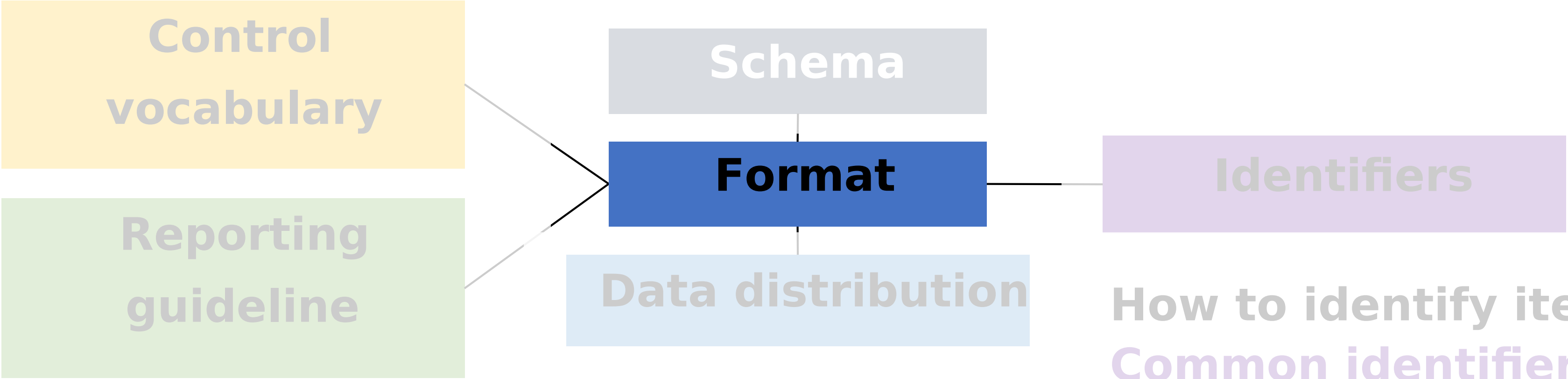
six males mixed
stallion
steer
sterile male
strictly male
tetraploide male
type i males
type ii males
virgin male
winged and wingless males
young male

female*

Standards

How to annotate it?
Controlled vocabularies (ontologies)

How/where to store it?
Database, common schemas



How to identify items?
Common identifiers

What to store? How to put the data in context?
Minimum information guidelines, Metadata

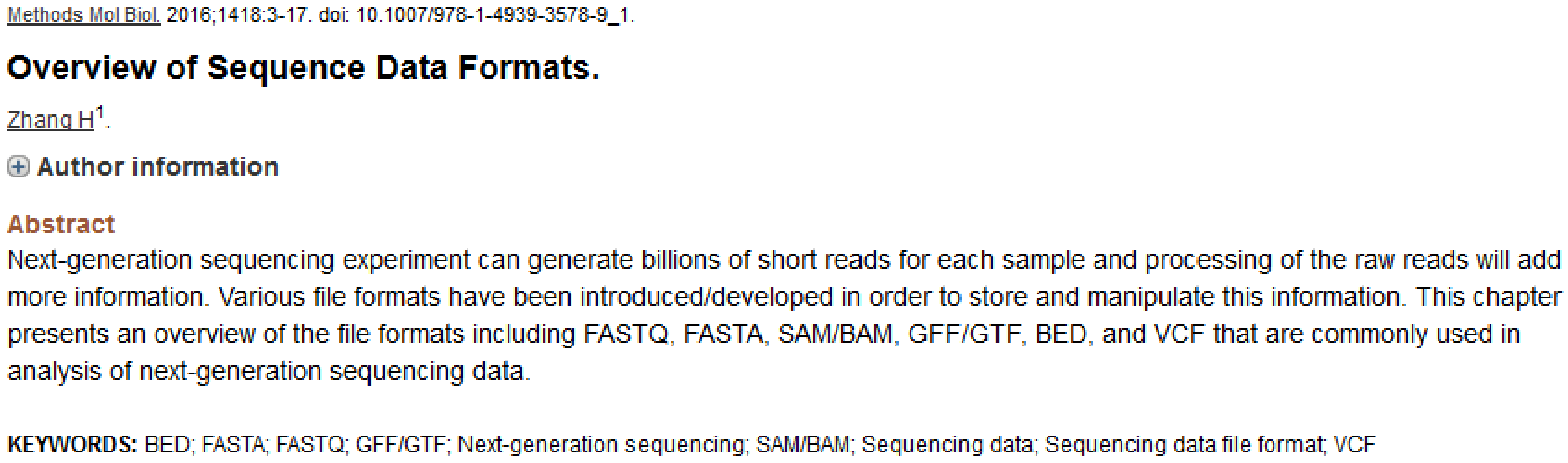
How to make it available?
Common formats, common query interfaces

Data formats

- **Merge data from different data sources with minimal effort**
- **Parse the data**
- **Build tools to use the data**

- **PSI-MI XML format (molecular interactions)**

Proteomics Standards Initiative



Methods Mol Biol. 2016;1418:3-17. doi: 10.1007/978-1-4939-3578-9_1.

Overview of Sequence Data Formats.

Zhang H[1].

⊕ Author information

Abstract

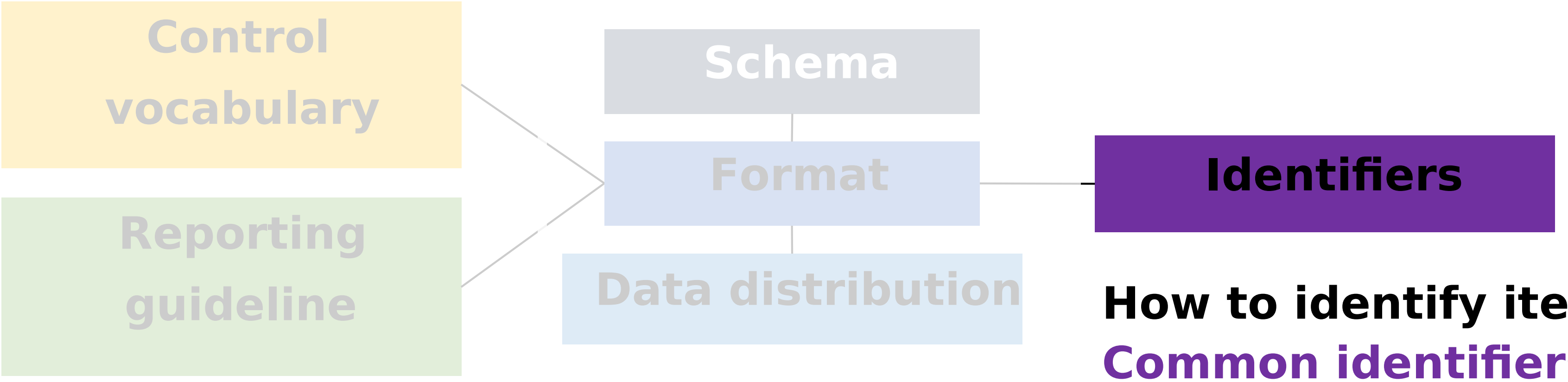
Next-generation sequencing experiment can generate billions of short reads for each sample and processing of the raw reads will add more information. Various file formats have been introduced/developed in order to store and manipulate this information. This chapter presents an overview of the file formats including FASTQ, FASTA, SAM/BAM, GFF/GTF, BED, and VCF that are commonly used in analysis of next-generation sequencing data.

KEYWORDS: BED; FASTA; FASTQ; GFF/GTF; Next-generation sequencing; SAM/BAM; Sequencing data; Sequencing data file format; VCF

Standards

How to annotate it?
Controlled vocabularies (ontologies)

How/where to store it?
Database, common schemas



How to identify items?
Common identifiers

What to store? How to put the data in context?
Minimum information guidelines, Metadata

How to make it available?
Common formats, common query interfaces

Identifiers & ID mapping

- Identifiers: The use of identifiers allows for unambiguous identifications of molecules or conceptual entities, and their representation in databases
- ID mapping: There is a large number of identifiers that aim to represent the same entity

ENSP00000269305 **P04637** **NP_001119584.1**

P53_HUMAN

HGNC:11998

```
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

NX_P04637

Antigen NY-CO-13 **Cellular tumor antigen p53** **TP53**

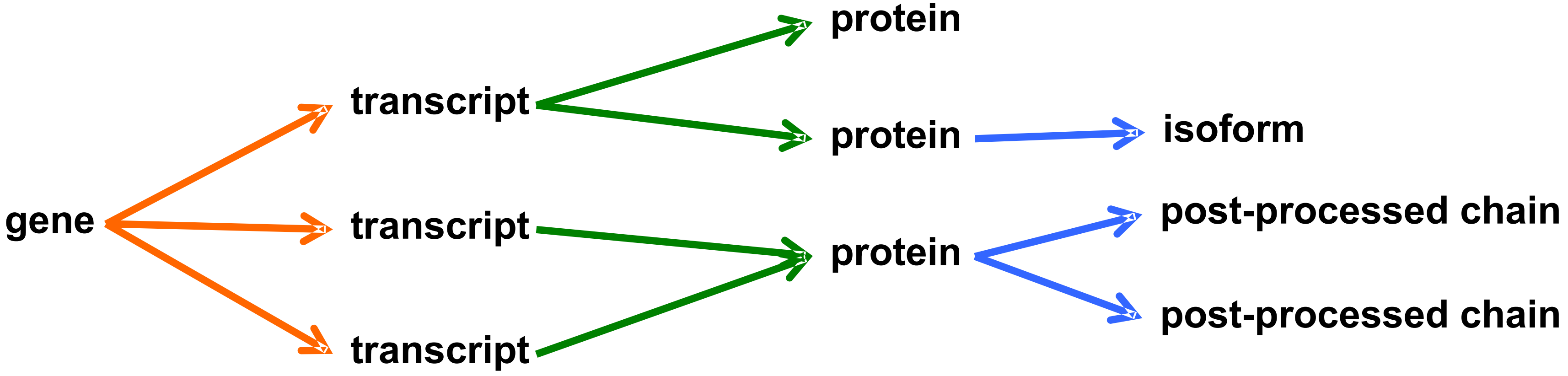
P04637-1 **uc002gii.2** **CHEMBL2221344**

Identifiers & ID mapping

- Most commonly used accessions:
  - **Entrez GeneIDs:** Gene-centered identifier: DNA consensus sequence, no isoform or variants.
  - **UniProt:** Represents proteins, taking into account isoforms. Additional identifiers for variants and post-processed chains.
  - **RefSeq:** Represents sequences of DNA, RNA and proteins.
  - **Ensembl:** Identifiers that represent genes and their different products: gene, gene tree, protein, regulatory feature, transcript, exon and protein family.
  - **International Protein Index:** Proteomics reference database (protein sequences). Now obsoleted, but still used in proteomics.
  - **HUGO gene symbols:** Unique symbols and names for human loci (protein-coding genes, RNA genes and pseudogenes).
  - **Organism centered databases:** FlyBase, TAIR, WormBase, SGD, …

Identifiers & ID mapping

- Identifiers most of the times do not have a 1:1 relationship: **gene ≠ transcript ≠ protein**



- Identifiers (and sequences!) **disappear** and **get updated**
- Identifiers are **misused**; *e.g.*, gene identifiers used to represent proteins

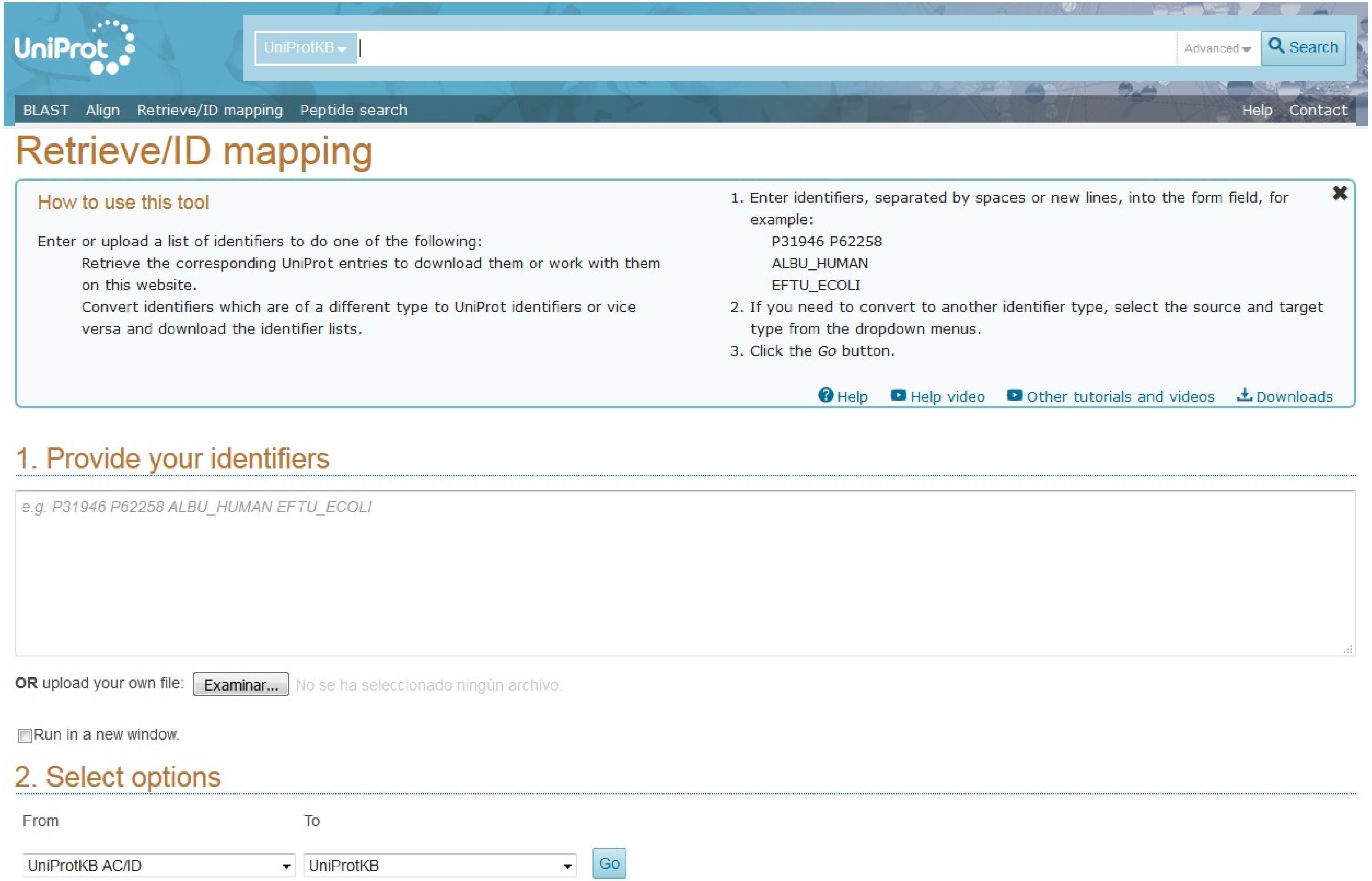
Identifiers & ID mapping

- ID mapping tools / mapping tables

UniProt ID mapping	https://www.uniprot.org/uploadlists/
Ensembl BioMart	https://www.ensembl.org/biomart/martview
DAVID GeneID Conversion Tool	https://david.ncifcrf.gov/conversion.jsp
HGNC project	http://www.genenames.org/

- Common problems:
  - You are not able to map all the IDs
  - You get 1:n mappings

Identifiers & ID mapping



Retrieve/ID mapping

How to use this tool

Enter or upload a list of identifiers to do one of the following:

Retrieve the corresponding UniProt entries to download them or work with them on this website.

Convert identifiers which are of a different type to UniProt identifiers or vice versa and download the identifier lists.

1. Enter identifiers, separated by spaces or new lines, into the form field, for example:
   P31946 P62258
   ALBU_HUMAN
   EFTU_ECOLI
2. If you need to convert to another identifier type, select the source and target type from the dropdown menus.
3. Click the *Go* button.

1. Provide your identifiers

e.g. P31946 P62258 ALBU_HUMAN EFTU_ECOLI

OR upload your own file: Examinar... No se ha seleccionado ningún archivo.

□ Run in a new window.

2. Select options

From: UniProtKB AC/ID

To: UniProtKB

http://www.uniprot.org/uploadlists/

Identifiers & ID mapping

Frameworks for finding and mapping equivalent database identifiers for genes, proteins, and metabolites

- **BridgeDb**
  - Cytoscape app: http://apps.cytoscape.org/apps/bridgedb
  - bridgeDbR Bioconductor package: http://bioconductor.org/packages/devel/bioc/html/BridgeDbR.html
  - Web service API: http://webservice.bridgedb.org/

- **BioMart**
  - Cytoscape app: http://apps.cytoscape.org/apps/biomartwebserviceclient
  - biomaRt Bioconductor package: https://bioconductor.org/packages/release/bioc/html/biomaRt.html
  - Web service API: http://www.biomart.org/martservice.html

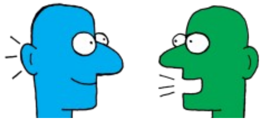
Points to remember

- **In order to merge/integrate data it needs to have as many similarities as possible – the same entity/concept described in different ways will not merge**
- **Data transfer/conversion almost always means data loss – if you want detail, go to the source**
- **More resources does not necessarily mean more data – many databases import from the few that curate**
- **If using a resource which imports data – check the date of the last import**
- **Web-services run directly from the most recent release – may be a better source of data than a data-warehouse, which imports infrequently**
- **One advantage of database submission – data will then be available in standard formats**

Some final recommendations

- Curation, curation, curation
  - Contribute to databases that accept submissions
  - Use credible data sources

- Collaboration, constant feedback, agree on standards
  - Use standard data formats
  - Use ontology terms to annotate metadata



- Openness, reproducibility
  - Use open-source software
  - Manage and share code
  - Transparent data sources and analysis procedures
  - Document as much as you can