Mappings @ EBI

James McLaughlin, PhD

Ontology Project Lead

EMBL-EBI

EMBL-EBI

- We host and manage biomedical DBs including Ensembl, UniProt, GWAS Catalog

- Our data is annotated using ontology terms
  - Our application ontology: the **Experimental Factor Ontology (EFO)** imports terms from domain ontologies such as the Monarch disease ontology (MONDO) and the Gene Ontology (GO)

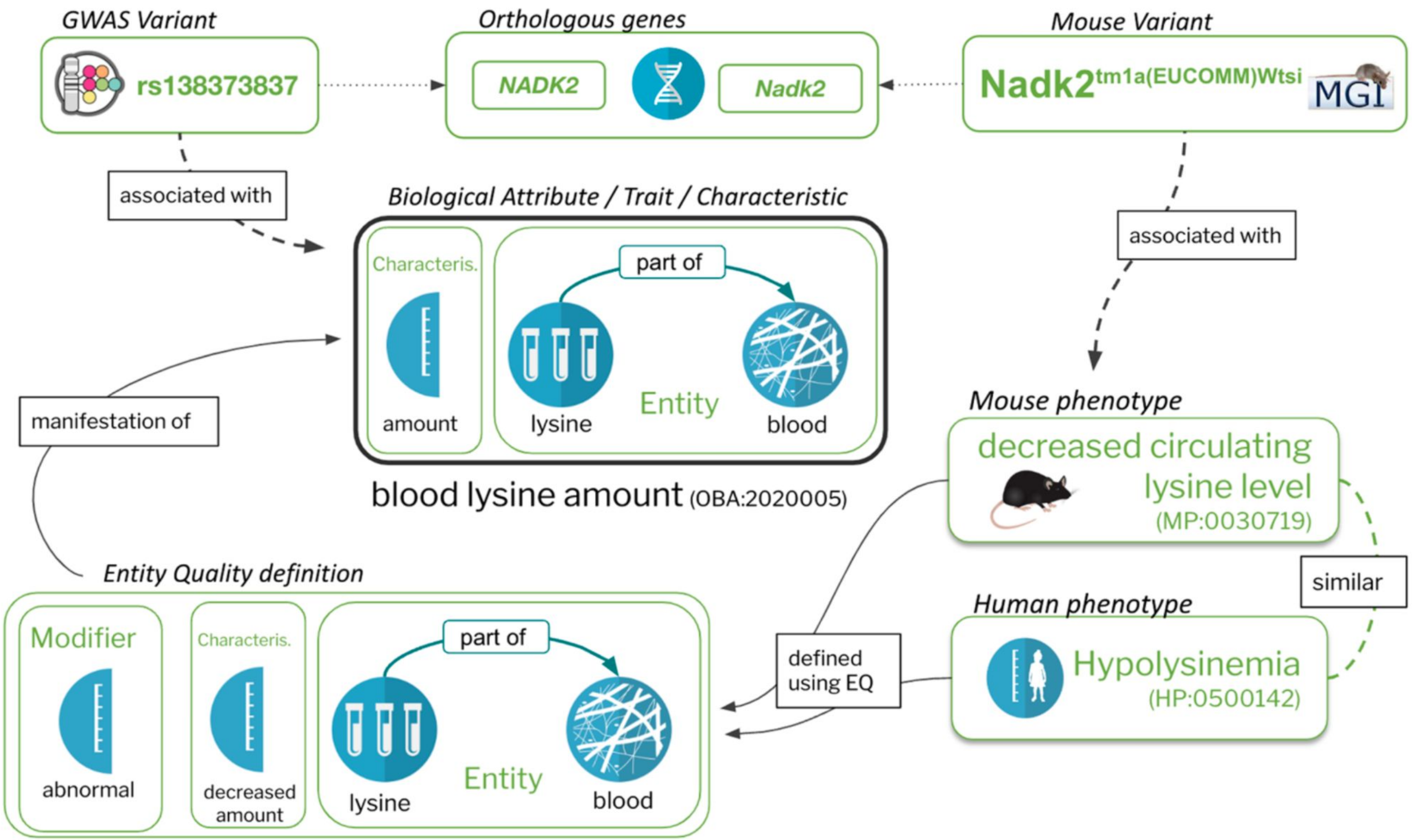
- We also provide web services for working with ontologies:
  - The **Ontology Lookup Service (OLS)**
  - The **Ontology Xref Service (OxO)**
  - **ZOOMA**
  - **+ OntoString** (beta)

Mapping use cases @ EBI

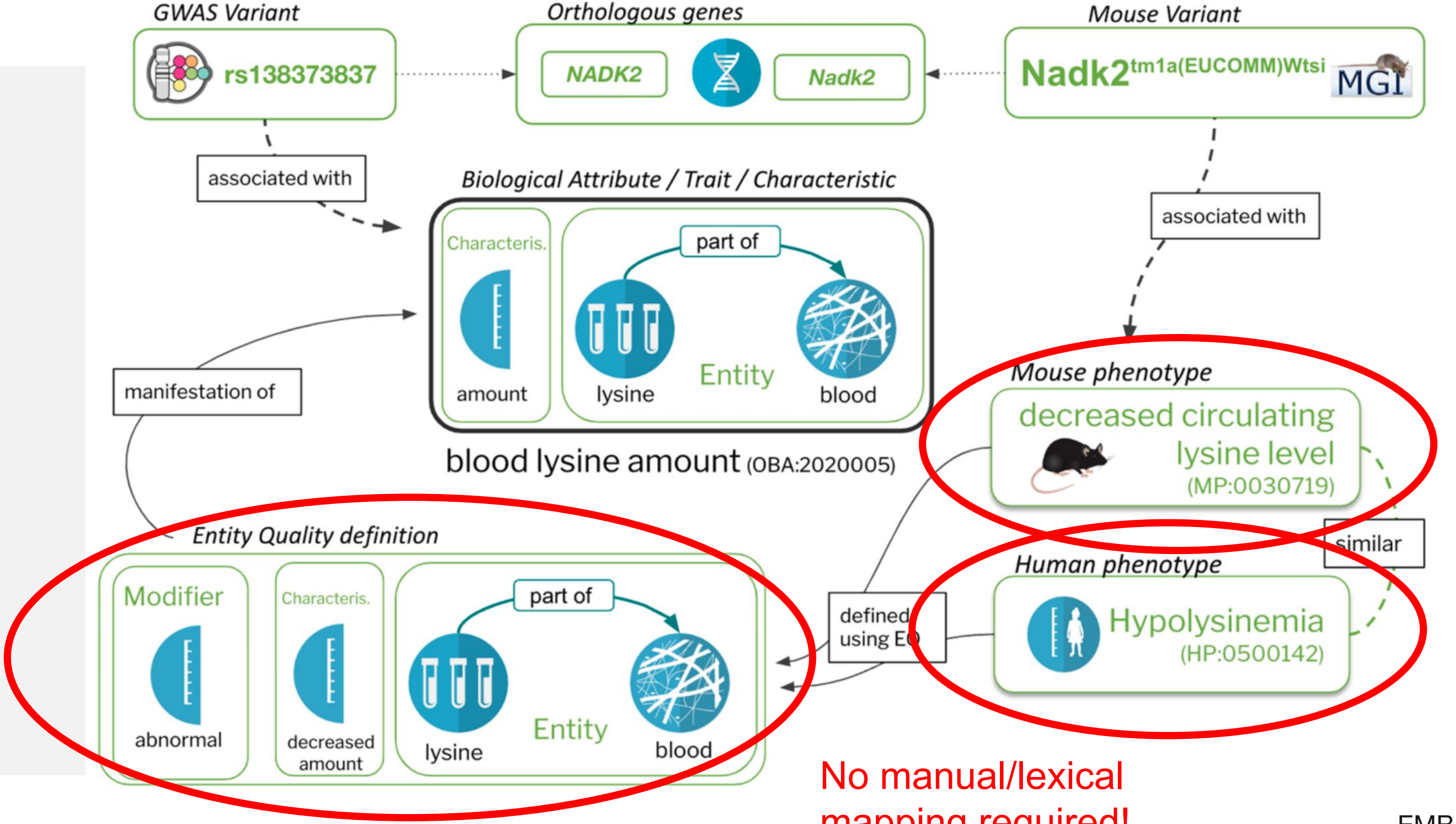
- There is LOTS of direct overlap in the ontology ecosystem, e.g. DOID and MONDO are both popular disease ontologies
  - Literally different terms which represent **exactly the same thing**, and to make datasets comparable (e.g. align a patient study with known info about a disease) we need to align them.

- Also less direct overlap e.g. disease-phenotype
  - A phenotype is an observable trait, and a disease is *generally* considered to be a collection of phenotypes
  - Sometimes there is a direct disease->phenotype mapping, sometimes a lot more nuanced.

- Mappings can be scientifically interesting!
  - E.g. mapping between mouse phenotypes in **MP** and human phenotypes in **HP** can make a mouse and human study comparable


GWAS Variant
rs138373837
Orthologous genes
NADK2
Nadk2
Mouse Variant
Nadk2tm1a(EUCOMM)Wtsi
MGI
associated with
Biological Attribute / Trait / Characteristic
Characteris.
amount
part of
lysine
Entity
blood
associated with
manifestation of
blood lysine amount (OBA:2020005)
Mouse phenotype
decreased circulating lysine level
(MP:0030719)
similar
Entity Quality definition
Human phenotype
Modifier
abnormal
Characteris.
decreased amount
part of
lysine
Entity
blood
defined using EQ
Hypolysinemia
(HP:0500142)

Mapping using semantic definition equivalence

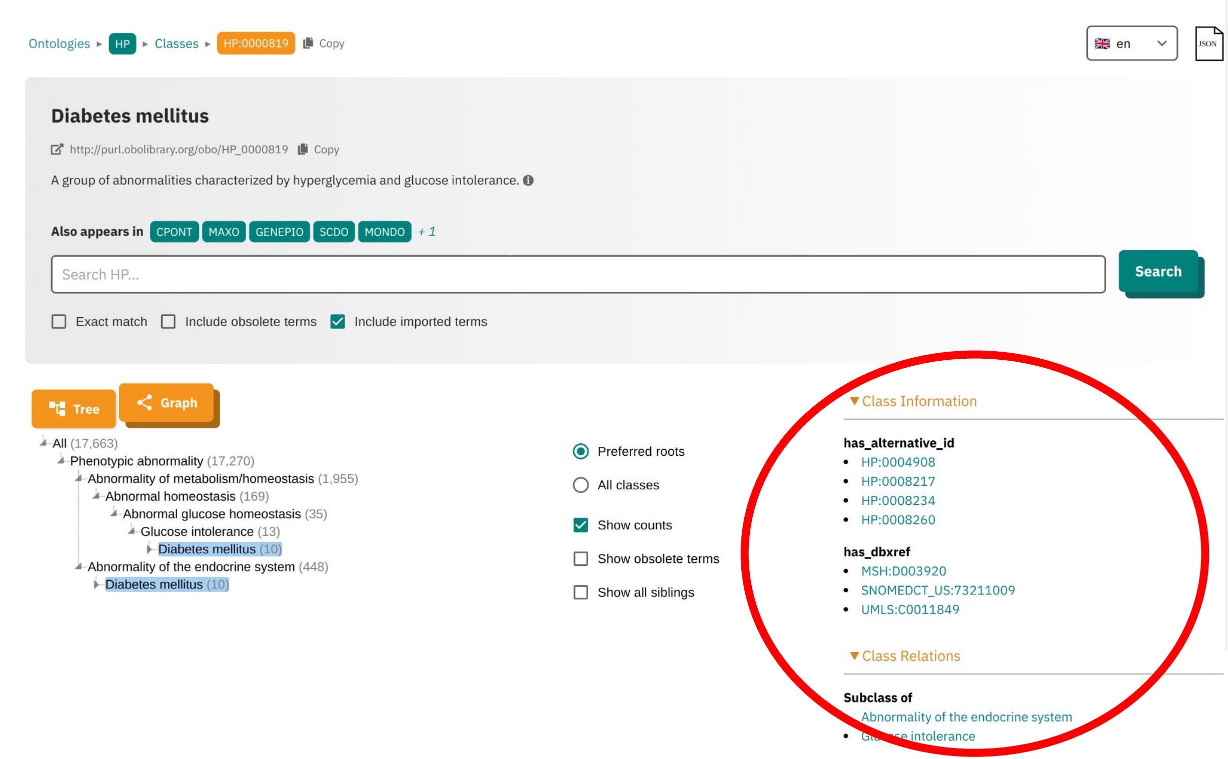


No manual/lexical mapping required!

EMBL-EBI

Explicitly defined mappings

- Ontologies often define mappings **(especially using the “oboInOwl:hasDbXref” predicate)**



Diabetes mellitus

http://purl.obolibrary.org/obo/HP_0000819 Copy

A group of abnormalities characterized by hyperglycemia and glucose intolerance.

Also appears in CPONT MAXO GENEPIO SCDO MONDO + 1

Search HP...

□ Exact match □ Include obsolete terms ☑ Include imported terms

- All (17,663)
  - Phenotypic abnormality (17,270)
    - Abnormality of metabolism/homeostasis (1,955)
      - Abnormal homeostasis (169)
        - Abnormal glucose homeostasis (35)
          - Glucose intolerance (13)
            - Diabetes mellitus (10)
    - Abnormality of the endocrine system (448)
      - Diabetes mellitus (10)

● Preferred roots

○ All classes

☑ Show counts

□ Show obsolete terms

□ Show all siblings

▼Class Information

has_alternative_id

- HP:0004908
- HP:0008217
- HP:0008234
- HP:0008260

has_dbxref

- MSH:D003920
- SNOMEDCT_US:73211009
- UMLS:C0011849

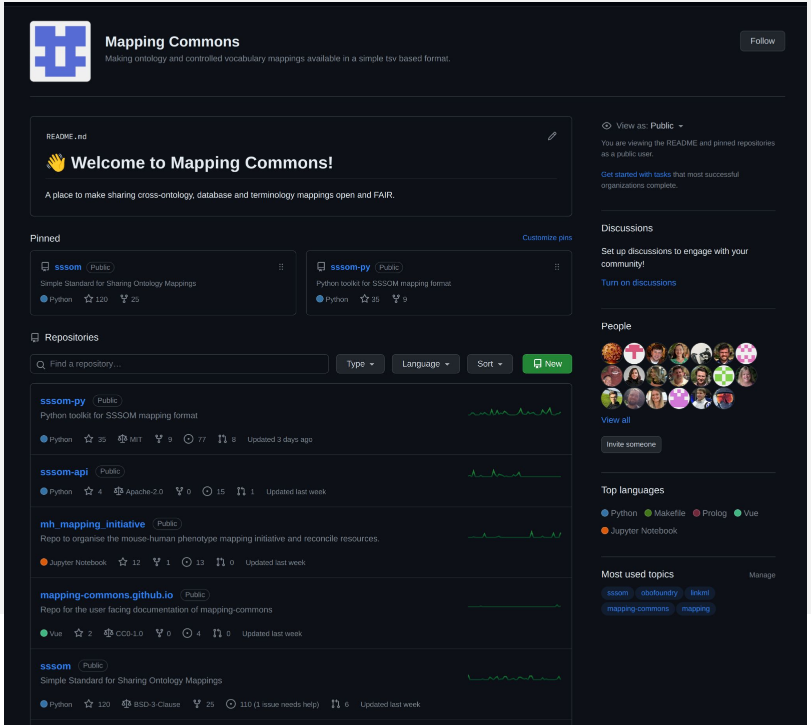
▼Class Relations

Subclass of

- Abnormality of the endocrine system
- Glucose intolerance

Explicitly defined mappings

- We are gradually trying to externalise mappings rather than store them in ontologies
  - Because you might want to make mappings between vocabularies you did not develop
  - And mappings can be subjective: the same vocabularies can be mapped in different ways



Mapping Commons

Making ontology and controlled vocabulary mappings available in a simple tsv based format.

README.md

👋 Welcome to Mapping Commons!

A place to make sharing cross-ontology, database and terminology mappings open and FAIR.

Pinned

sssom Public
Simple Standard for Sharing Ontology Mappings
Python 120 25

sssom-py Public
Python toolkit for SSSOM mapping format
Python 35 9

Repositories

Find a repository…

sssom-py Public
Python toolkit for SSSOM mapping format
Python 35 MIT 9 77 8 Updated 3 days ago

sssom-api Public
Python 4 Apache-2.0 0 15 1 Updated last week

mh_mapping_initiative Public
Repo to organise the mouse-human phenotype mapping initiative and reconcile resources.
Jupyter Notebook 12 1 13 0 Updated last week

mapping-commons.github.io Public
Repo for the user facing documentation of mapping-commons
Vue 2 CC0-1.0 0 4 0 Updated last week

sssom Public
Simple Standard for Sharing Ontology Mappings
Python 120 BSD-3-Clause 25 110 (1 issue needs help) 6 Updated last week

View as: Public

You are viewing the README and pinned repositories as a public user.

Get started with tasks that most successful organizations complete.

Discussions

Set up discussions to engage with your community!

Turn on discussions

People

Top languages

Python Makefile Prolog Vue Jupyter Notebook

Most used topics

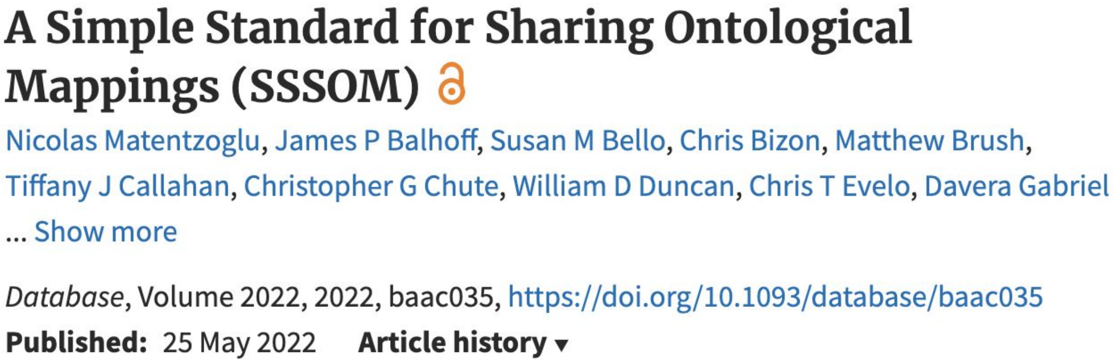
sssom obofoundry linkml mapping-commons mapping

SSSOM: a simple standard for storing mappings in TSV files

Mapping-commons: an organisation on github for people who love mappings (come and join us!)

SSSOM

SIMPLE STANDARD FOR SHARING ONTOLOGY MAPPINGS



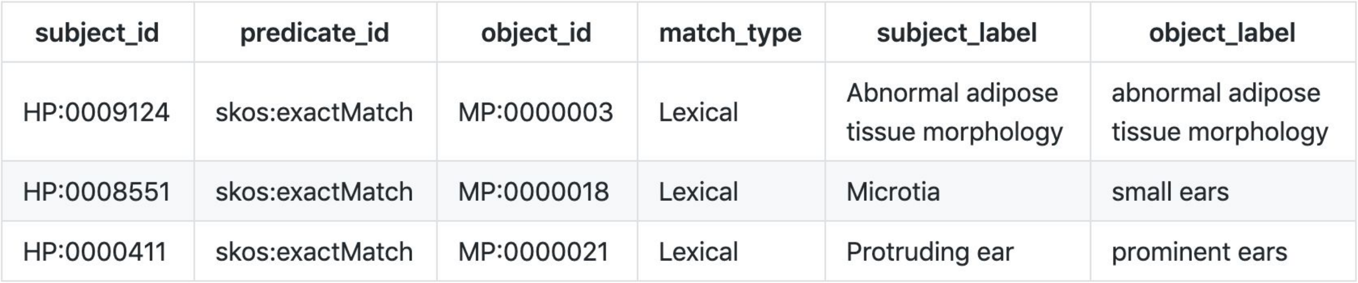
A Simple Standard for Sharing Ontological Mappings (SSSOM)

Nicolas Matentzoglu, James P Balhoff, Susan M Bello, Chris Bizon, Matthew Brush, Tiffany J Callahan, Christopher G Chute, William D Duncan, Chris T Evelo, Davera Gabriel ... Show more

Database, Volume 2022, 2022, baac035, https://doi.org/10.1093/database/baac035

Published: 25 May 2022 **Article history** ▾

- TSV standard with standard metadata elements to describe mapping



subject_id	predicate_id	object_id	match_type	subject_label	object_label
HP:0009124	skos:exactMatch	MP:0000003	Lexical	Abnormal adipose tissue morphology	abnormal adipose tissue morphology
HP:0008551	skos:exactMatch	MP:0000018	Lexical	Microtia	small ears
HP:0000411	skos:exactMatch	MP:0000021	Lexical	Protruding ear	prominent ears

Mapping predicates include owl:EquivalentClass; OBO xref
Other columns include confidence score; mapping tool

https://github.com/mapping-commons/SSSOM

EMBL-EBI

But how do we define the mappings?!

- **Manual curation:** an expert in the subject matter looks at the sets of terms and asserts that they map to each other
- **Lexical matching:** String similarity of labels
- **Semantic similarity:** e.g. Jaccard; what is the semantic overlap of the terms
- And everything in between, e.g. combining these methods with human supervision to make a semi-automated mapping pipeline
- **Next: LLMs ?!**

Types of mappings

- **Clear correspondence**
  - E.g. skos:exactMatch, owl:equivalentClass
- **Fuzzy correspondence**
  - E.g. gene <codes for> protein
  - Glucose <is measured by> glucose level measurement
- **Associations** however are different from mappings
  - E.g. student <enrolled in> university
  - Phone <manufactured in> china

Source: Nico Matentzoglu
https://github.com/mapping-commons/sssom/issues/324

EMBL-EBI

Slightly different: String to term mappings

- Mappings can be **term to term**….
  - **DOID:162** (cancer) —> **MONDO:0004992** (cancer)
  - (Like all of the examples we have just seen)

- Another class of mappings: **string to term**
  - *"Type 2 diabetes"* **—> EFO:0001360** "type II diabetes mellitus"

  - Important for
    - extracting ontology terms from bodies of text
    - Importing datasets that use free text rather than terms

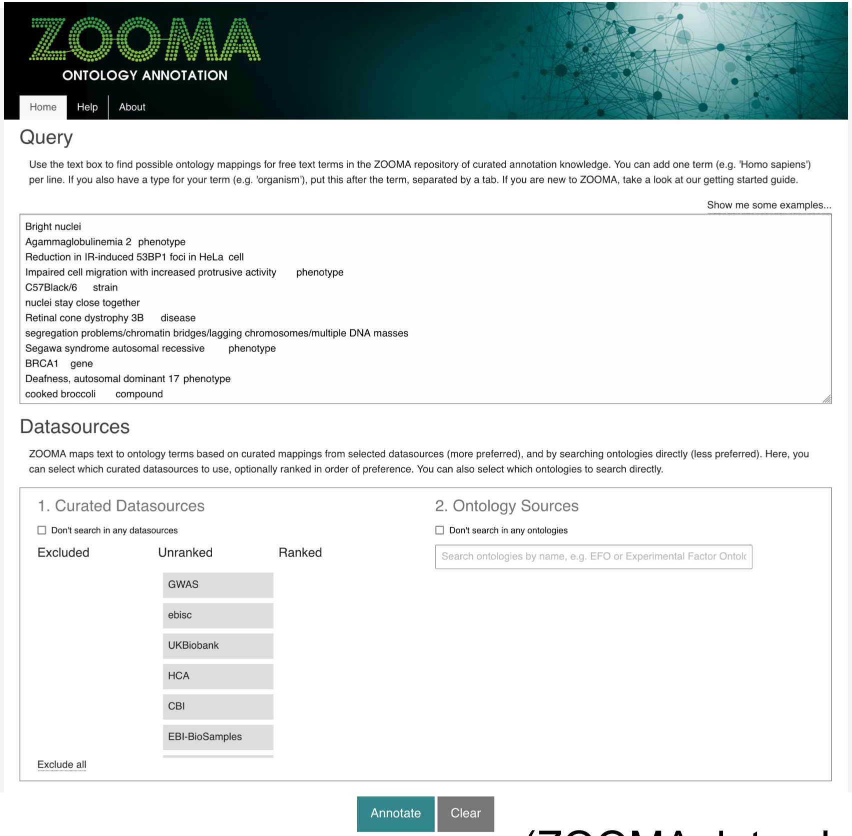
String to term mapping in SSSOM

SSSOM

SIMPLE STANDARD FOR SHARING
ONTOLOGY MAPPINGS

```
literal mapping:
  description: Represents an individual mapping between a a literal and an entity
  slots:
  - literal
  - literal_datatype
  - predicate_id
  - predicate_label
  - predicate_modifier
  - object_id
  - object_label
  - object_category
  - mapping_justification
  - author_id
  - author_label
  - reviewer_id
  - reviewer_label
  - creator_id
  - creator_label
  - license
  - literal_source
  - literal_source_version
  - object_type
  - object_source
  - object_source_version
  - mapping_provider
```

String to term mappings : ZOOMA



Query

Use the text box to find possible ontology mappings for free text terms in the ZOOMA repository of curated annotation knowledge. You can add one term (e.g. 'Homo sapiens') per line. If you also have a type for your term (e.g. 'organism'), put this after the term, separated by a tab. If you are new to ZOOMA, take a look at our getting started guide.

Show me some examples...

```
Bright nuclei
Agammaglobulinemia 2  phenotype
Reduction in IR-induced 53BP1 foci in HeLa  cell
Impaired cell migration with increased protrusive activity      phenotype
C57Black/6      strain
nuclei stay close together
Retinal cone dystrophy 3B       disease
segregation problems/chromatin bridges/lagging chromosomes/multiple DNA masses
Segawa syndrome autosomal recessive        phenotype
BRCA1    gene
Deafness, autosomal dominant 17 phenotype
cooked broccoli       compound
```

Datasources

ZOOMA maps text to ontology terms based on curated mappings from selected datasources (more preferred), and by searching ontologies directly (less preferred). Here, you can select which curated datasources to use, optionally ranked in order of preference. You can also select which ontologies to search directly.

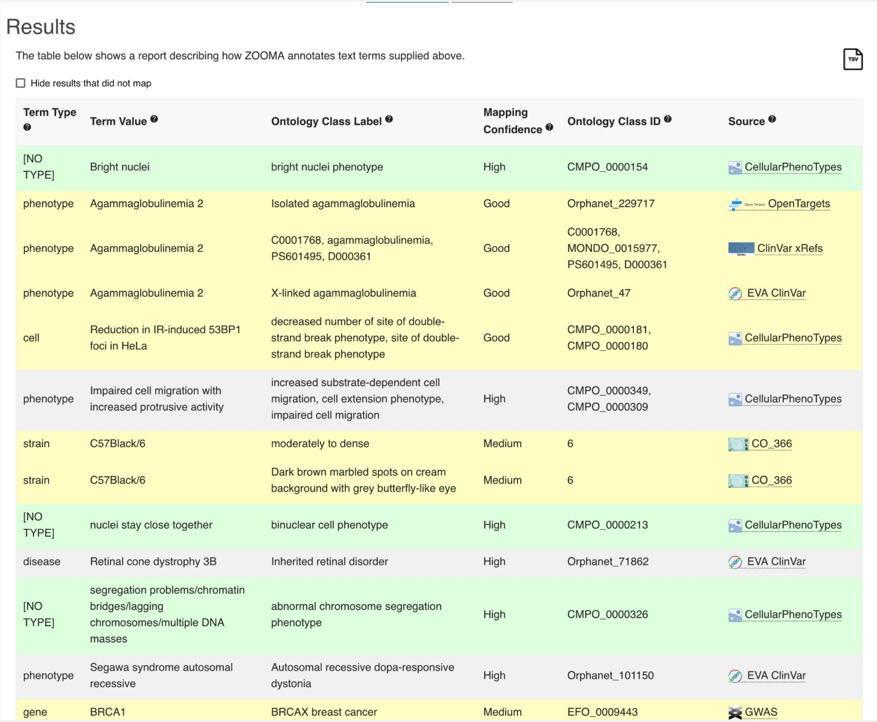
1. Curated Datasources

☐ Don't search in any datasources

2. Ontology Sources

☐ Don't search in any ontologies

Search ontologies by name, e.g. EFO or Experimental Factor Ontolo
Annotate
Clear



Results

The table below shows a report describing how ZOOMA annotates text terms supplied above.

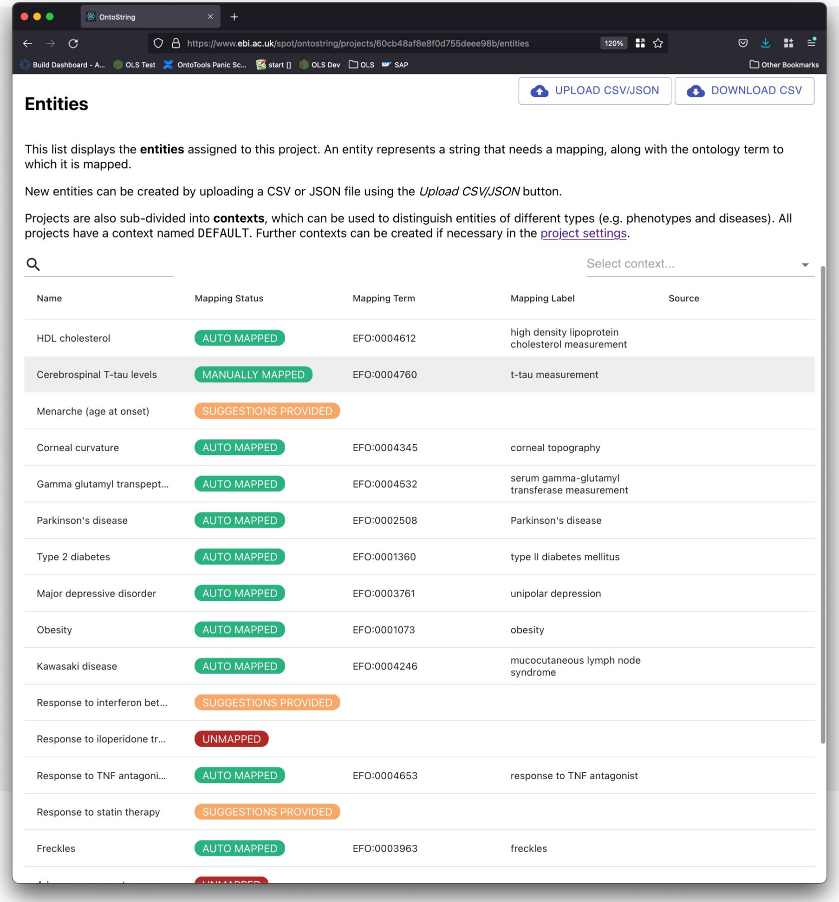
☐ Hide results that did not map

Term Type	Term Value	Ontology Class Label	Mapping Confidence	Ontology Class ID	Source
[NO TYPE]	Bright nuclei	bright nuclei phenotype	High	CMPO_0000154	CellularPhenoTypes
phenotype	Agammaglobulinemia 2	Isolated agammaglobulinemia	Good	Orphanet_229717	OpenTargets
phenotype	Agammaglobulinemia 2	C0001768, agammaglobulinemia, PS601495, D000361	Good	C0001768, MONDO_0015977, PS601495, D000361	ClinVar xRefs
phenotype	Agammaglobulinemia 2	X-linked agammaglobulinemia	Good	Orphanet_47	EVA ClinVar
cell	Reduction in IR-induced 53BP1 foci in HeLa	decreased number of site of double-strand break phenotype, site of double-strand break phenotype	Good	CMPO_0000181, CMPO_0000180	CellularPhenoTypes
phenotype	Impaired cell migration with increased protrusive activity	increased substrate-dependent cell migration, cell extension phenotype, impaired cell migration	High	CMPO_0000349, CMPO_0000309	CellularPhenoTypes
strain	C57Black/6	moderately to dense	Medium	6	CO_366
strain	C57Black/6	Dark brown marbled spots on cream background with grey butterfly-like eye	Medium	6	CO_366
[NO TYPE]	nuclei stay close together	binuclear cell phenotype	High	CMPO_0000213	CellularPhenoTypes
disease	Retinal cone dystrophy 3B	Inherited retinal disorder	High	Orphanet_71862	EVA ClinVar
[NO TYPE]	segregation problems/chromatin bridges/lagging chromosomes/multiple DNA masses	abnormal chromosome segregation phenotype	High	CMPO_0000326	CellularPhenoTypes
phenotype	Segawa syndrome autosomal recessive	Autosomal recessive dopa-responsive dystonia	High	Orphanet_101150	EVA ClinVar
gene	BRCA1	BRCAX breast cancer	Medium	EFO_0009443	GWAS

(ZOOMA data also now available in SSSOM)

EMBL-EBI

String to term mapping : OntoString (beta)



Entities

This list displays the **entities** assigned to this project. An entity represents a string that needs a mapping, along with the ontology term to which it is mapped.

New entities can be created by uploading a CSV or JSON file using the *Upload CSV/JSON* button.

Projects are also sub-divided into **contexts**, which can be used to distinguish entities of different types (e.g. phenotypes and diseases). All projects have a context named DEFAULT. Further contexts can be created if necessary in the project settings.

Name	Mapping Status	Mapping Term	Mapping Label	Source
HDL cholesterol	AUTO MAPPED	EFO:0004612	high density lipoprotein cholesterol measurement	
Cerebrospinal T-tau levels	MANUALLY MAPPED	EFO:0004760	t-tau measurement	
Menarche (age at onset)	SUGGESTIONS PROVIDED			
Corneal curvature	AUTO MAPPED	EFO:0004345	corneal topography	
Gamma glutamyl transpept...	AUTO MAPPED	EFO:0004532	serum gamma-glutamyl transferase measurement	
Parkinson's disease	AUTO MAPPED	EFO:0002508	Parkinson's disease	
Type 2 diabetes	AUTO MAPPED	EFO:0001360	type II diabetes mellitus	
Major depressive disorder	AUTO MAPPED	EFO:0003761	unipolar depression	
Obesity	AUTO MAPPED	EFO:0001073	obesity	
Kawasaki disease	AUTO MAPPED	EFO:0004246	mucocutaneous lymph node syndrome	
Response to interferon bet...	SUGGESTIONS PROVIDED			
Response to iloperidone tr...	UNMAPPED			
Response to TNF antagoni...	AUTO MAPPED	EFO:0004653	response to TNF antagonist	
Response to statin therapy	SUGGESTIONS PROVIDED			
Freckles	AUTO MAPPED	EFO:0003963	freckles	

- Semi-automated mapping tool
- Users can upload a list of terms, OntoString will guess (using ZOOMA and OLS) a mapping based on previously manually curated mappings
- If no mapping is available, the user can explicitly map it
- This data (string -> term mappings) will eventually inform ZOOMA

(Migrating to SSSOM.)

Conclusion

- We at EBI use mappings extensively to link biomedical datasets
- We are trying to move away from explicit mappings to **semantic equivalence** by better defining terms in ontologies
- But we still use lots of methods for mapping e.g. manual curation, lexical matching, semantic similarity
    - And maybe LLMs soon?
- We use mappings for relationships that are correspondences rather than associations
- We are aggressively adopting SSSOM in all of our mapping infrastructure!
    - And please don’t forget about string to term mappings