https://w3id.org/sssom

sssom

SIMPLE STANDARD FOR SHARING ONTOLOGY MAPPINGS

FAIR Impact Workshop: SSSOM - a machine actionable model for simple entity mappings

Nicolas Matentzoglu, Why Mappings Matter and how to make them FAIR?
Workshop, 13.04.2023

https://www.lifewatch.eu/events/why-mappings-matter-and-how-to-make-them-fair-a-fair-impact-workshop/

monarch INITIATIVE

* SSSOM can be pronounced "sessom"

What do we mean by “mapping”?

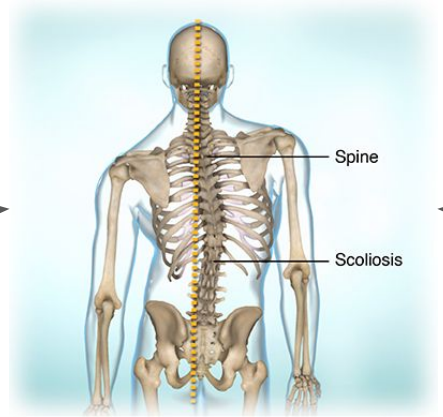
- **Entity mapping:** Determining and documenting the correspondence of an entity in one semantic space to another.
- **Schema mapping:** Determining and documenting the translation rules for converting an entity from one semantic space to another.

What do we mean by “mapping”?

- **Entity mapping:** Determining and documenting the correspondence of a entity in one semantic space to another.
- ~~**Schema mapping:** Determining and documenting the translation rules for converting an entity from one semantic space to another.~~

Today is all about entity mapping - this is not to say that schema mapping is not super relevant as well for us.

What are entity mappings?



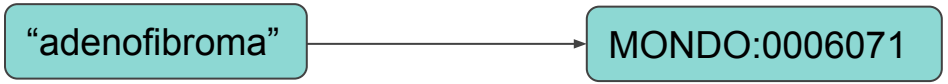
“Friedreich's Ataxia”

OMOP:441554

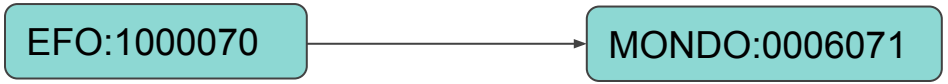
Entities are symbols, such as codes in a terminology, classes in an ontology, enums in a data model, or simply strings in a text field that are **intended to refer to a real world thing.**

Different types of entity mappings

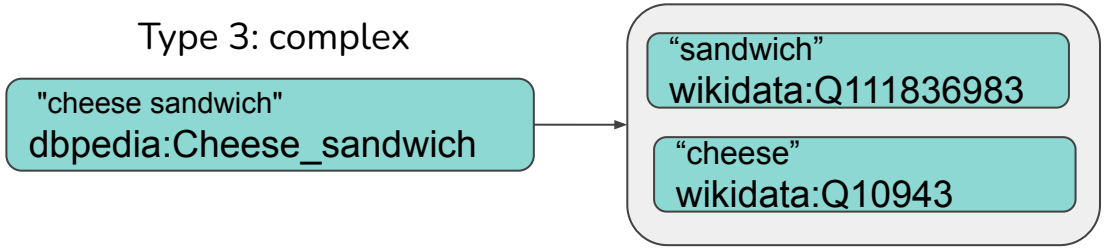
Type 1: string - identifier



Type 2: identifier - identifier



Type 3: complex



What is the problem?

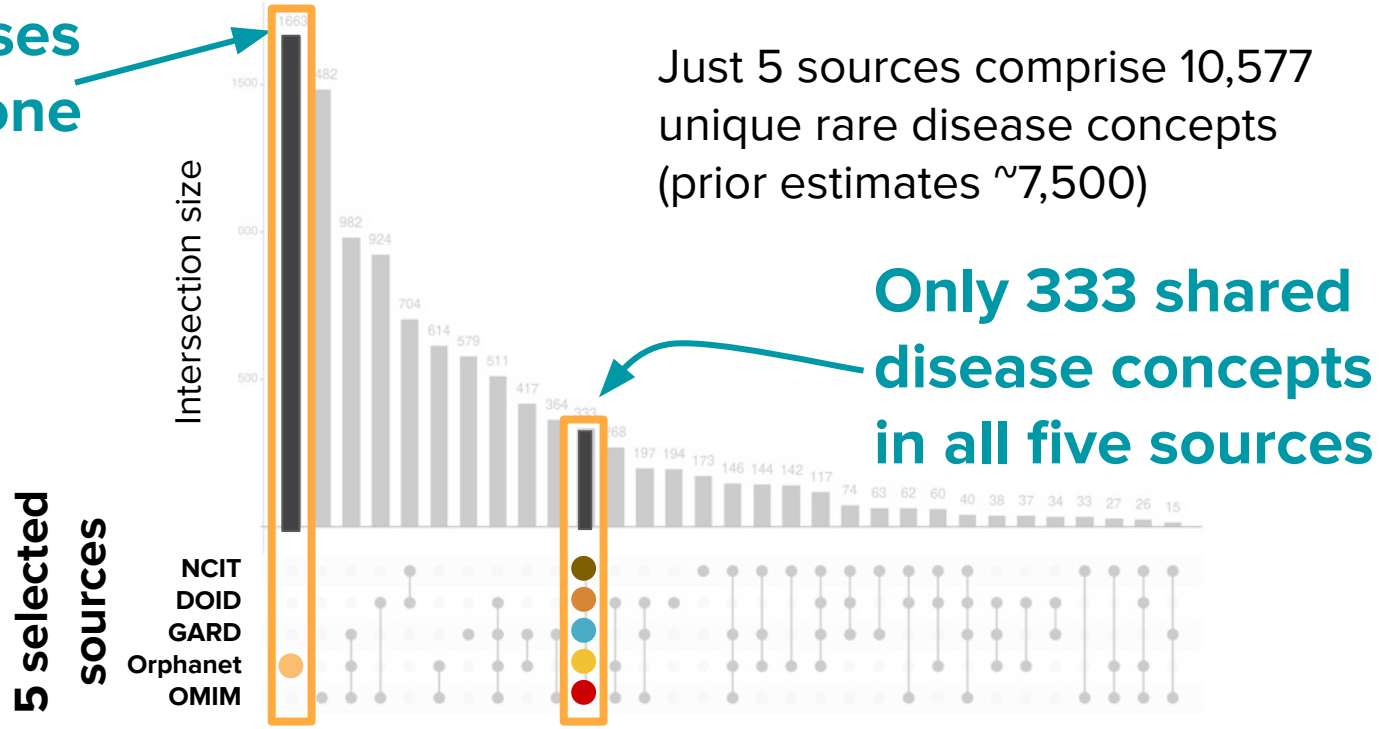
Mappings frequently lack semantic precision

ICD10CM Code	ICD10CM Label	Relation	OMOP Label	OMOP ID
A06	Amebiasis	Maps to	Amebic infection	438959
D46.A	Refractory anaemia with multi-lineage dysplasia	Maps to	Refractory anemia	4003185

https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:mapping

We need to enable analytics across semantic spaces

Many diseases are in only one source



Just 5 sources comprise 10,577 unique rare disease concepts (prior estimates ~7,500)

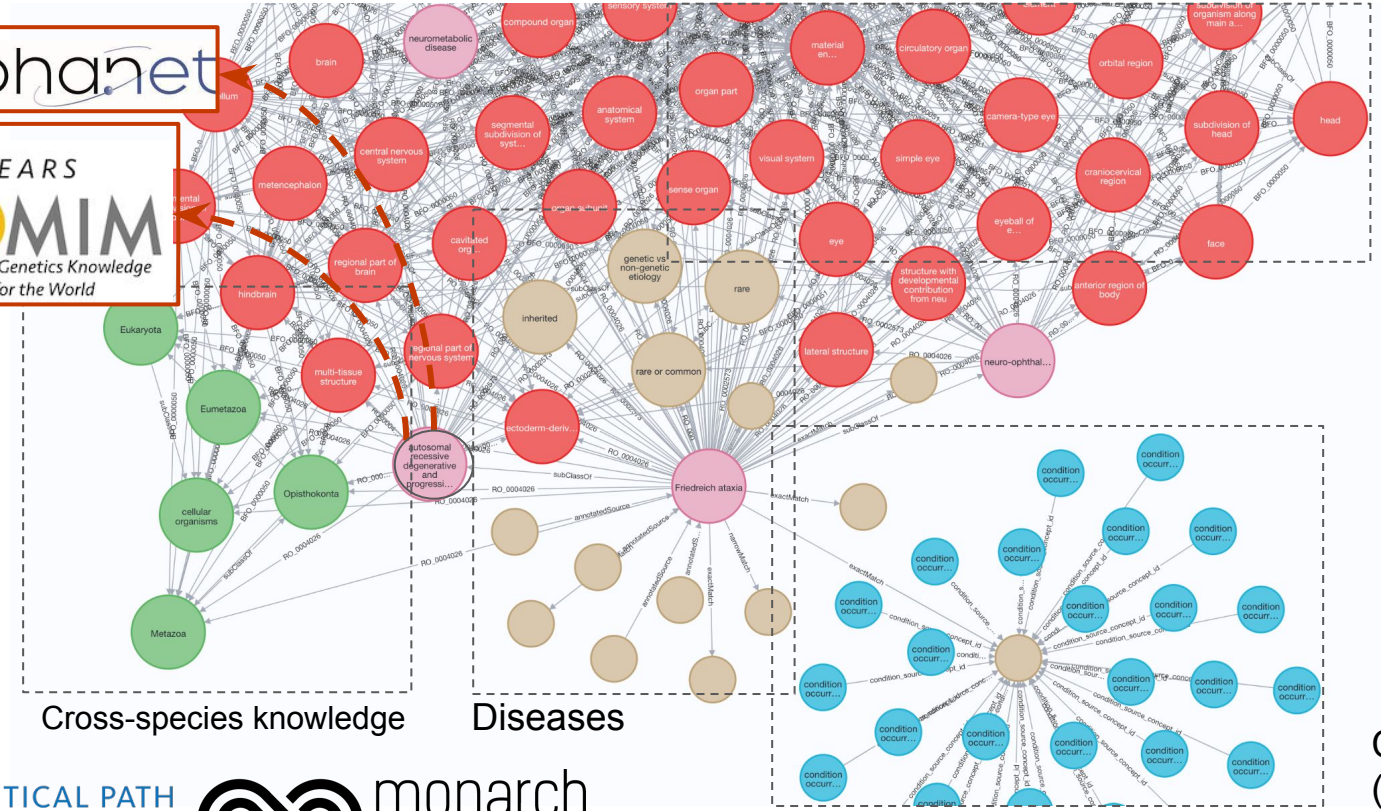
Only 333 shared disease concepts in all five sources

hpo

Widely used for rare disease, est. 25% coverage in SNOMED

Nature Reviews Drug Discovery (bit.ly/nature-rare-diseases)

We need to integrate genomics and clinical data across diverse semantic spaces



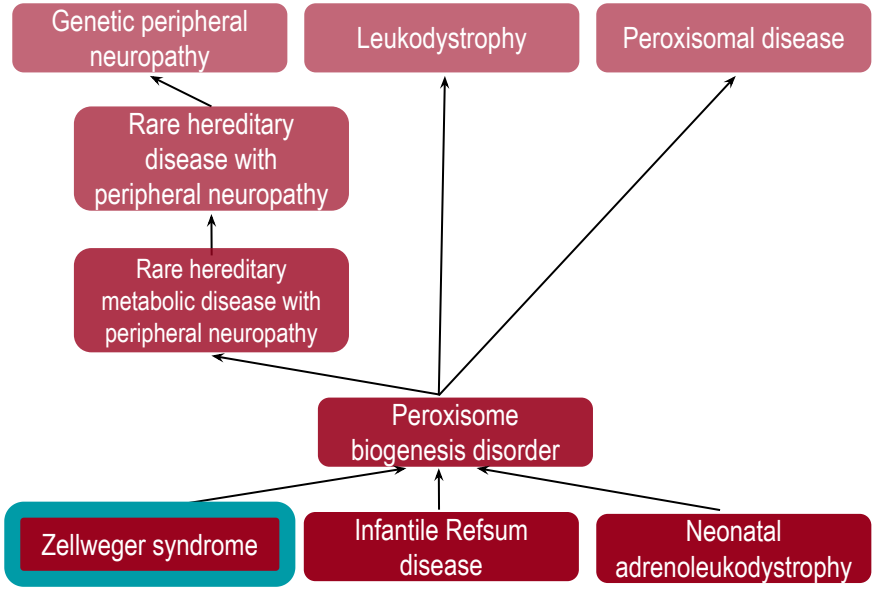
Anatomical reference models

CRITICAL PATH INSTITUTE

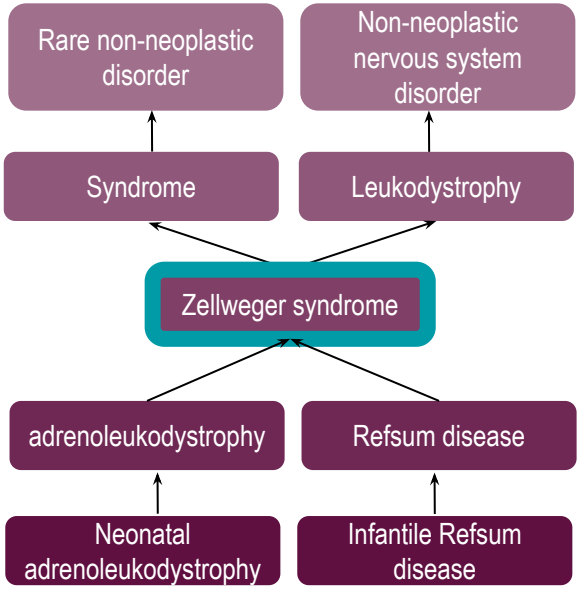
monarch INITIATIVE

Curating good mappings is hard - and therefore costly

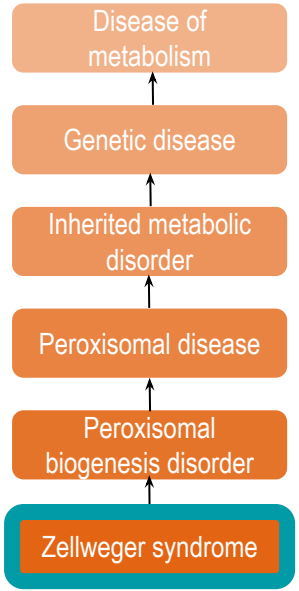
Orphanet



NCIt



DO



https://github.com/monarch-initiative/mondo/issues/61

We can do better.

- **Share** mappings to avoid creating the same mapping over and over again
- **Enrich** mappings with metadata to enable the combination of mappings from different sources
- **Build** a coordinated decentralised effort of human curators, similar to what we do for (open) ontologies.

FAIRSFAIR
Fostering Fair Data Practices in Europe

The SSSOM Metadata Model

Rich YAML schema powered by LinkML

```
267     - other
268     - comment
269   mapping:
270     description: Represents an individual
271     slots:
272     - subject_id
273     - subject_label
274     - subject_category
275     - predicate_id
276     - predicate_label
277     - object_id
278     - object_label
279     - object_category
```

Shex shapes for validating rdf

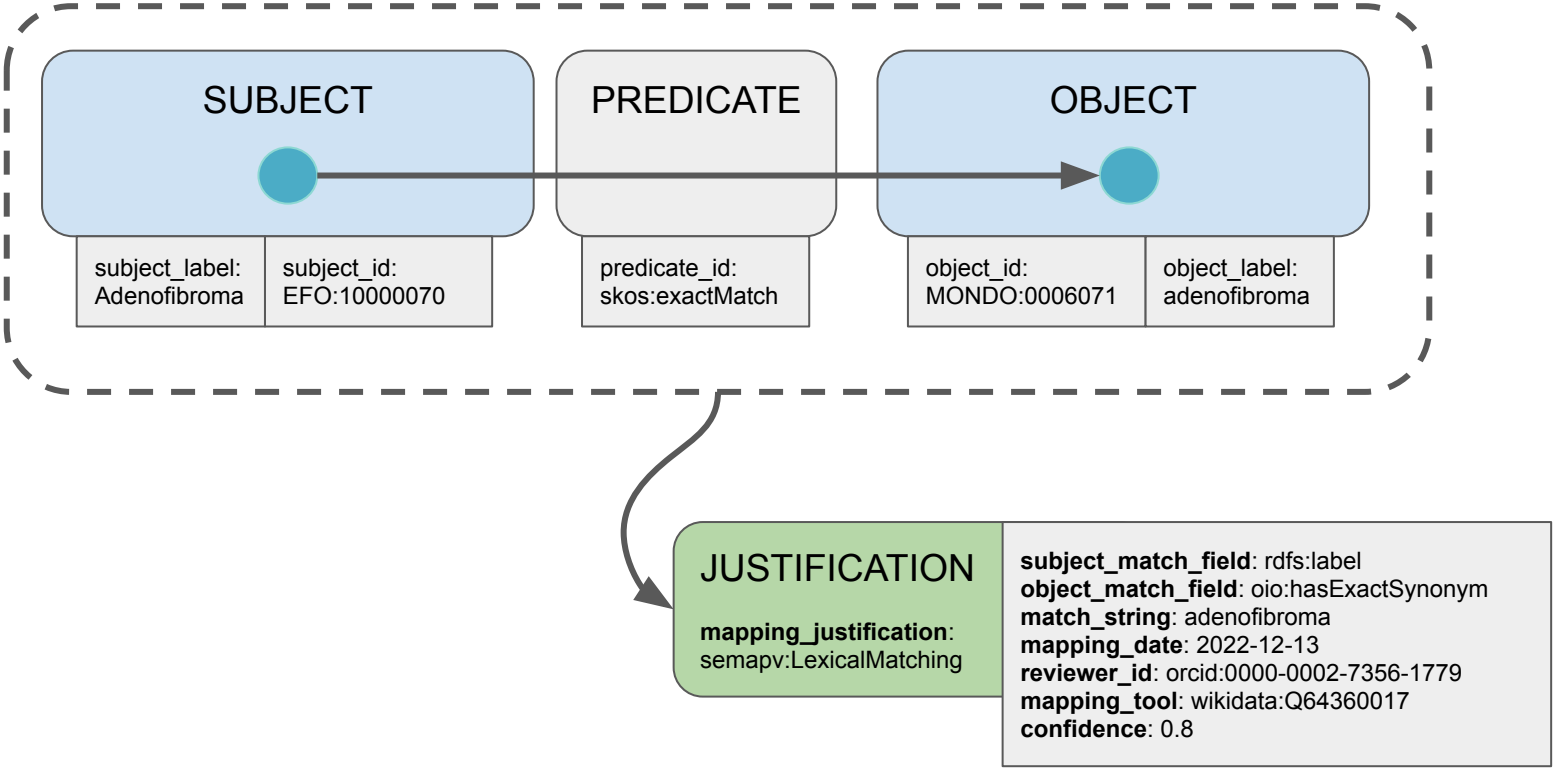
JSON Schema

Markdown docs

Browse: https://w3id.org/sssom/spec

- **subject_id**
- subject label
- subject category
- **predicate_id**
- predicate_label
- **object_id**
- object label
- object category
- **match_type**
- creator id
- creator_label
- license
- subject source
- subject source_version
- object source
- object source version
- mapping provider
- mapping cardinality
- mapping tool
- mapping date
- **confidence**
- subject match field
- object match_field
- match string
- subject preprocessing
- object preprocessing
- match term type
- semantic_similarity_score
- see also
- other
- comment

The anatomy of a semantic entity mapping



Example SSSOM TSV file

Can be exported to JSON, RDF, etc.

```
#mapping_set_id: MGI_Full_MP_HPO
#mapping_set_title: All mappings of MP terms to HPO terms generated by MGI
#mapping_set_description: "Consolidated list of all HPO to MP mappings done by MGI…."
#creator_id:
# - orcid:0000-0003-4606-0597
# - orcid:0000-0002-6490-7723
# - orcid:0000-0003-2307-1226
# - ror:021sy4w91
# - wikidata:Q1951035
#license: https://creativecommons.org/licenses/by/4.0/
#object_source: obo:hp
#subject_source: obo:mp
#curie_map:
# HP: http://purl.obolibrary.org/obo/HP_
# MP: http://purl.obolibrary.org/obo/MP_
```

Provenance and descriptions

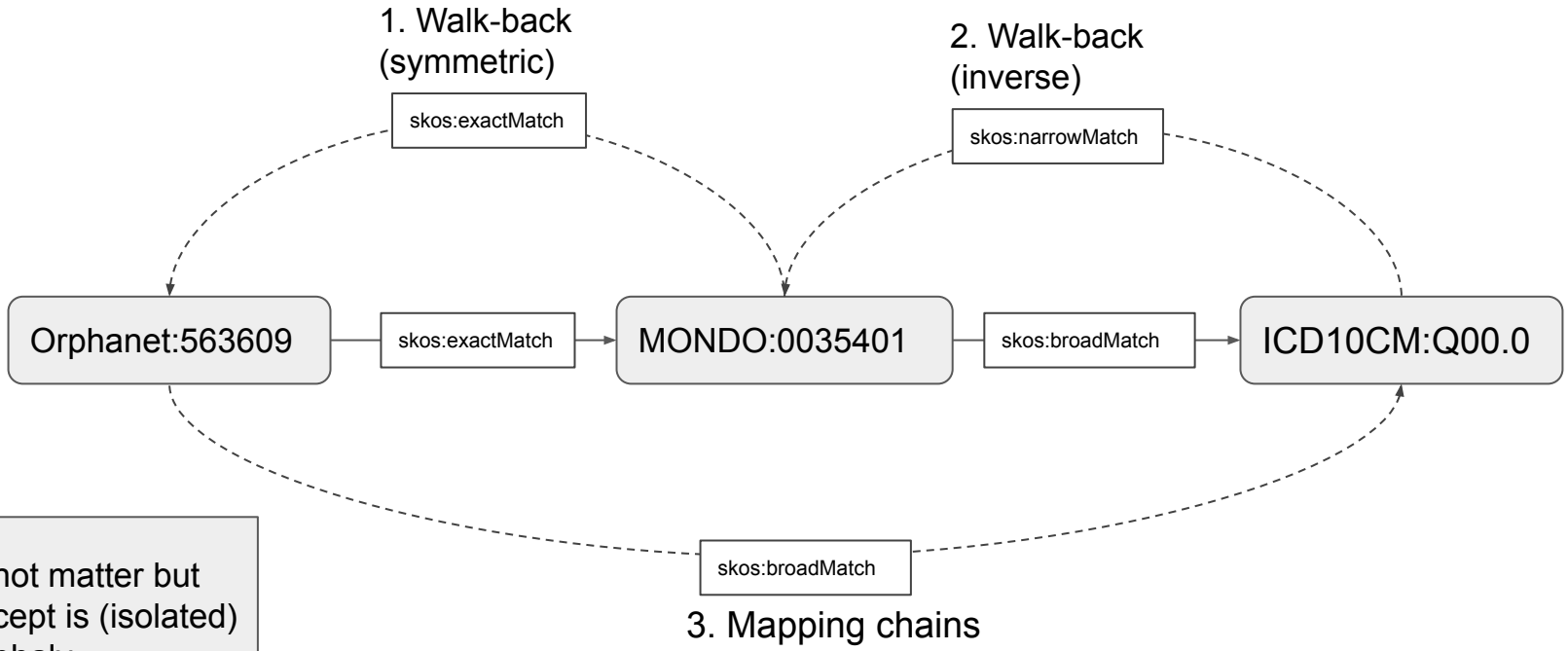
https://bit.ly/ohdsi-sssom-example

Licensing information in header

Mapping Table

object_id	object_label	predicate_id	confidence	subject_id	subject_label	mapping_justification	author_id	mapping_date	comment
HP:0000016	Urinary retention	skos:exactMatch	1	MP:0003622	ischuria	semapv:ManualMapp	orcid:0000-0003	2022-08-02	scoliosis
HP:0000023	Inguinal hernia	skos:exactMatch	1	MP:0006077	inguinal hernia	semapv:ManualMapp	orcid:0000-0003	2021-05-27	KidsFirst
HP:0000028	Cryptorchidism	skos:exactMatch	1	MP:0002286	cryptorchism	semapv:ManualMapp	orcid:0000-0003	2021-05-27	KidsFirst
HP:0000033	Ambiguous genital	skos:narrowMatch	1	MP:0009198	abnormal male	semapv:ManualMapp	orcid:0000-0003	2022-02-07	KidsFirst; ex
HP:0000034	Hydrocele testis	skos:narrowMatch	1	MP:0003623	hydrocele	semapv:ManualMapp	orcid:0000-0002	2021-05-27	KidsFirst; M

Semantic mapping predicates

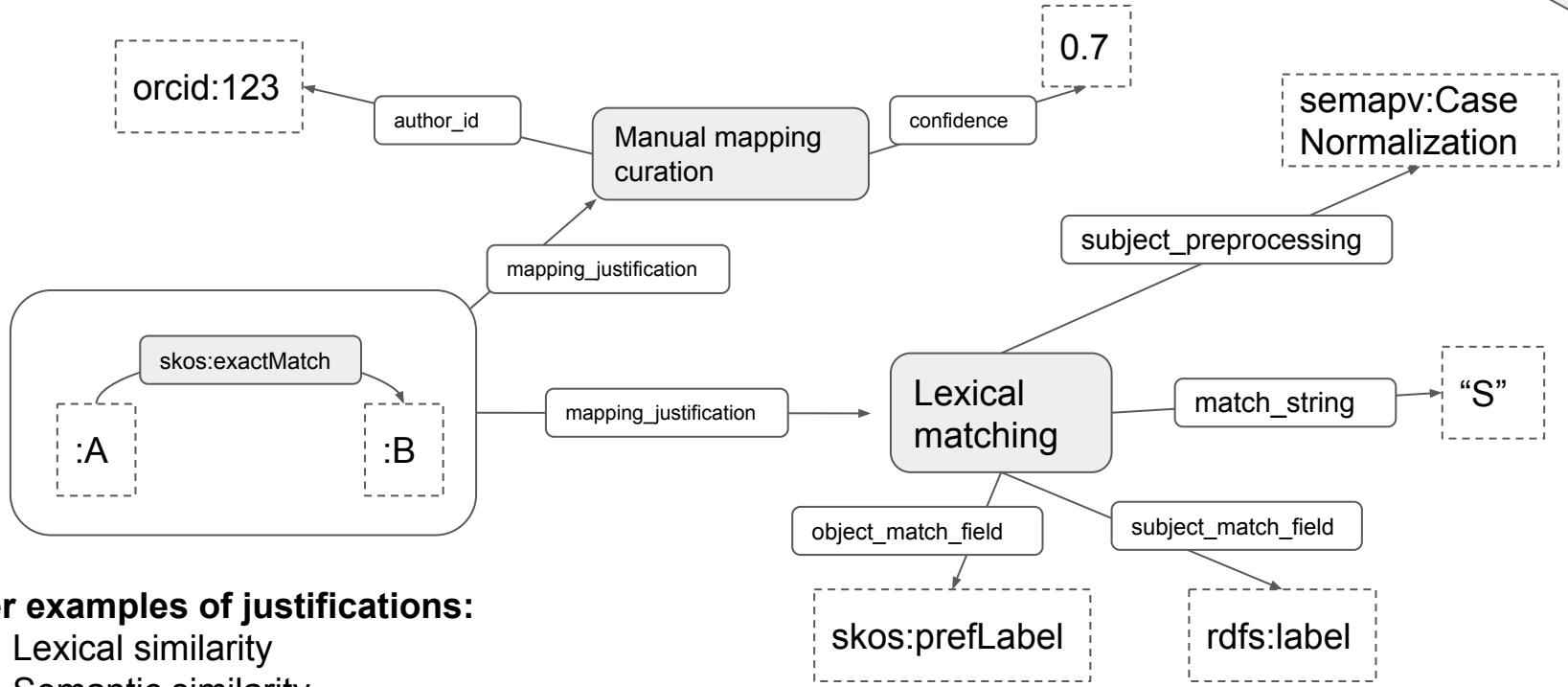


It does not matter but the concept is (isolated) anencephaly

Tutorial: https://mapping-commons.github.io/sssom/mapping-predicates/

Mapping justifications

A mapping can have more than one justification!



Other examples of justifications:

- Lexical similarity
- Semantic similarity
- Mapping Chaining

Dereferencable identifiers

- Most metadata elements in SSSOM require the use of “entity references” rather than simple strings
- An “entity reference” should be a globally unique, persistent (and resolvable) identifier (GUPRI)
- The entity reference itself is usually recorded as a Compact URI, or CURIE, which can be resolved to a URI using a special “curie_map”.

Tip of the day: Use dereferencable identifiers **to refer to people**, rather than labels!

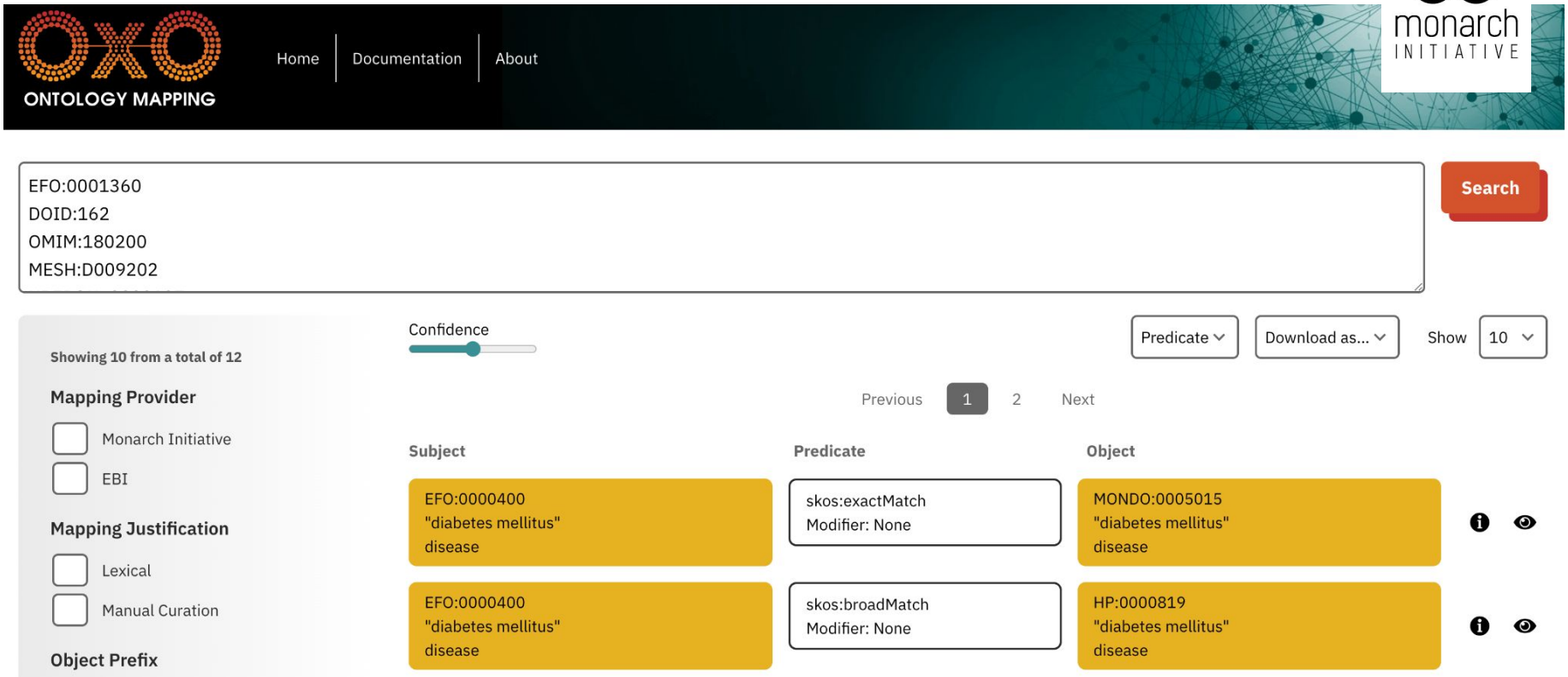
SSSOM Toolkit and other SSSOM related tools

- SSSOM toolkit (https://github.com/mapping-commons/sssom-py)
  - Design philosophy of SSSOM to not require any special tooling
  - Utility methods such as
    - "merge" (to merge two mapping sets)
    - "parse" (to convert a different format, such as EDOAL, into SSSOM)
    - "validate" (to check that a mapping set is legal SSSOM)
    - "filter" command allows to filter a mapping set based on any of its metadata elements
- The Ontology Access Kit (OAK) implements functionality to do basic lexical matching based on term synonyms and extracting SSSOM mappings from ontologies

A first SSSOM aware mapping browser is developed at EBI: OxO 2

EMBL-EBI

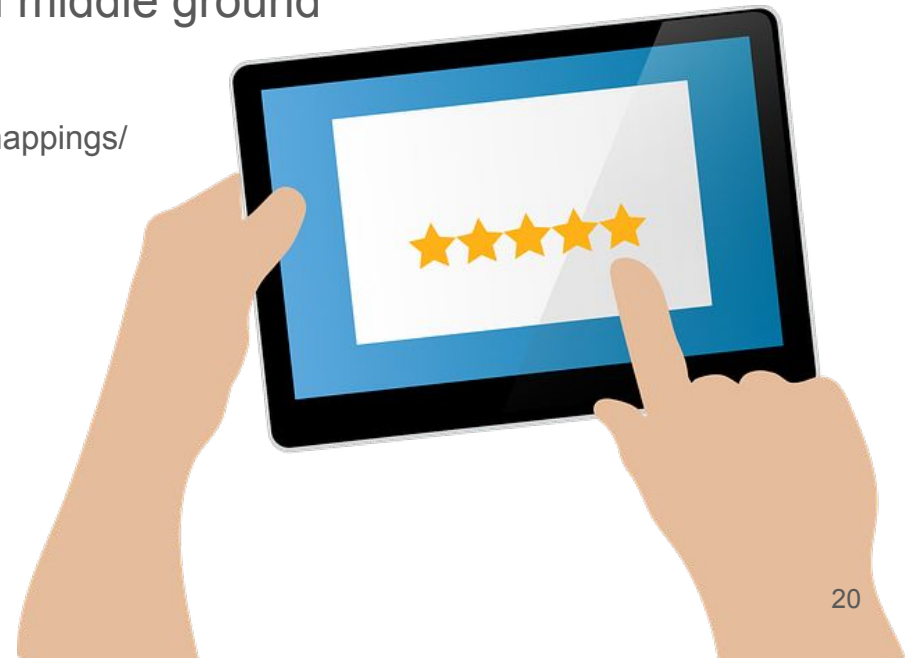
monarch INITIATIVE



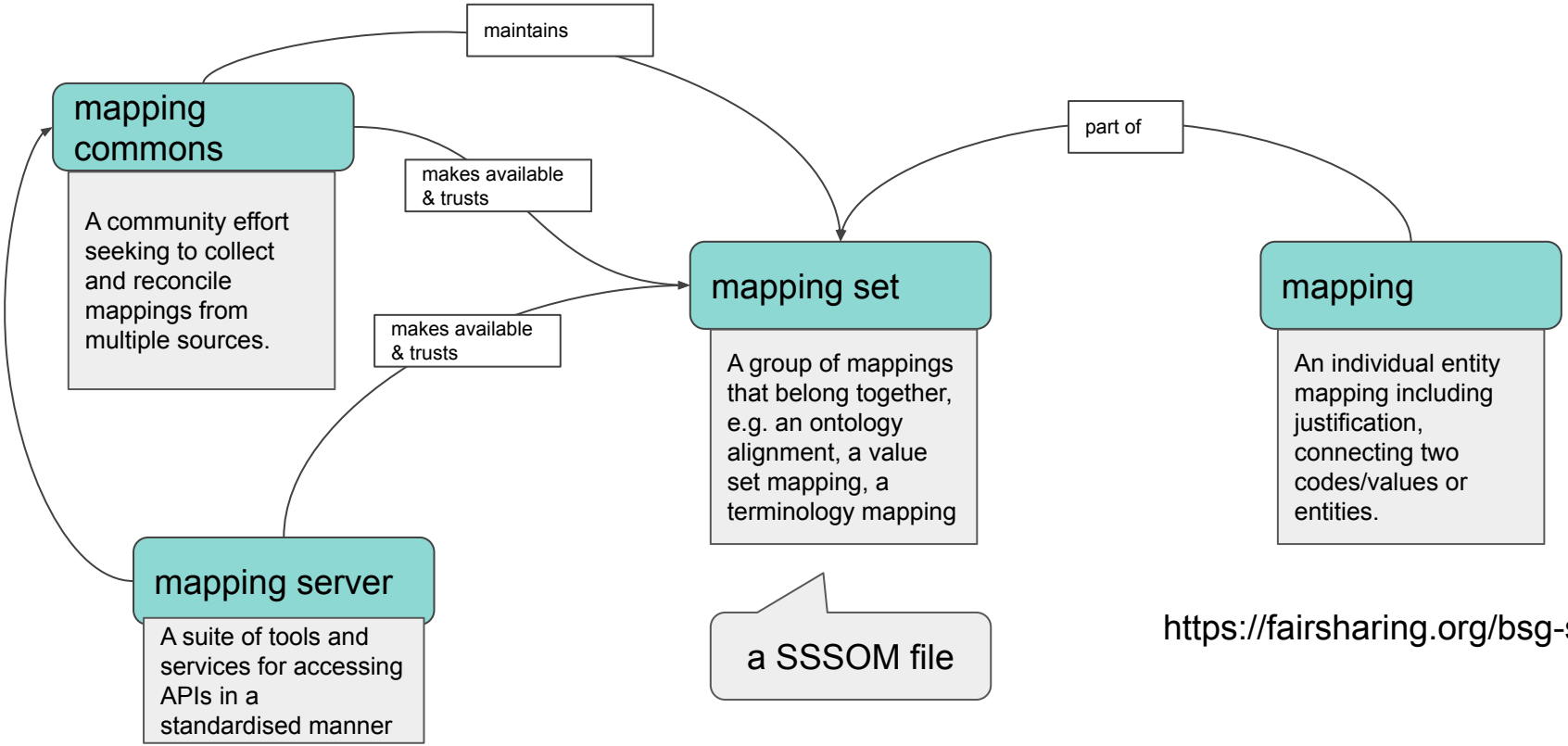
What do we do now? A Five-Star system for mappings

- Standardisation of mappings is costly
- Sometimes, perfect is the enemy of the good enough
- As a community we need to find a good middle ground for what is “good enough”.
- See https://mapping-commons.github.io/sssom/5star-mappings/

Please lobby all mapping providers out there to publish mappings using CC-0 or CC-BY licenses under a public URL!

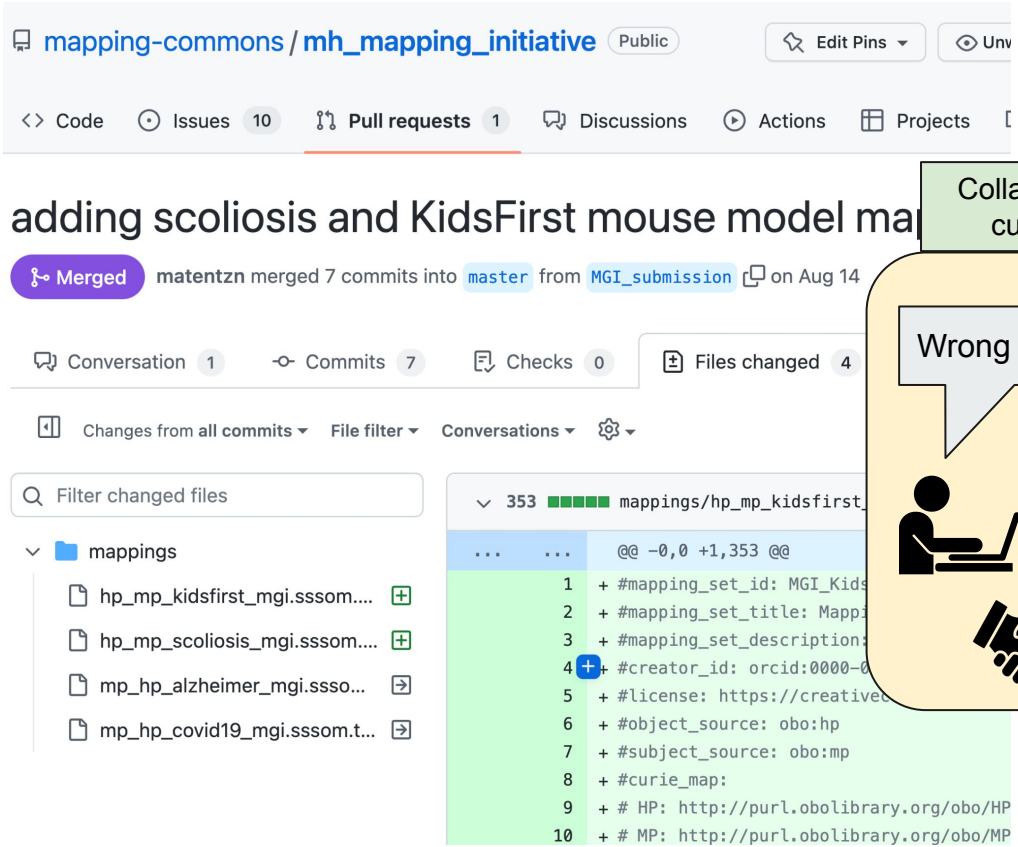


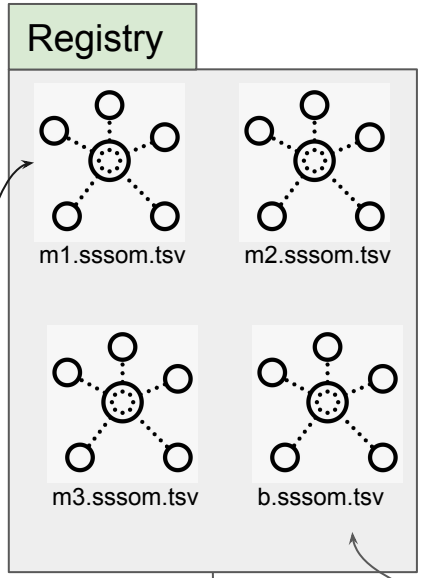
The bigger picture

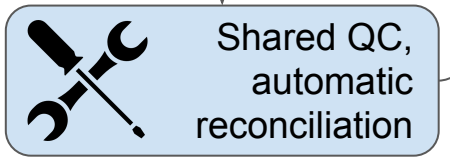


https://fairsharing.org/bsg-s001618

Building FAIR mapping registries: Mapping Commons







Acknowledgements

Funding

Phenomics First (NIH / NHGRI #1RM1HG010860-01): Spec, Mondo integration, sssom-py CLI

Monarch (NIH / OD #5R24OD011883): Cross-species mappings, outreach, knowledge graph integration

Bosch Gift to LBNL: sssom-py IO, testing, converters, tutorials

DARPA: Young Faculty Award W911NF2010255 (PI: Benjamin M. Gyori)

BOSCH EMBL-EBI

Community contributions:
https://w3id.org/sssom

Core Team, alphabetical order

(https://github.com/orgs/mapping-commons/teams/sssom-core)

- Alex H. Wagner (Nationwide Children's Hospital)
- **Anita Caron (EMBL-EBI)**
- Charlie Hoyt (Harvard Medical School)
- Chris Mungall (LBNL)
- **Damien Goutte-Gattat (Flybase)**
- David Osumi-Sutherland (EMBL-EBI)
- **Emily Hartley (C-Path)**
- Ernesto Jimenez-Ruiz (City, Univ. of London)
- Harshad Hegde (LBNL)
- Henriette Harmse (EMBL-EBI)
- Hyeongsik Kim (Bosch)
- Ian Harrow (Pistoia Alliance)
- James McLaughlin (EMBL-EBI)
- Jim Balhoff (RENCI)
- John Graybeal (Stanford)
- Melissa Haendel (CU Anschutz)
- Nicolas Matentzoglu (EMBL-EBI)
- Nicole Vasilevsky (CU Anschutz)
- Núria Queralt Rosinach
- Simon Jupp (SciBite)
- **Sophie Aubin (INRAE)**
- Thomas Liener (Pistoia Alliance)
- Tiffany Callahan (Columbia University)
- Tim Putman (CU Anschutz)
- William Duncan (UFlorida)
- ...many more contributors, see publication

*recent joiners highlighted in bold

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