A Simple Standard for Sharing Ontology Mappings (SSSOM)

A plaidoyer for caring more about the standardisation and dissemination of ontology mappings.

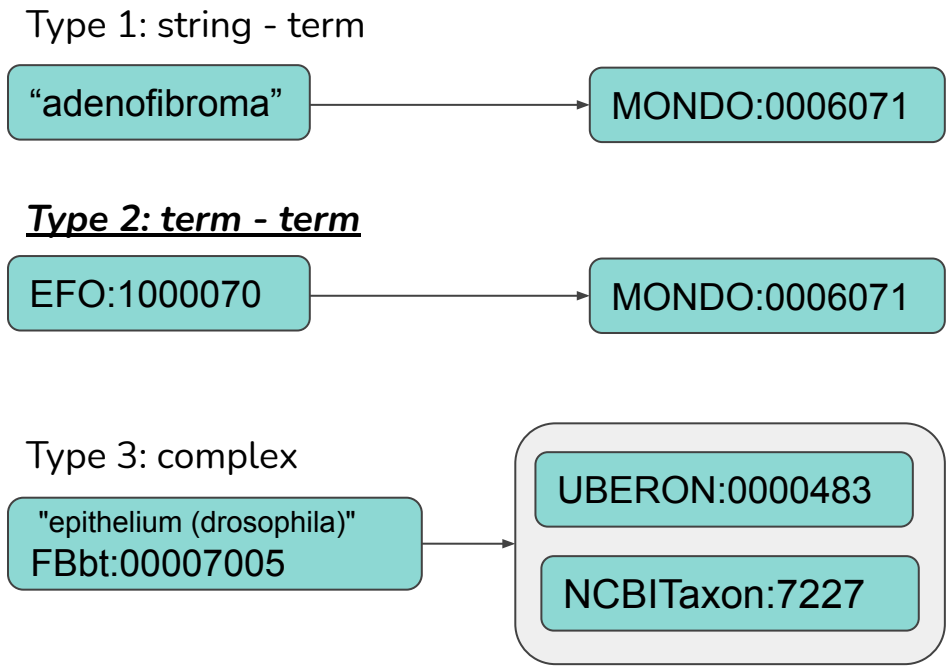
Matentzoglu, Nicolas; Balhoff, James P.; Callahan, Tiffany; Chute, Christopher; Dahzi, Jiao; Duncan, William; Gabriel, Davera; Graybeal, John; Haendel, Melissa; Harmse, Henriette; Harold, Solbrig; Harris, Nomi; Hegde, Harshad; Hoyt, Charles Tapley; Jimenez-Ruiz, Ernesto; Jupp, Simon; Kim, Hyeongsik; Koehler, Sebastian; Liener, Thomas; Malone, James; McLaughlin, James; Munoz-Torres, Monica; Osumi-Sutherland, David; Overton, James; Thessen, Anne; Vasilevsky, Nicole; Mungall, Chris

WSBO - 14.07.2021

https://fairsharing.org/bsg-s001618

What is a mapping (in the sense of this talk)?

- a mapping is a *correspondence* of two terms
- a mapping set is a set of term mappings, usually assembled for a particular purpose
- an alignment is a special kind of mapping set which contains all mappings between two ontologies



Convergence vs Mapping

- **Mapping** and **Convergence** are two complementary approaches to achieve **synergy**
  - Convergence: continuous ontological unification (OBO Foundry)
  - Mappings: bridging across ontologies (UMLS, OxO)
- Convergence is the ideal, but the idea of total convergence is unrealistic
  - Extremely expensive, hard to agree on, “political” issues
  - In Medical Terminology, it is not even considered anymore
- Mapping is the reality, and they are are essential for bridging semantic spaces/silos

How are mappings curated and used in practice?

- **Monarch Initiative, IMPC and MOD databases**: Cross-species mappings
  - Increasing the clinical relevance of model organism data: search, variant prioritisation
  - Using an automated logic-based approach to identify similar phenotypes
- **IHCC**: Curators map data dictionaries for description of rare disease cohorts to an application ontology for genomics cohorts
- **Bosch**: Enriching Bosch's internal ontologies with external ones such as FoodOn, Wikidata, and GAO, etc. (to support smart assistant agents and recommendation platforms)

Mappings are essential to bridge semantic spaces, but they are hard to use.. and share.

owl:sameAs

owl:equivalentClass

owl:equivalentProperty

rdfs:subClassOf

rdfs:subPropertyOf

skos:relatedMatch

skos:closeMatch

skos:exactMatch

skos:narrowMatch

skos:broadMatch

oio:database_cross_reference

rdfs:seeAlso

Non-transparent imprecision: Mappings are rarely exact equivalents.. but we often don't know that

- Precision: exact, narrow, broad, close, related
- **Imprecise is not necessarily bad...**
- **... but non-transparent imprecision is**
- Aggravated when considering "crosswalks"
  - inverse walking (Alzheimer's → Alzheimer's 2):
  - multihop walking (Alzheimer's 2 → Alzheimer's → Alzheimer's 3)
  - hairball problem
- <u>You cannot use bidirectional nor multi-hop crosswalks without knowing the precision</u>

Inaccurate and incomplete: The consequence of relying on tools or mass manual review

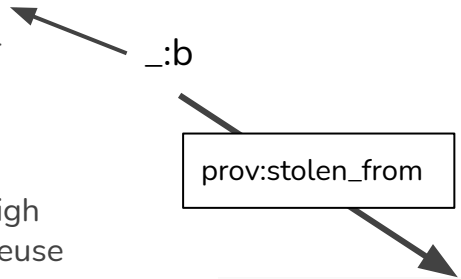
- New tooling make large scale matching more viable - but precision and accuracy will continue to be questionable
- Curators are basically insanely accurate at this - but for 20K mappings?
- The moment a mapping is “complete”, it is basically out of date.
- (Aside: we keep on redoing the same mappings over and over again - which does not help “mapping” convergence and costs so many hours)

An overview of mapping systems in the biomedical domain (excerpt)

- Alignment API and EDOAL: Expressive and Declarative Ontology Alignment Language
- Bioportal Mappings
- OBO hasDbXref (-> OxO)
- Others:
  - SNOMED simple and complex maps, Common Terminology Services 2
  - OpenPHACTS Linksets, BridgeDB Mapping Vocabulary
  - SEMAF: Flexible Semantic Mapping Framework (emerging)
    - SEMAF registries / Mapping Commons
    - Formalising crosswalks great idea
  - Biohackathon 2015: Ontology Mapping Metadata

Mappings are pretty unFAIR

- Findable:
  - Publish {mappings} in dedicated repositories
  - Minimum {mappings} metadata
  - Use unique Ids to identify concepts, relations and {mapping sets}
- Accessible:
  - Use simple APIs to provide access to {ontology} content and metadata (Bioportal, OLS)
- Interoperable:
  - Use common practices at syntactic level
  - Use common structure for {concepts and relations}
- Reusable:
  - Reuse existing related {mappings}
  - **Provide rich metadata to describe {mappings}**
    - without knowing the curation rules, many high precision scenarios would not even permit reuse

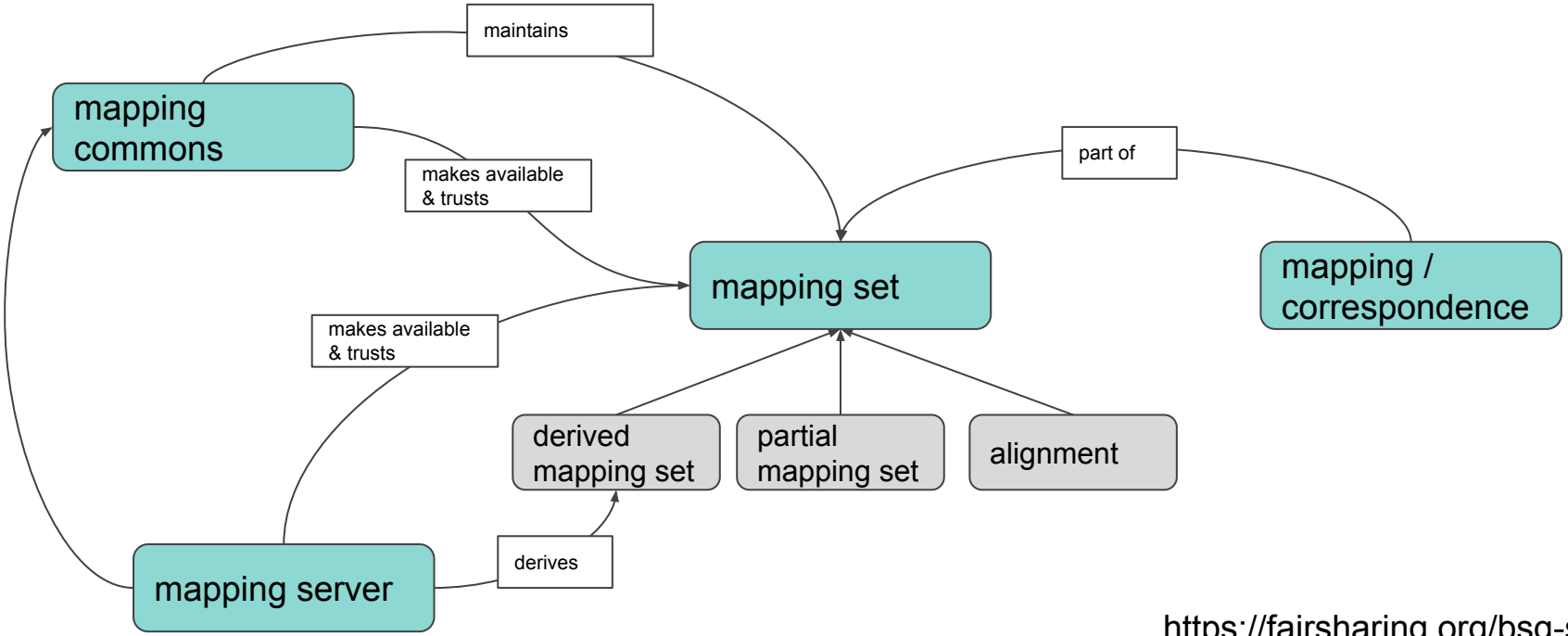


FAIRSFAIR
Fostering Fair Data Practices in Europe

Mappings could be:

- transparently inaccurate
- transparently incomplete
- transparently imprecise
- transparently conflicting
- open
- FAIR and easy to use

Making mappings FAIR and Open: A Simple Standard for Sharing Ontology Mappings (SSSOM)



https://fairsharing.org/bsg-s001618

Mapping model

orcid:0001-.....

confidence

0.9

match_type

HumanCurated

mapping_date

29.05.2021

match_type

HumanCurated

subject_id

EFO:1000070

predicate_id

skos:exactMatch

object_id

MONDO:0006071

match_type

HumanCurated

subject_label

Adenofibroma

subject_match_field

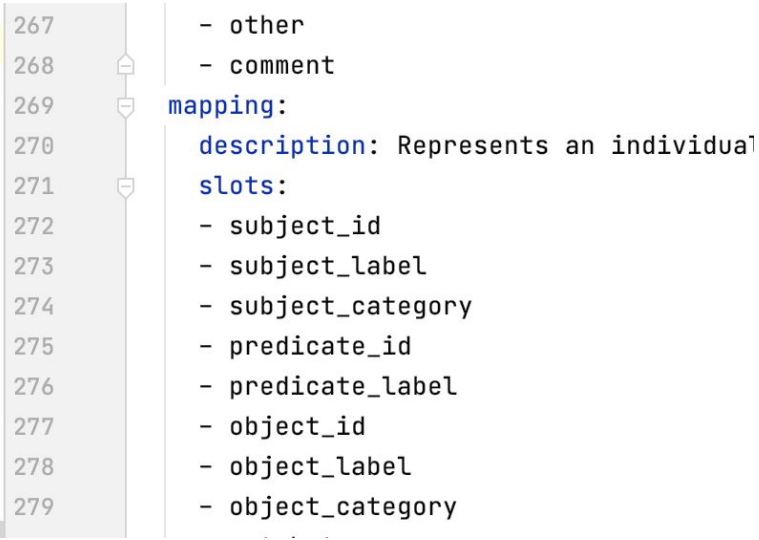
label

object_label

adenofibroma

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

- **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

https://github.com/mapping-commons/sssom-py/tree/master/schema

Example SSSOM tsv file

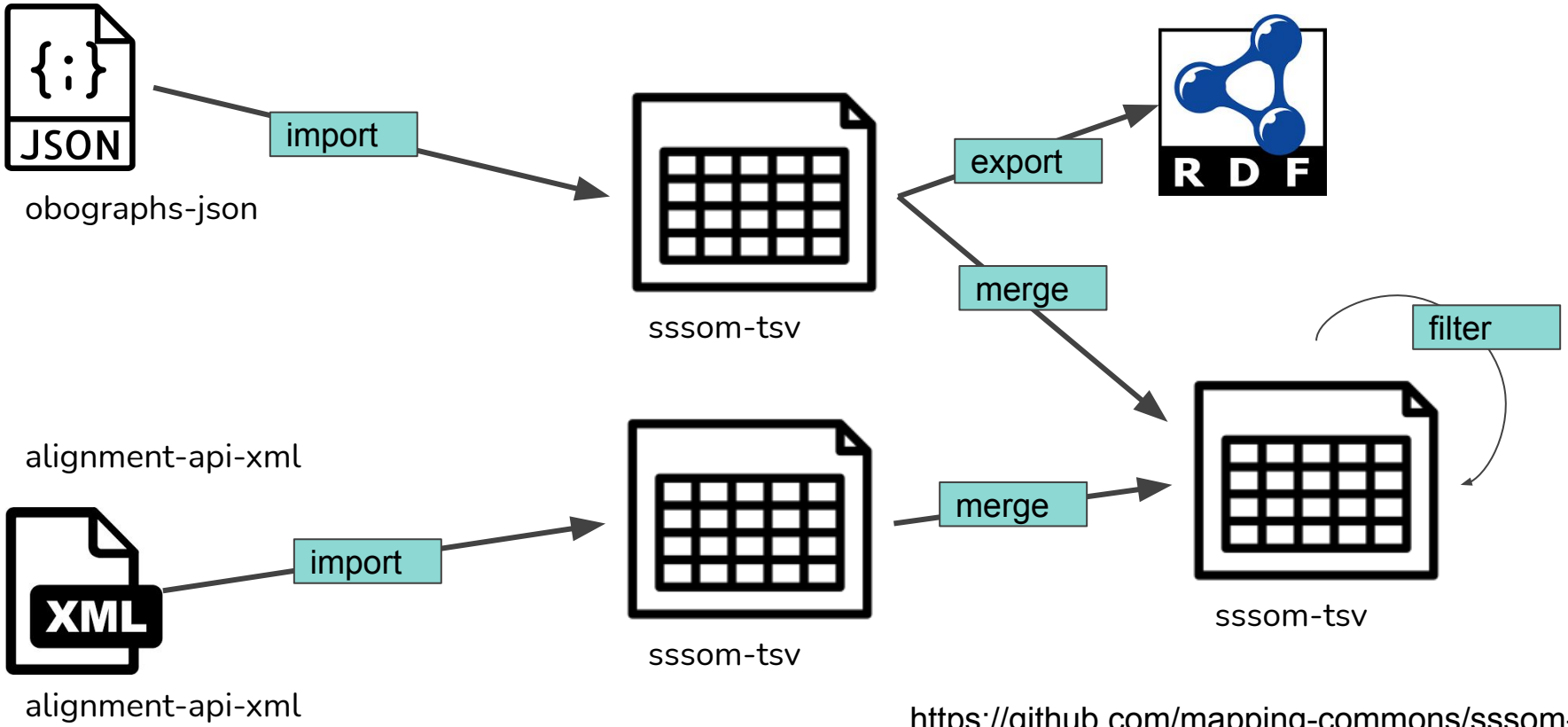
```
#license: "https://creativecommons.org/publicdomain/zero/1.0/"
#mapping_tool: "https://github.com/cmungall/rdf_matcher"
#mapping_date: "2021-07-07"
#curie_map:
#  EFO: "http://www.ebi.ac.uk/efo/EFO_"
#  MONDO: "http://purl.obolibrary.org/obo/MONDO_"
#  dc: "http://purl.org/dc/elements/1.1/"
#  dcterms: "http://purl.org/dc/terms/"
#  oio: "http://www.geneontology.org/formats/oboInOwl#"
#  skos: "http://www.w3.org/2004/02/skos/core#"
subject_id» subject_label»predicate_id» object_id»object_label» match_type» subject_source» object_source»m
MONDO:0008897»hyperphosphatemic familial tumoral calcinosis»skos:exactMatch»EFO:0009383»tumoral calcinosis,
MONDO:0008897»hyperphosphatemic familial tumoral calcinosis»skos:exactMatch»EFO:0009384»tumoral calcinosis,
EFO:1000280»Gastrointestinal Hamartoma» skos:exactMatch»MONDO:0006231»gastrointestinal hamartoma» Lexical»E
EFO:1000280»Gastrointestinal Hamartoma» skos:exactMatch»MONDO:0006231»gastrointestinal hamartoma» Lexical»E
EFO:1000280»Gastrointestinal Hamartoma» skos:exactMatch»MONDO:0006231»gastrointestinal hamartoma» Lexical|S
EFO:1000280»Gastrointestinal Hamartoma» skos:exactMatch»MONDO:0006231»gastrointestinal hamartoma» Lexical|S
EFO:1000280»Gastrointestinal Hamartoma» skos:exactMatch»MONDO:0006231»gastrointestinal hamartoma» Lexical»E
EFO:1000280»Gastrointestinal Hamartoma» skos:exactMatch»MONDO:0006231»gastrointestinal hamartoma» Lexical»E
EFO:1000280»Gastrointestinal Hamartoma» skos:exactMatch»MONDO:0006
EFO:1000280»Gastrointestinal Hamartoma» skos:exactMatch»MONDO:0006
```

Metadata header with curie map, and various mapping set level metadata fields.

Actual mappings with metadata in TSV form

pd.read_csv(f,comment="#", sep="\t")

sssom-py: a python toolkit and CLI to process SSSOM files



https://github.com/mapping-commons/sssom-py

Example uses

- Mouse-Human phenotype Mapping Commons
  - Pistoia Alliance contributed a lot of automated mappings
    - Mapping across disease and phenotype ontologies (merged into FAIR implementation project)
    - Strong focus on automated matching approaches to retain currency
  - IMPC example
  - Similar efforts for other MODs
- The National Microbiome Data Collaborative (NMDC, https://microbiomedata.org/):
  - To provide standardized metadata for data provided by GOLD, map GOLD database fields to MIxS and environmental data elements to ENVO
- Others:
  - CCDH (mapping clinical data models and value sets to ontologies)
  - biomappings
  - omop2obo

A simple guide to make your mappings FAIR.. and open

- Publish mappings:
  a. in a standard file format (tsv, csv, json, xml) - not (only) buried in a database
  b. using a standard open license
  c. referring to a standard metadata schema such as SSSOM or EDOAL
  d. in a public space with an issue tracker (GitHub)
  e. with at least the following metadata:
     - MUST: precision, confidence, match type, prefix map, license
     - SHOULD: curation rules / mapping justification
     - (It is hard to stop here as there are about a dozen or so essential things)
- Nico's personal amendment: Publish mappings
  a. in CC0 or CC-BY (provenance!)
  b. as CSV or TSV to enable straight forward integration into standard data science pipelines
  c. in a designated Mapping Commons which provides exports from your mappings into SSSOM
  d. with rich metadata, such as versions of the term sources, provenance, mapping tools used
  e. Treating mappings with the same love and care as our ontologies and vocabularies

Takeaways

Join us for the next SSSOM user meetup on 3rd September..

Acknowledgements:

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

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

Mapping community: Lots of volunteering contributions, e.g. Charlie Hoyt (sssom-py), Pistoia Alliance (Thomas Liener), John Graybeal (spec), James McLaughlin (infrastructure), ...

Core team:
https://w3id.org/sssom/SSSOM.md

- Striving for **convergence** does not mean we should dismiss **mapping** - the two are **complementary processes** and need to evolve side-by-side
- **Your mappings are valuable, sharing them in a standard format** on a public repository does them and your hard work justice - and saves other people a ton or work (OAEI & manual curation)
- When mapping terms, **please curate at the very least precision, confidence, prefix maps and curation rules**.
- Get involved in SSSOM and influence its development as a community standard