A Simple Standard for Ontological Mappings 2022: Updates of data model and outlook

Nicolas Matentzoglu[1[0000-0002-7356-1779]], Joe Flack[2[0000-0002-2906-7319]], John Graybeal[3[0000-0001-6875-5360]], Nomi L. Harris[4[0000-0001-6315-3707]], Harshad B. Hegde[4[0000-0002-2411-565X]], Charles T. Hoyt[5[0000-0003-4423-4370]], Hyeongsik Kim[6[0000-0002-3002-9838]], Sabrina Toro[7[0000-0002-4142-7153]], Nicole Vasilevsky[7[0000-0001-5208-3432]], Christopher J. Mungall[4[0000-0002-6601-2165]]

[1] Semanticly, Athens, Greece
[2] Johns Hopkins University, Baltimore, MD 21218, USA
[3] Stanford University, Stanford, CA, 94305, USA
[4] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[5] Harvard Medical School, Boston, MA 02115, USA
[6] Robert Bosch LLC
[7] University of Colorado Anschutz Medical Campus, Aurora, CO 80217, USA
cjmungall@lbl.gov

Abstract. The Simple Standard for Ontological Mappings (SSSOM) was first published in December 2021 (v. 0.9). After a number of revisions prompted by community feedback, we have published version 0.10.1 in August 2022. One of the key new features is the use of a controlled vocabulary for mapping-related processes, such as preprocessing steps and matching approaches. In this paper, we give an update on the development of SSSOM since v. 0.9, introduce the Semantic Mapping Vocabulary (SEMAPV) and outline some of our thoughts on the establishment of mapping commons in the future.

Keywords: standards, mappings, ontologies, ontology mapping, FAIR data.

1 Introduction

The problem of mapping between entities in databases and ontologies is ubiquitous - from automatically establishing mapping using ontology matching or entity resolution techniques to applying them in the context of data transformation, ontology merging or knowledge graph integration. We define a mapping as the correspondence of one entity (a record in a database, a class in an ontology), i.e., the subject, to another entity, i.e., the object. A semantic mapping is a mapping that further specifies a predicate describing how the subject maps to the object e.g., is it an “exact match”, or merely “close”? Are the two entities “logically equivalent” in the sense of the Web Ontology Language (OWL)? Despite their importance, semantic mappings are rarely shared widely outside the tool-developing communities. Standards like the EDOAL [1] and the Alignment

API [2] have had a huge impact on the field of ontology alignment, being the reference format for the Ontology Alignment Evaluation Initiative (OAEI) community, but are conceptually focused on ontologies, and have only very limited ability to express detailed metadata on mapping sets (alignments), such as provenance, licensing information and attribution.

The Simple Standard for Ontological Mappings (SSSOM[1]) has been proposed as a community standard for sharing semantic mappings between information entities [3]. A mapping in this context is a statement <s, p, o> that establishes a correspondence of a subject entity (s) to an object entity (o) via a mapping predicate (p). Individual mappings can be grouped into mapping sets. A special kind of mapping set is an "alignment" comprising all mappings between two data spaces (ontologies/databases). SSSOM specifies a schema and data model for mapping sets and mappings, and a rich set of metadata elements to describe them. Publishing mappings in a standard format with appropriate provenance, whether they are automatically generated by ontology matchers, manually curated or both, drastically reduces the burden of re-curating mappings between the same ontologies over and over again, and provides a "point of convergence", standard mapping sets curated and maintained much the same way as ontologies and vocabularies are curated: as FAIR semantic artefacts [4].

In this paper, we will describe the most significant changes to the SSSOM standard since its initial publication in December 2021, including:

1. The move from modelling mapping metadata as "type" information, e.g. this mapping corresponds to a "lexical match", to modelling them as "mapping justifications", e.g. this mapping is supported by "lexical matching".
2. The establishment of a bespoke semantic mapping vocabulary that defines certain key mapping activities, such as types of matching approaches, formally.
3. A first concrete proposal to define mapping registries, collections of mapping sets, as a backbone to defining what we envision as mapping commons - community-driven spaces that curate and reconcile mappings and enrich them with metadata.

2 Updates to the SSSOM standard since December 2021

The following changes have been made since version 0.9 of the SSSOM standard. At the time this paper was written, the version is 0.10.1.

2.1 Key changes to the SSSOM data model

Requiring sources to be recorded as entity references. We now require entity references instead of strings to denote source information for the subjects and objects in the

[1] https://github.com/mapping-commons/sssom

mapping (subject_source, object_source). For example, we previously permitted the string "UBERON" to be used to declare that the subject is part of the Uberon ontology [5], but now require (ideally resolvable) entity references instead, e.g. obo:uberon to denote the Uberon ontology or wikidata:Q465 to denote DBpedia. While this practice by itself does not guarantee that sources are documented uniformly across mapping sets (both obo:uberon and wikidata:Q7876491 refer to the Uberon ontology), it does ensure that users can at least figure out which was the intended source without having to resort to manual web searches. We are currently discussing how to ensure that "sources" can be referred to by a unique, consistent identifier scheme.

Requiring entity references for documenting preprocessing techniques. Previously we allowed preprocessing techniques, for example, applying to lexical matching activities (Stemming, Lemmatisation, etc), to be recorded using natural language strings. To facilitate the standardisation of these references, we now require preprocessing techniques to be recorded using controlled vocabularies. One such vocabulary is the Semantic Mapping Vocabulary (SEMAPV), which we will introduce later in this report.

Splitting match term type into separate type fields for the subject and the object of the mapping. To facilitate mappings between different ontological types (classes and individuals, for example), we split the previous match_term_type property (which used a controlled vocabulary with terms like ConceptMatch, ClassMatch, IndividualMatch) into separate metadata elements for the subject (subject_term_type) and object (object_term_type), using the standard Semantic Web types as values, such as owl:Class, rdfs:Resource or skos:Concept. This ensures, for example, that we can map between elements that are not strictly the same type, such as "skos:Concept", for example "Alzheimers" in a SKOS taxonomy, and "owl:Class", for example "Alzeimhers" in an OWL Ontology. As another positive side effect, we can use this kind of information to make assumptions about the intended semantic framework for the mapping: if, for example, a mapping property is used that is formally an owl:ObjectProperty, and both the subject and the object are owl:Class, then we can export the mapping as an owl:SomeValuesFrom restriction when the user uses the SSSOM toolkit to transform their mapping into OWL.

Modelling mapping metadata as "mapping justifications", rather than "match type" information. In SSSOM v. 0.9, the "match_type" property was used to express that the mapping between the subject entity (subject_id) and the object entity (object_id), using a particular mapping predicate (predicate_id), is of type "lexical match". During the first phase of the design, this felt more natural, as we would often use phrases like "this is a lexical match" when talking about a specific mapping. But it turned out that there were a few problems with this metadata element. Firstly, we realised that there was a confusingly inconsistent use of the "matching" vs "mapping" terminology across the SSSOM specification. For example, the name "SSSOM" suggests that we are talking about a standard for mappings - but one of its central metadata elements ("match_type") uses the term "match" instead. The core team has settled on the

following conventions for using the terms “mapping” and “matching” across the specification: a *term mapping* is a statement <s,p,o> comprising a subject entity (s), an object entity (o) and a mapping predicate (p); this is synonymous to what the ontology matching literature refers to as correspondence. We refer to the process that results in a mapping between a subject and an object entity as “matching”. Obviously, these are merely conventions to ensure consistent communication when talking about SSSOM, and not in any way normative - there are many different valid uses of the terms “mapping”, “match” and “matching” across the literature. Secondly, a single mapping can be lexical, logical and expert-curated at the same time. Stating that the mapping is of “multiple types” is awkward and confusing. Instead, aligning SSSOM a bit more closely with the provenance model of PROV[2] (another request by the community) appeared more natural: describing how the mapping came into being as “activities that generate or confirm a mapping” (“activity” is a term from the PROV data model denoting a process “that occurs over a period of time and acts upon or with entities”). We decided to refer to these processes as “mapping justifications”.

2.2 Tooling related updates

The SSSOM toolkit offers a number of utility methods such as “merge” (to merge two mapping sets), “parse” (to convert a different format, such as EDOAL, into SSSOM) and “validate” (to check that a mapping set is legal SSSOM). While it was always part of the design philosophy of SSSOM to not require any special tooling for reading and writing SSSOM files (i.e. it should be possible to use the normal data science toolbox, such as pandas), it can be convenient to have a special toolkit that covers some of the more frequently used operations of mapping sets on the command line, and provides a convenient API for data pipelines in Python. Since SSSOM 0.9 a number of new features have been added. The “annotate” function allows adding mapping set level metadata, such as license or a version, directly using the command line. This supports, for example, use cases like automated mapping extraction or matching pipelines which automatically assign versions. The “validate” command has been significantly improved and now covers JSON schema validation. Lastly, the “filter” command allows to filter a mapping set based on any of its metadata elements: for example, mapping sets can easily be filtered by predicate id, subject id prefix or mapping provider. Some infrastructure developers have also started making pull requests. For example, a converter of SSSOM to OntoPortal mappings has been provided by the AgroPortal developers; a converter for the FHIR ConceptMap is provided by a FHIR developer (under construction).

[2] https://www.w3.org/TR/prov-o/

The Ontology Access Kit[3] (OAK) implements functionality to do basic lexical matching based on term synonyms, including a system for specifying mapping rules such as: if a label of the subject matches an exact synonym, then we declare the presence of an "skos:exactMatch". Term matches are exported as SSSOM mapping sets, including detailed justifications such as "subject match field" and "match string". OAK furthermore allows retrieving ontology mappings from various endpoints such as OxO or BioPortal and exporting it as SSSOM mapping sets.

The next version of the Ontology Development Kit[4] (ODK) will implement direct support for managing SSSOM mappings alongside ontologies by providing functionality for automatically exporting mappings curated as part of the ontology into SSSOM for easier consumption, copying mappings relevant to the ontology sourced from elsewhere and curating mappings (manually or with matching tools).

2.3 Key changes related to outreach and governance

Contribution guidelines and Code of Conduct. Since April 2022, we have defined contribution guidelines [6], a Code of Conduct [7] and proposed a few guidelines for general governance[5] , such as for joining the core team, voting on changes to the data standard and resolving conflicts between team members.

New SSSOM tutorial. We also developed our first comprehensive tutorial for the curation of mappings [8]. The tutorial is the first in a series for getting familiar with SSSOM as a standard to capture mappings and improve practices for mapping curation in general. We believe that it is key to sensitise curators to the idea of mapping precision (exact, narrow, close, broad) and capture this precision as part of the metadata, in particular their choice of a concrete mapping predicate such as skos:exactMatch. Much of our training efforts focus around teaching curators how to choose appropriate mapping predicates for each use case [9].

SSSOM logo design. After a few rounds of proposals and feedback, we settled on a circular Sankey diagram style solution developed by Julie McMurry that illustrates cross-mappings between resources, similar to what the Ontology Xref Services uses [10].

---

[3] https://github.com/INCATools/ontology-access-kit

[4] https://github.com/INCATools/ontology-development-kit

[5] (https://github.com/mapping-commons/sssom/issues/82)

3 The Semantic Mapping Vocabulary (SEMAPV)

In the first public version of SSSOM, most metadata elements were either defined as enums (hardcoded values that are part of the SSSOM standard itself, see above discussion on mapping preprocessing and justifications) or simple open-ended strings (for example the "source" element). Now, to make certain metadata more easily customisable and extensible (while still being semantically meaningful, with rich metadata), we decided to develop a new controlled vocabulary independently of SSSOM to capture the required terms. The Semantic Mapping Vocabulary (SEMAPV) is in a very early stage of development, covering at the moment primarily the terms required for the use cases of the SSSOM user community, in particular to document mapping justifications (e.g. mapping justifications, widely used preprocessing methods). We are, however, discussing the possibility to include additional mapping relationships such as cross-species mappings for connecting biological databases covering diverse species from Drosophila and zebrafish to Homo Sapiens. These are not covered well by the skos mapping vocabulary (which, alongside OWL is the preferred vocabulary for semantic mappings in SSSOM), unless we give up a lot of precision and model these all as skos:closeMatch or skos:relatedMatch. An early version of SEMAPV can be inspected[6]. SEMAPV is open for community feedback and new term requests.

3.1 Mapping registries and mapping commons: how to manage collections of mapping sets

Traditionally, a lot of focus in and around modelling mappings has been about the individual term mappings and mappings sets (or alignments). Very little work is done to deal with sets of mappings, either in terms of collection, indexing and retrieval (mapping registries), or their reconciliation. Matching tools are typically designed to generate a one-off alignment, which is difficult to reconcile with partial alignments from other sources, such as human curated subsets. The role of human review in general is typically seen in service to the automated approaches either by providing a training set for machine learning based approaches, or a validation set for other more traditional approaches ("my matcher is 70% correct compared to a human reviewed subset"). In reality mapping sets are co-developed by curators, automated matchers and, importantly, "crosswalks" or "mapping chains". The latter exploits the fact that if a subject A is an exactMatch to two objects (B and C) then we can reasonably assume that B and C are also exact matches to each other. Manual curation may often not be feasible due to the scale of the data, but can often be captured by user feedback ("this does not make sense!"). It is our firm belief that all these mapping approaches should be applied in concert, which means that a good model for dealing with collections of mapping sets

[6] https://mapping-commons.github.io/semantic-mapping-vocabulary

must be developed. This model answers questions like how to reconcile conflicting mappings (picking the one with higher confidence?) or mappings whose application leads to nonsensical knowledge graphs and ontology structures (equivalence hairballs, inconsistencies). Many of the latter issues are covered by the ontology merging literature [11–15], but their outputs have yet to be standardised to be useful to a wider audience.

With the growing number of projects developing terminology servers, we have recently started looking into formalising collections of mappings sets as mapping registries. An early version of this is included in the current SSSOM release. Mapping registries allow capturing metadata about mappings sets from the perspective of the registry maintainers. For example, while a specific matching tool will output a confidence value for a mapping, the registry maintainer may trust the tool more or less (curate their own confidence in the mapping set). These confidence values can then be used by reconciliation tools to create "harmonised mappings" (a concept that is not clearly defined that the authors usually intend to mean a "mapping set that does not induce equivalence hairballs" or "lead to logical inconsistencies). An example of a mapping registry instance can be found on GitHub (https://github.com/mapping-commons/mh_mapping_initiative/blob/master/mappings.yml). Ultimately, we hope that mapping registries could form the backbone of mapping commons, social endeavours that seek to collect and harmonise all mappings covering a specific domain. For example, the mouse-human mapping commons (https://github.com/mapping-commons/mh_mapping_initiative) seeks to collect and maintain mappings relevant to the integration of mouse model organism research data with human clinical data.

4 Discussion and Conclusions

The adoption of SSSOM is still in the early stages. It is increasingly clear that precise and well documented terminological mappings are required for many use cases, such as bridging the chasm between the world of clinical terminology, which dominates the domain of clinical data, with the world of biological research (such as genomics) which is dominated by open biomedical ontologies (such as OBO Foundry ontologies), and simply relying on automated tooling won't work for many use cases. Furthermore, it is likely that millions of taxpayer dollars are wasted on recreating the exact same mappings over and over again due to the lack of a FAIR and open mapping culture along the lines of what we already have for biomedical ontologies. What is needed is a holistic approach that integrates (partial) mappings created by different groups for specific use cases with the results from automated matchers (which enable coverage and scalability) and human curation (user feedback, biocuration). To achieve this our community (Ontology Matching) should evolve from a primarily tool-centric view (with a focus on algorithmic precision and fully automated matching) to a more data-centric view (integrated development processes, continuous updates of mapping sets, reuse and sharing, hybrid automated and manual mapping curation). Not only will the publication and curation of mappings in mapping commons improve the user experience; for the Ontology Matching community, manually curated mapping sets can ultimately evolve

from silver to gold standard corpora for evaluation as well. Tools should be retrofitted to export “mapping justifications” alongside their results, documenting preprocessing steps, mapping decisions and more in a way that allows downstream users to accept or reject a mapping based on the justification alone.

References

1. EDOAL: Expressive and declarative ontology alignment language. [cited 12 Aug 2022]. Available: https://moex.gitlabpages.inria.fr/alignapi/edoal.html
2. David J, Euzenat J, Scharffe F, dos Santos CT. The Alignment API 4.0. Semantic Web. 2011. pp. 3–10. doi:10.3233/sw-2011-0028
3. Matentzoglu N, Balhoff JP, Bello SM, Bizon C, Brush M, Callahan TJ, et al. A Simple Standard for Sharing Ontological Mappings (SSSOM). Database . 2022;2022: baac035.
4. Franc YL, Coen G, Essen JP, Bonino L, Lehväslaiho H, Staiger C. D2.2 FAIR semantics: First recommendations. 2020 [cited 12 Aug 2022]. Available: https://www.narcis.nl/publication/RecordID/oai:pure.knaw.nl:publications%2F8e193436-dd29-40e5-8e60-1b6a3cf43e8f
5. Haendel MA, Balhoff JP, Bastian FB, Blackburn DC, Blake JA, Bradford Y, et al. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. J Biomed Semantics. 2014;5: 21.
6. CONTRIBUTING.md at master · mapping-commons/sssom. Github; Available: https://github.com/mapping-commons/sssom
7. CODE_OF_CONDUCT.md at master · mapping-commons/sssom. Github; Available: https://github.com/mapping-commons/sssom
8. Basic tutorial - A simple standard for sharing ontology mappings (SSSOM). [cited 12 Aug 2022]. Available: https://mapping-commons.github.io/sssom/tutorial/
9. How to use mapping predicates - A Simple Standard for Sharing Ontology Mappings (SSSOM). [cited 12 Aug 2022]. Available: https://mapping-commons.github.io/sssom/mapping-predicates/
10. Ontology Xref Service. [No title]. [cited 12 Aug 2022]. Available: https://www.ebi.ac.uk/spot/oxo/
11. Lambrix P, Edberg A. Evaluation of ontology merging tools in bioinformatics. Pac Symp Biocomput. 2003; 589–600.
12. Stumme, Maedche. Ontology merging for federated ontologies on the semantic web. OIS@ IJCAI. Available: https://openreview.net/pdf?id=ryWUgQz_-B
13. Dou D. Ontology Translation by Ontology Merging and Automated Reasoning. Yale University; 2004.
14. Raunich S, Rahm E. Towards a Benchmark for Ontology Merging. On the Move to Meaningful Internet Systems: OTM 2012 Workshops. Springer Berlin Heidelberg; 2012. pp. 124–133.
15. Noy, Musen. Algorithm and tool for automated ontology merging and alignment. Proceedings of the 17th National Conference on. Available: https://www.aaai.org/Papers/AAAI/2000/AAAI00-069.pdf?ref=https://githubhelp.com