A Simple Standard for Ontological Mappings 2023: Updates on data model, collaborations and tooling

Nicolas Matentzoglu[1], Ian Braun[5], Anita R. Caron[10], Damien Goutte-Gattat[2], Benjamin M. Gyori[3], Nomi L. Harris[4], Emily Hartley[5], Harshad B. Hegde[4], Sven Hertling[6], Charles Tapley Hoyt[3], HyeongSik Kim[7], Huanyu Li[8], James McLaughlin[10], Cassia Trojahn[9], Nicole Vasilevsky[5] and Christopher J. Mungall[4]

[1] *Semanticly, Athens, Greece*
[2] *University of Cambridge, Cambridge, CB2 3DY, UK*
[3] *Harvard Medical School, Boston, MA 02115, USA*
[4] *Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[5] *Critical Path Institute, Tucson, AZ 85718, USA*
[6] *Data and Web Science Group, University of Mannheim, Germany*
[7] *Robert Bosch LLC*
[8] *Linköping University, Linköping, Sweden*
[9] *Universite Toulouse 2, Toulouse, France*
[10] *European Bioinformatics Institute (EMBL-EBI), Hinxton, UK*

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 0.15.0 in July 2023. Here we report on the progress made since August 2022, in particular changes to tooling, data model and summary of ongoing standardisation efforts.

Keywords

standards, mappings, ontologies, ontology mapping, FAIR data

1. Introduction

Entity mappings define correspondences between entities in different semantic spaces. A "semantic space" in this context can be anything from an ontology, terminology, database or controlled vocabulary to enumerations in a data model. Entities identify/represent a real-world concept or instance in that space. Many such entities refer to the exact same, or similar, real-world concept or instance.

Our most urgent global problems, including rare disease, climate change and ocean pollution require the integration of data that span many semantic spaces across countries and modalities. Efforts that aim to connect semantic spaces typically pursue two main avenues: the definition of powerful, all-encompassing ontologies and terminologies (vocabularies) that are adopted at a global scale, and entity mapping efforts that connect spaces by translating entities from one vocabulary to another. In practice, we rarely see true convergence to the same vocabulary at a global scale. For example, there are dozens of widely used disease terminologies in the clinical domain. This establishes the need for entity mappings.

Proceedings of the 18th International Workshop on Ontology Matching
Nicolas.matentzogly@gmail.com (N. Matentzoglu); cjmungall@lbl.org (C. J. Mungall)
0000-0002-7356-1779 (NM); 0000-0002-8361-2795 (JML); 0000-0002-6523-4866 (ARC); 0000-0002-6095-8718 (DGG); 0000-0001-9439-5346 (BMG); 0000-0001-6315-3707 (NLH); 0000-0001-5839-2535 (EH); 0000-0002-2389-9288 (IB); 0000-0002-2411-565X (HBH); 0000-0003-0333-5888 (SH); 0000-0003-4423-4370 (CTH); 0000-0002-3002-9838 (HK); 0000-0003-1881-3969 (HL); 0000-0003-2840-005X (CT); 0000-0001-5208-3432 (NV); 0000-0002-6601-2165 (CJM)

The Simple Standard for Sharing Ontological Mappings (SSSOM) was created to enable the sharing of semantic entity mappings [2]. The standard provides a rich set of metadata elements to describe entity mappings, such as mapping "justifications" (processes that generate evidence to support the mapping), "mapping predicate" (the mapping relationship, e.g. skos:exactMatch), "confidence" (the probability or certainty an agent has in the truthfulness of a mapping) and "author" (the agent asserting the mapping). The data model is specified using a LinkML schema (https://linkml.io/), which enables well-specified translations into a variety of serialisations, from the default TSV format to JSON and Turtle. In contrast to existing mapping standards like Alignment API (https://moex.gitlabpages.inria.fr/alignapi/), SSSOM is targeted at use cases that require the documentation of (1) rich metadata about mappings (2) in a simple manner, such as a spreadsheet. Tools that convert other standards (such as Alignment API) into SSSOM exist. Here, we report on the progress made since August 2022 [3], describing updates to the data model, SSSOM related efforts in the community, existing and new emerging tooling, and outline the vision for the coming years.

2. Updates to the SSSOM standard and documentation

Compared to the last few years, only few new metadata elements were added to SSSOM. This suggests that the standard is largely stabilising, and most requirements to the current user base are met. Four new metadata elements were added to the core model: mapping_set_title (a human-readable title for the mapping set), issue_tracker (a link to the issue tracker to be used for reporting problems related to a mapping set), issue_tracker_item (a field to track an ongoing discussion about a specific mapping) and curation_rule. A curation rule is a (potentially) complex condition executed by an agent that led to the establishment of a mapping. Curation rules often involve complex domain-specific considerations, which are hard to capture in an automated fashion. For example, when curation a mapping between two vocabularies about phenotypic abnormalities, we may define a curation rule such as "The two phenotypes inhere in homologous anatomical structures and exhibit the same phenotypic quality (increased amount, length, morphology)", or "Two genes are considered matching if they are orthologous".

We have improved the documentation in various ways, by creating a page listing all SSSOM talks [4] (including slides and recordings), providing a reference for "chaining rules" that can be applied to mappings to infer additional mappings [5], and releasing two new tutorials: a guide for setting up a new Mapping Commons [6] and a guide on how to document mapping justifications [7].

3. Tooling-related updates

The **sssom-py library and CLI tool** for processing (converting, validating, etc.) SSSOM mapping files has been migrated to supporting extended prefix maps [8] (EPM) supplied by the Bioregistry [9] for processing SSSOM files, especially during Extract-Transform-Load (ETL) operations. EPMs are a novel way to manage messy identifiers in data that, in combination with the curies [10] package, enable the standardisation and conversion between RDF/URI-based identifiers and CURIEs, for example when converting SSSOM between different file formats. sssom-py also has a new "invert" command for changing the direction of semantic mappings.

The successor of the European Bioinformatics Institute's Ontology Xref Service [11] (OxO) is developed entirely based on the SSSOM data model. **OxO2** [12] is a Docker-based UI system that is built on top of an entirely independent REST API for the retrieval of SSSOM mappings.

The **Ontology Access Kit** [13] (OAK) has a number of methods that provide mappings in SSSOM format, including the "mappings" command for retrieving mappings, for example from databases or ontologies, and the "lexmatch" command for doing basic lexical matching. Various

improvements have been implemented to provide standard conformant SSSOM, and improve automated matching by supporting a standard system for encoding synonym rules ("synonymiser"). A new command, boomerang, was implemented to analyse the output of boomer [14], a tool that uses a combined logical and probabilistic approach to translate mappings into logical axioms that can be used to merge ontologies. boomerang effectively translates the output of boomer, which is focussed on analysing and de-conflicting mapping cliques, into harmonised SSSOM mapping sets.

The **sssom-java library** aims to enable basic SSSOM processing in Java. An independent implementation of the SSSOM standard is now available for the Java programming language [doi:10.5281/zenodo.8192579]. The SSSOM-Java project provides a Java library to read, write, and manipulate SSSOM mappings, and allows rule-based arbitrary transformations of mappings into any kind of objects. It also provides an experimental ROBOT plugin (to be used with future versions of ROBOT where support for such plugins will be available) to inject SSSOM-derived axioms into an OWL ontology.

The operationalization of semantic mappings in order to support the standardised identification of biomedical entities requires the ability to assemble, reason over, and process semantic mappings at scale, to choose the single "best" reference within a given configurable context, and to be able to assess the confidence of such mappings. The **Semantic Reasoning Assembler** (SeMRA [15,16]) is a software tool that addresses these needs, in part by building on the SSSOM standard and SSSOM-py implementation for standardised input and output of a combination of primary and derived mappings.

Biomappings [17] (https://github.com/biopragmatics/biomappings) is a repository of predicted and community curated semantic mappings not available from primary resources like ontologies. It keeps detailed provenance information about curations of predictions, including predictions that are non-trivially untrue. It now exports all uncurated predictions (curr. 94 mappings), predictions curated as true (curr. 10306 mappings), and predictions curated as false (curr. 1566 mappings) into a combined SSSOM that leverages detailed provenance features and now the predicate_modifier field for representing untrue mappings.

The **Mapping Commons repository template** [18] is a cookiecutter [19] template for creating repositories of SSSOM mapping files. Mapping commons registries are created from the template via cruft (https://cruft.github.io/cruft/), which is also used to keep the repository in-sync with any updates made to the template. The template provides the basic infrastructure for managing SSSOM files, including a Makefile with commands to update the repo and validate the mapping files, and CI workflows for either GitHub or GitLab to automatically run the validation tests. Using a standardised layout for mapping registries will help make the publication of mappings more FAIR and transparent, but also ensure that mappings are published using and validated against the latest metadata model versions. SSSOM itself comes with a lightweight metadata model for defining mapping registries, which furthermore makes it easier to combine multiple registries, for example for the purpose of building a centralised mapping server. An example implementation can be found at https://gitlab.c-path.org/c-pathontology/mapping-commons.

4. Collaborations and Community updates

SSSOM at Biocuration 2023. The 2nd Mapping Commons Workshop on Simple Standard for Sharing Ontology Mappings [20] was held in Padua, Italy. The workshop was about the limits of the applicability of SSSOM, and ways to work around it. The discussion focussed on four important corner cases: literal mappings (mappings between a literal string and an identifier), complex mappings (mappings involving more than two entities), data structure mappings

(mappings that involve translation rules for converting an entity from one data schema to another) and value set mappings (two-level mappings where two sets, such as enums, clinical concept sets, etc are associated with each other on set-level, but also pairwise at element-level). It was generally consensus that capturing data structure mappings was entirely out of scope for SSSOM, with references to alternative solutions such as the LinkML transformer [21]. For literal mappings, a new SSSOM profile is being discussed that reuses most of the metadata elements from the main SSSOM standard, but adds a small number of additional elements to describe the mapped literals. Value set mappings are likely to be natively supportable by the SSSOM standard in the future, requiring only small tweaks to the mapping set level metadata (i.e. specifying the mapped value sets). Complex mappings were the most contentious subject discussed at the workshop. While no universal agreement could be reached by the workshop participants, it is likely that at least one proposal will emerge that involves specifying a complex expression in a URI string (that can then be referenced in a SSSOM file just as any individual identifier).

SSSOM at OAEI 2023. Since 2004, OAEI (Ontology Alignment Evaluation Initiative) organises yearly evaluation campaigns for ontology matching technologies. OAEI provides several test sets in different domains (e.g., anatomy and conference), including ontologies to be matched and reference alignments. SSSOM will be introduced at OAEI in stages. In 2023, the OAEI benchmarking team (MELT) agreed to adding SSSOM as an optional output format. MELT [24] is a framework used by system implementers and track organizers to develop, submit, and evaluate matching systems. We have published a draft guide for implementing SSSOM [22] for matching tool developers, and three matching tool developers have agreed to implement the standard for this year's OAEI campaign.

SSSOM and FAIR IMPACT. The SSSOM developer community has established a fruitful collaboration with members of the FAIR IMPACT project. We work together on promoting the publication of FAIR mappings using the SSSOM standard in the wider life science communities and co-organised a workshop to that end [1].

SSSOM at OHDSI 2023 European Symposium. The SSSOM developer community has established a collaboration with the OHDSI program. OHDSI maintains the widely used Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), an open community data standard for representing observational medical data. The goal of the collaboration is to improve the metadata collected for terminological mappings, both for OHDSI managed mappings and community contributions. So far, key SSSOM elements such as confidence, predicate_id, mapping_source, mapping_justification and mapping_tool have been adopted for the OMOP Community Contribution template [23]. At the symposium, we also presented a collaborative work on flowsheet mappings.

Emerging mapping registries for managing collections of mapping sets. The Critical Path Institute (C-Path) Data Collaboration Center (https://c-path.org/programs/dcc/) translated a subset of the mappings available through the OMOP2OBO [24] project into SSSOM-conformant mappings and metadata. This work entailed creating a simple translation between the provided metadata elements and standard SSSOM metadata elements (e.g. "Mapping Evidence Component'' to sssom:mapping_justification), and from the provided metadata values to allowable terms from the ontologies supported by the SSSOM standard (e.g., "Hand Mapping'' to semapv:ManualMappingCuration). This translation and the resulting mappings were managed within the C-Path Mapping Commons repository, providing access to the SSSOM Toolkit functionality for validating the mappings and exporting them in the supported file formats. This process enabled loading the resulting mappings into a graph database, as well as providing interoperability with additional mappings created by C-Path by conforming to the SSSOM standard.

5. Discussion and Conclusions

Compared to the last update a year ago, changes to the SSSOM standard are getting fewer, which means that the metadata model is finally stabilising. Our intention is to launch version 1.0 by the year's end. Efforts have shifted notably to community engagement and collaboration, and expansion of SSSOM related tooling. The main aspiration of the data and terminology mapping community should be to share well defined semantic mappings in completely open Mapping Commons, and support domain experts to curate better, semantically meaningful mappings.

Acknowledgements

NM was supported by NIH National Human Genome Research Institute Phenomics First Resource, NIH-NHGRI # 5RM1 HG010860, a Center of Excellence in Genomic Science, and a kind gift from Bosch LLC to LBNL; Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy [DE-AC0205CH11231 to N.L.H., H.B.H. and C.J.M.]; CTH and BMG were funded under the Defense Advanced Research Projects Agency (DARPA) Young Faculty Award [W911NF-20-1-0255] and the DARPA Automating Scientific Knowledge Extraction and Modeling program [HR00112220036]; DGG was supported by grant BB/T014008/1 from the UK Biotechnology and Biological Sciences Research Council (BBSRC) and the US National Science Foundation Directorate of Biological Sciences (NSF/BIO). The Critical Path Institute is supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) (54.2%) and non-government source(s) (45.8%). The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS or the U.S. Government.

References

[1] Why Mappings Matter and how to make them FAIR? [cited 30 Jul 2023]. Available: https://www.fair-impact.eu/events/fairimpact-events/why-mappings-matter-and-how-make-them-fair

[2] 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. doi:10.1093/database/baac035

[3] Matentzoglu N, Flack J, Graybeal J, Harris NL, Hegde HB, Hoyt CT, et al. A Simple Standard for Ontological Mappings 2022: Updates of data model and outlook. 2022. doi:10.5281/zenodo.7672104

[4] Presentations. [cited 30 Jul 2023]. Available: https://mapping-commons.github.io/sssom/presentations/

[5] Overview of chaining rules. [cited 30 Jul 2023]. Available: https://mapping-commons.github.io/sssom/chaining_rules/

[6] Set up a mapping registry/commons - A Simple Standard for Sharing Ontology Mappings (SSSOM). [cited 30 Jul 2023]. Available: https://mapping-commons.github.io/sssom/mapping-commons/

[7] Mapping justifications - A simple standard for sharing ontology mappings (SSSOM). [cited 30 Jul 2023]. Available: https://mapping-commons.github.io/sssom/mapping-justifications/

[8] Data Structures — curies documentation. [cited 31 Jul 2023]. Available: https://curies.readthedocs.io/en/latest/struct.html

[9] Hoyt CT, Balk M, Callahan TJ, Domingo-Fernández D, Haendel MA, Hegde HB, et al. Unifying the identification of biomedical entities with the Bioregistry. Sci Data. 2022;9: 714.

[10] Hoyt CT. curies: Idiomatic conversion between URIs and compact URIs (CURIEs). Github; Available: https://github.com/cthoyt/curies

[11] Ontology Crossref Service (OxO). [cited 29 Jul 2023]. Available: https://www.ebi.ac.uk/spot/oxo/
[12] Oxo2: UI for OxO 2. [cited 29 Jul 2023]. Available: https://github.com/EBISPOT/oxo2
[13] Ontology Access Kit: A python library and command line application for working with ontologies. [cited 29 Jul 2023]. Available: https://github.com/INCATools/ontology-access-kit
[14] boomer: Bayesian OWL ontology merging. Github; Available: https://github.com/INCATools/boomer
[15] semra: Semantic Mapping Reasoning Assembler (SeMRA): tooling for semantic mappings. [cited 29 Jul 2023]. Available: https://github.com/biopragmatics/semra
[16] Hoyt CT. biopragmatics/semra: v0.0.2-alpha. 2023. doi:10.5281/zenodo.8192829
[17] Hoyt CT, Hoyt AL, Gyori BM. Prediction and curation of missing biomedical identifier mappings with Biomappings. Bioinformatics. 2023;39. doi:10.1093/bioinformatics/btad130
[18] Mapping Commons Cookiecutter. [cited 29 Jul 2023]. Available: https://github.com/mapping-commons/mapping-commons-cookiecutter
[19] Cookiecutter. In: PyPI [Internet]. [cited 29 Jul 2023]. Available: https://pypi.org/project/cookiecutter/
[20] 2nd mapping commons workshop on simple standard for sharing ontology mappings (SSSOM). [cited 30 Jul 2023]. Available: https://mapping-commons.github.io/sssom/events/mc2023/
[21] linkml-transformer: ALPHA data model mapping with linkml. [cited 30 Jul 2023]. Available: https://github.com/linkml/linkml-transformer
[22] Matching tool implementation guide - A Simple Standard for Sharing Ontology Mappings (SSSOM). [cited 30 Jul 2023]. Available: https://mapping-commons.github.io/sssom/matching-tool-implementation-guide/
[23] OHDSI Vocabulary: Community contribution. [cited 30 Jul 2023]. Available: https://github.com/OHDSI/Vocabulary-v5.0/wiki/Community-contribution
[24] Callahan TJ, Stefanski AL, Wyrwa JM, Zeng C, Ostropolets A, Banda JM, et al. Ontologizing health systems data at scale: making translational discovery a reality. NPJ Digit Med. 2023;6: 89.
[25] Hertling, Sven, Jan Portisch, and Heiko Paulheim. "Melt-matching evaluation toolkit." International conference on semantic systems. Cham: Springer International Publishing, 2019.