Ontology Mapping for Life Science Data

Amrapali Zaveri and Michel Dumontier

Stanford Center for Biomedical Informatics Research, Department of Medicine,
Stanford University, United States,
amrapali|michel.dumontier@stanford.edu

Abstract. Bio2RDF is an open-source project that offers a large and connected knowledge graph of Life Science Linked Data. Each dataset is expressed using its own vocabulary, thereby hindering integration, search, query, and browse data across similar or identical types of data. With growth and content changes in source data, a manual approach to maintain mappings has proven untenable. The aim of this work is to develop a (semi)automated procedure to generate high quality mappings between Bio2RDF and SIO using BioPortal ontologies. Our preliminary results demonstrate that our approach is promising in that it can find new mappings using a transitive closure between ontology mappings. Further development of the methodology coupled with improvements in the ontology will offer a better-integrated view of the Life Science Linked Data.

Keywords: biomedical data, life science, ontology, integration, semantic web, linked data, RDF, Bio2RDF, SIO

1 Life Science Data Integration

The life sciences have a long history of producing open data. Unfortunately, these data are represented using a wide variety of incompatible formats (e.g. CSV, XML, custom flat files etc.). Linked Data (LD) offers a new paradigm of using community standards to represent and provide a data in a uniform and semantic manner [5]. Bio2RDF is an open-source project that provides Linked Data across 35 life science datasets [1][2]. Although Bio2RDF effectively offers a large and connected knowledge graph, each dataset is expressed using its own vocabulary, thereby hindering search, query, and browse data across similar or identical types of data. In previous work, we explored the use of mappings between Bio2RDF types and relations to the Semanticscience Integrated Ontology (SIO[1]) [4], an integrated upper level ontology (types, relations) for consistent knowledge representation across domains. Our work made it possible to query across the network of linked life science data using a single ontology. However, with growth and content changes in source data, a manual approach to maintain such mappings has proven untenable.

[1] Available at https://bioportal.bioontology.org/ontologies/SIO

The aim of this work is to develop a (semi)automated procedure to generate high quality mappings between Bio2RDF and SIO. Specifically, we infer Bio2RDF-SIO mappings by mapping Bio2RDF and SIO classes to biomedical ontologies contained in the NCBO BioPortal [11], and consequently use their hierarchies to find indirect Bio2RDF type to SIO class mappings. We evaluate our approach with 319 Bio2RDF classes to be mapped with 1500 SIO classes and 475 BioPortal ontologies.

2 Related Work

Numerous methods have been developed for ontology matching [9], including string-based distance metrics [3], graph-based [6] and instance-based [8] matching. BioPortal [11], a repository of biomedical ontologies, also offers limited quality mappings generated from exact literal matches, but does have a facility to enable users to store and share their mappings [10]. However, developing high quality mappings is time-consuming and few open source tools are available to perform this in an effective and iterative manner.

We have previously reported the generation and use of manually curated mappings between Bio2RDF and SIO [2]. This mapping was performed for 19 datasets and resulted in mappings for 136 classes and 407 object properties[2]. Mappings were formalized using `rdfs:subClassOf`, `owl:equivalentProperty` and `owl:superProperty` relations as appropriate. However, while that work demonstrated the utility of such mappings in query answering across 19 Bio2RDF datasets, it falls short of a reproducible methodology for the now 35 datasets that are currently part of Bio2RDF. Thus, the aim of this work is to develop a semi-automated approach to map Bio2RDF classes to SIO classes.

3 Methodology

Bio2RDF. (June 2013) contains 11 billion triples from 35 data sets across a wide variety of biomedical data types such as genes, drugs, and clinical trials. Each converted dataset is semantically described using its own dataset-specific types and relations. There are 6093 classes across all the datasets.

Semanticscience Integrated Ontology. SIO is an upper and mid-level ontology that covers essential types (objects, processes, attributes) and relations for the rich description of arbitrary (real, hypothesized, virtual, fictional) objects, processes and their attributes. Version 1.31.0 contains 1500 classes and 208 properties.

Mapping Process. Figure 1 provides an overview of the methodology we used for the mapping process. We first pruned the set of 6093 Bio2RDF classes to 319

[2] Available at `http://github.com/bio2rdf/bio2rdf-mapping/tree/master/2/`

classes by removing blank nodes, general resources (e.g. biogrid:Resource*[3]), OWL vocabulary (e.g. owl:Class), and other ontologies such as FOAF[4]. We then used the NCBO Annotator [7][5] to find direct mappings via exact lexical matches from the dct:title of each of the 319 classes (Step 1). We used the parameters to set the ontology to be SIO, 'longest_only=true', which meant that only the longest match for a given phrase will be returned and 'max_level = 3' to specify the depth of the hierarchy to use when performing an annotation. We also removed the 'longest_only=true' parameter to find partial matches to the ontology terms.

Additional processing was performed to augment the number of matches including removing the underscore (e.g. Change_of_expression_level), hyphen (e.g. Affected-organism), separating the camel case, and removing dataset specific type declarations e.g. KEGG, MGI, SGD etc.

We used the NCBO Recommender[6] to retrieve a ranked list of mappings to classes from the BioPortal ontologies for those Bio2RDF classes for which direct mappings were not retrieved (Step 2) since the recommender enables partial matches. We used the parameter input_type=1 to indicate that the input was a list of comma separated keywords and output_type=1 to indicate that the output should be a ranked list of individual ontologies.

We then used LogMap [6], an ontology matching tool which enables large-scale ontology matching, to map SIO classes to the BioPortal ontologies (Step 3). We sought to map Bio2RDF classes to SIO classes through the Bio2RDF-BioPortal and SIO-BioPortal mappings. To do so, we traversed the ancestors of the mapped BioPortal class to the first super class that is mapped to a SIO class (Step 4). In this way, the Bio2RDF type becomes a candidate subclass of the SIO class. Thereafter, we evaluated the mappings manually.

4 Preliminary Results

In Step 1, 174 of the 319 Bio2RDF classes were mapped to SIO using the NCBO Annotator. Examples of correct direct mappings include clinicaltrials:Organization with sio:000012 (organization). We also obtained partial phrase matches including sgd:GlycineCount with sio:000794 (count). While this mapping is correct, it is only by virtue that there is no Glycine class in SIO that would generate a type mis-match.

In Step 2, the remaining 145 classes were mapped to one or more of the BioPortal ontologies using the NCBO Recommender. Matches for 94 classes were obtained, leaving 50 classes with no matches. Examples of classes not matched to SIO or any of the BioPortal ontologies include drugbank:Biotech, sgd:CodonBias, clinicaltrials:Treatment_Comparison.

[3] Note: All prefixes are abbreviated from http://bio2rdf.org/(dataset)_vocabulary:(Class Name) to (dataset):(Class Name).

[4] http://xmlns.com/foaf/spec/

[5] https://bioportal.bioontology.org/annotator, last accessed 19 July, 2016.

[6] https://bioportal.bioontology.org/recommender, last accessed 19 July, 2016.

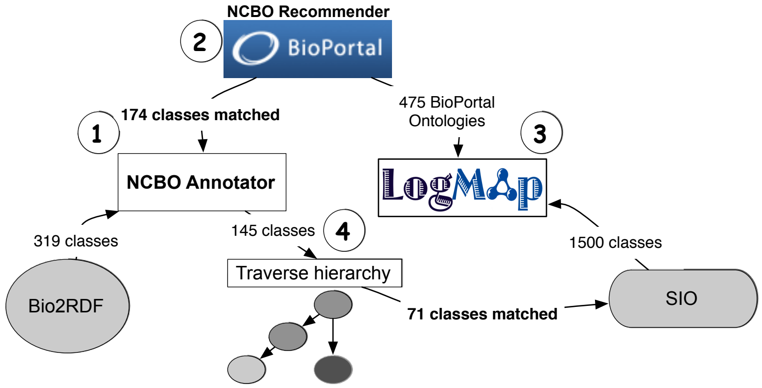

Fig. 1. Overview of the methodology used for mapping Bio2RDF classes with SIO.

In Step 3, we used LogMap to generate mappings between SIO and 393 ontologies (of a total of 475 ontologies). Ontologies such as CPO, FAO had no mappings, while others (e.g. GAZ, COGPO) were inconsistent and could not be used by LogMap.

In Step 4, we analyzed the 94 mappings to BioPortal ontologies to find a super class that could be matched to SIO. Of these, 71 classes were mapped to SIO. For example, `clinicaltrials:Clincial-Study` was mapped to `edda:clinical_trial`[7]. The parent class in that ontology is 'Study Design', which was mapped to `sio:001041` (study design). Other mappings resulted in a matching of top level SIO category. For instance, `clinicaltrials:Baseline` was mapped to the NCI Thesaurus concept `EVS:C25213` (Baseline). The NCI class has a direct super class of 'Conceptual Entity', which maps to the SIO class `sio:000000` (entity). Little can be done in this case because the NCI hierarchy is shallow. In other cases, the mapping was close, but semantically imprecise. For instance, `clinicaltrials:Sham_Comparator` mapped to `co-oed:Sham_comparator`[8]. By traversing the CO-ODE ontology hierarchy, the parent class 'Clinical Trial Study' maps to `sio:001000` (clinical trial). Since a sham comparator refers to an arm of a clinical trial where the patients are not given the intervention, the sham comparator is a part of a clinical trial rather that a type of clinical trial. Therefore, the mapping is not precisely that of subsumption, but formally related as a part, or perhaps as conceptually narrower than a trial.

Some Bio2RDF classes mapped to more than one SIO class. 64 Bio2RDF classes mapped to SIO's base type of *entity*, 21 to *attribute*, 7 to *process* and 27 to other classes such as *role, drug, gene, disease, clinical trial, study, product* etc. For instance, `sider:Drug-Indication-Association` mapped to three of the SIO classes `sio:010038` (drug) and `sio:010299` (disease) and `sio:000897` (association). There were matches that were semantically incorrect such as the

[7] edda prefix for `http://ontologies.dbmi.pitt.edu/edda/StudyDesigns.owl`.

[8] co-ode prefix for `http://www.co-ode.org/ontologies/ont.owl#`.

Bio2RDF class `sgd:Complex` was mapped to the SIO class`sio:010035` (state). This was because the recommended class was `sweet:Complex`[9], the parent of which is 'State', which in turn matched to `sio:000662` (state). However, this SIO class 'State' refers to the geographical boundary of a place, whereas the class 'Complex' refers to the particular condition that something is in at a particular time. From the 94 mappings, 23 could not be matched with SIO. These include `drugbank:Food-interaction`, `clinicaltrials:Analysis`, and `drugbank:Clearance`.

In total, 245 of the 319 classes of Bio2RDF were matched to one or more of the SIO classes. Table 1 lists examples of exact, partial and incorrect mappings. All the input data, NCBO Annotator and Recommender code, BioPortal ontologies, LogMap mapping results, Bio2RDF to SIO mapping results are available at `https://goo.gl/eiijmQ`.

Table 1. Examples of exact, partial and semantically incorrect mappings between the Bio2RDF classes and SIO.

Bio2RDF Class	**SIO Class**	**Annotation**
`clinicaltrials:Organization`	`sio:000012` (organization)	exact
`drugbank:toxicity`	`sio:001008` (toxicity)	exact
`clinicaltrials:Category`	`sio:000137` (category)	exact
`sgd:GlycineCount`	`sio:000794` (count)	partial
`wormbase:Genetic-Interaction`	`sio:010035` (gene)	partial
`pharmgkb:disease-variantlocation-Association`	`sio:010299` (disease)	partial
`sgd:Complex`	`sio:000662` (state)	incorrect
`clinicaltrials:Serious-Event`	`sio:000614` (attribute)	incorrect
`drugbank:Source`	`sio:000510` (model)	incorrect

5 Conclusions, Limitations and Future Work

In this paper, we investigated a methodology to map Bio2RDF classes to SIO via BioPortal ontologies. This work is significant because the lack of semantic integration of the Life Science Linked Data makes it more difficult to find and retrieve data of interest across different datasets.

We found that while SIO has by far the largest number of matching concepts over other biomedical ontologies in BioPortal, neither SIO nor the other ontologies have complete coverage. Our preliminary results demonstrate that our approach is promising in that it can find new mappings using a transitive closure between ontology mappings, but that substantial problems remain concerning mapping to i) semantically incompatible classes, overly general concepts, iii) altogether incorrect classes, and iv) over 50 unmatched classes.

Future work will focus on developing a semi-automated methodology in which discovered mappings can be quickly validated by experts, and potentially by

[9] sweet prefix for `http://sweet.jpl.nasa.gov/2.3/stateSystem.owl#`

non-expert workers from crowdsourcing environments. We will also investigate approaches that extend the mid-level portion of SIO to eliminate root level mappings. Finally, we will extend SIO to include those classes for which mappings currently do not exist or are ultimately not found.

Ultimately, we plan to integrate a revised SIO and vetted mappings to Bio2RDF SPARQL endpoints, where they can serve as a mechanism to improve nascent semantic search and query answering capabilities over the data provided by Bio2RDF.

References

1. Francois Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706 – 716, 2008.
2. Alison Callahan, José Cruz-Toledo, and Michel Dumontier. Ontology-based querying with bio2rdf's linked open data. *Journal of Biomedical Semantics*, 4(1):1 – 13, 2013.
3. W. W. Cohen, P. D. Ravikumar, and Fienberg S. E. A Comparison of string distance metrics for name-matching tasks. *IIWeb*, pages 73 – 78, 2003.
4. Michel Dumontier et. al. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*, 5(1):1 – 11, 2014.
5. Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space (1st edition)*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2011.
6. Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. *LogMap: Logic-Based and Scalable Ontology Matching*, pages 273 – 288. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
7. Clement Jonquet, Nigam H. Shah, Cherie H. Youn, Mark A. Musen, Chris Callendar, and Margaret-Anne Storey. NCBO Annotator: Semantic Annotation of Biomedical Data. In *8th International Semantic Web Conference (ISWC 2009) Posters and Demonstrations*, pages 706 – 716, 2009.
8. V. Loia, G. Fenza, C. De Maio, and S. Salerno. Hybrid methodologies to foster ontology-based knowledge management platform. In *Intelligent Agent (IA), 2013 IEEE Symposium on*, pages 36 – 43, April 2013.
9. Alma Gomez-Rodriguez Lorena Otero-Cerdeira, Francisco J. Rodriguez-Martinez. Ontology matching: A literature review. *Expert Systems with Applications*, 42:949 – 971, 2015.
10. Natalya F. Noy, Nicholas Griffith, and Mark A. Musen. *Collecting Community-Based Mappings in an Ontology Repository*, pages 371 – 386. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
11. P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39:541 – 545, 2011.