The Semantic Data Dictionary Approach to Data Annotation & Integration

Sabbir M. Rashid[1], Katherine Chastain[1], Jeanette A. Stingone[2], Deborah L. McGuinness[1], and James P. McCusker[1]

[1] Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[2] Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

Abstract. A standard approach to describing datasets is through the use of data dictionaries: tables which contain information about the content, description, and format of each data variable. While this approach is helpful for a human readability, it is difficult for a machine to understand the meaning behind the data. Consequently, tasks involving the combination of data from multiple sources, such as data integration or schema merging, are not easily automated. In response, we present the Semantic Data Dictionary (SDD) specification, which allows for extension and integration of data from multiple domains using a common metadata standard. We have developed a structure based on the Semanticscience Integrated Ontology's (SIO) high-level, domain-agnostic conceptualization of scientific data, which is then annotated with more specific terminology from domain-relevant ontologies. The SDD format will make the specification, curation and search of data much easier than direct search of data dictionaries through terminology alignment, but also through the use of "compositional" classes for column descriptions, rather than needing a 1:1 mapping from column to class.

1 Introduction

A common challenge in scientific research involves finding data across databases with the same semantic meaning. This challenge arises since the labels of columns in data tables do not necessary reveal the meaning of the data. Furthermore, one to one mappings between columns from separate sources are not readily accessible. Column headers and traditional data dictionaries describe the conceptual structure underlying a dataset in a manner understandable by human readers, but it is difficult for computers to extract this same information. A single row in a dataset may contain data on multiple entities - for example, the subject, the subject's blood sample, and information about the subject's mother, such as whether or not she smoked during pregnancy. Understanding that these are separate but related entities and how they are related to each other facilitates finding other data that are relevant for comparison.

The Semantic Data Dictionary (SDD) specification is a way to represent implicit entities and their relationships using a general ontology, namely the Semanticscience Integrated Ontology (SIO). SIO provides general properties to

describe the relations between entities, and measured characteristics are represented as attributes of those entities [4]. Domain-specific ontologies, such as the Children's Health Exposure Analysis Resource (CHEAR) ontology [11], allow more fine-grained and dataset-specific annotation of concepts. A well-formed SDD contains information about the entity types represented and/or referred to by each column in a tabular dataset, utilizing the relevant ontology URIs in order to convey this information in a manner that is both machine-readable and unambiguous.

We use SIO's high-level conceptualization of data as our target semantic structure when constructing the SDD. Leveraging one particular structure as a basis helps to focus a user by providing a limited subset of relationships and entities for the user to consider. The SDD can express data against any SIO-compatible ontology, and can be used to describe tabular data where there are any number of entities, attributes, timepoints, roles, and relationships. Our intent is to create a process that is more accessible to domain scientists and data providers as it only requires knowledge of a limited number of ontologies. Properties in the SIO ontology can be used to describe characteristics of a data variable.

In this paper we demonstrate the utility of the SDD format and the use of the SIO and CHEAR ontologies by representing a number of relevant tables. We present an evaluation of the SDD approach by creating SDD specifications for the National Health and Nutrition Examination Survey (NHANES) 2013-2014 dataset. Finally, we describe the use of Semantic Data Dictionaries in existing research projects.

2 Related Work

2.1 Data Integration

Data integration involves the ability to unite data from multiple sources in such a way that results in a unified view of the combined data [8]. An increasingly used approach to data integration is the use of ontologies to annotate data. However, the success of this approach has lead to an increase in the number of existing ontologies, resulting in difficulties in deciding which ontologies to use and the consideration of possible interoperability issues between ontologies. For the biomedical domain, the Open Biomedical Ontologies (OBO) consortium is helping to address this problem by creating a family of logically well-formed ontologies which follow a set of shared design principles [16]. Ontologies contained in the OBO Foundry include the Gene Ontology (GO), Chemical Entities of Biological Interest (ChEBI) and Human Disease Ontology (DOID)[3]. Another important ontology used in biomedical research and science in general is the Semanticscience Integrated Ontology (SIO), which divides entities into three distinct categories: objects, events and processes [5]. Adhering to a common foundational model such as SIO or the OBO Foundry ontologies facilitates data integration.

[3] `http://www.obofoundry.org/`

2.2 Schema Merging

One approach to integrated information from multiple datasets is through Schema Merging. General methods for Schema Merging have involved either using a set of tools to alter multiple schemas such that the are consistent with each other, or using the multiple schemas to create one merged schema [1]. In order to successfully implement these approaches, however, it is important to know which corresponding elements in each schema should be aligned. Further considerations include possible union or intersection of schema elements, generalization of attributes described in a schema, and the removal of redundant attributes or relationships [10]. Aside from algorithmic approaches to Schema Merging or Ontology Alignment, other methods my take advantage of crowd sourcing in order to acquire human contributions. An example of such a platform is CrowdMap [15], which reduces complex alignment problems into individual alignment tasks, which are published online to be outsourced to a distributed group of contributors.

2.3 Semantic Annotation

Semantic Annotation refers to the practice of assigning metadata descriptions that describe information about entities in a database or in text [7]. Recent surveys on Semantic Annotation platforms describe architecture, methods, and performance on currently available tools that facilitate semantic annotations [13], [18]. Several of the most effective annotation platforms (in terms of F-Measure) include MUSE [9], Armadillo [3] and KIM [12]. MUSE is a information extraction system that performs named entity recognition using a tokeniser, sentence splitter, part of speech tagger, and a semantic tagger [9]. Armadillo is a generic and portable architecture for scraping information for websites [3]. KIM is a platform for semantic annotation, indexing, and retrieval that includes the use of an ontology, a server, and a front-end interface [12]. Using an algorithm called Taxonomy-Based Disambiguation, which involves Spotting, Learning and Tagging, SemTag was able to achieve automated large-scale semantic tagging of over 250 million web pages [2]. OntoAnnotate leverages existing conceptualizations from domain specific ontologies, but relies primarily on human annotation [17].

3 Methods

3.1 SDD Specification

The Semantic Data Dictionary is made up of a collection of tabular data which can be written in Excel or Google sheets, or tabular text format, such as Comma Separated Value (CSV) files. The first of these files is the infosheet, which contains information about the study as well as the location of the other tables. The tables referenced in the infosheet are the Semantic Data Dictionary, Codebook, Timeline and Code Mappings. The Semantic Data Dictionary contains columns following the SDD specification, which is shown in Table 1. The SDD contain

Table 1. Semantic Data Dictionary Specification

Column	Value	Type	Related Property	Description
Column	ID	all		Column header
Label	string	all	rdfs:label	Label for the column
Comment	string	all	rdfs:comment	Comment for the column
Definition	string	all	skos:definition	Text column definition
Attribute	URI	attribute	rdf:type	URI of the attribute type
attributeOf	ID	entity	sio:isAttributeOf	Entity having the attribute
Unit	ID	attribute	sio:hasUnit	Unit of Measure for attribute
Time	ID	attribute	sio:measuredAt	Time point attribute was measured
Entity	URI	entity	rdf:type	Type of the entity
Role	URI	entity	sio:hasRole	Type of the role the entity plays
inRelationTo	ID	entity	sio:inRelationTo	Entity that the role is linked to
wasDerivedFrom	ID	entity	prov:wasDerivedFrom	Entity from which the attribute was derived
wasGeneratedBy	ID	all	prov:wasGeneratedBy	Activity from which the attribute was produced

Table 2. Example Semantic Data Dictionary (Actual Columns)

Column	Attribute	attributeOf	Unit	Time	inRelationTo	wasDerivedFrom	wasGeneratedBy
id	sio:Identifier	??child					
race	sio:Race	??mother					
age	sio:Age	??mother	sio:Year	??visit1			
edu	chear:EducationLevel	??mother		??visit1			
bmi	chear:BMI	??mother	kgm2	??visit1		weight, height	
weight	sio:Mass	??mother	kg	??visit1			
height	sio:Height	??mother	cm	??visit1			
smoker	chear:SmokingStatus	??mother		??pregn			
pb_1	sio:Concentration	??pb_1	mgL	??visit1	??sample1	??sample1	hasco:ICP-MS
pb_2	sio:Concentration	??pb_2	mgL	??visit2	??sample2	??sample2	hasco:ICP-MS
ga	chear:GestationalAge	??child	sio:Week	??birth			
birthwt	chear:Weight	??child	kg	??birth			

actual columns derived from the dataset, as well as virtual columns. The actual columns contain mappings to the underlying attribute that is described by the dataset column as well as provenance information such as how that variable was generated or derived, as shown in Table 2. In order to describe the entity to which the attribute is describing or the time of measurement, virtual columns are used. One benefit of using virtual columns is that they allow for inclusion of mapping to concepts that are implicit to the data, such as the entity that an attribute belongs to. An example of virtual columns is shown in Table 3. Virtual columns involving time intervals should be stored in the Timeline table. Like standard codebooks used by the biomedical community, the Codebook table contains possible values of coded variables and their associated labels. We augment each possible value with mappings to corresponding ontological concepts, as shown in Table 4. Finally, the Code Mappings table contains mappings

Table 3. Example Data Semantic Data Dictionary (Virtual Columns)

Column	Entity	Role	Relation	inRelationTo	wasDerivedFrom	wasGeneratedBy
??mother	sio:Human	chear:Mother		??child		
??child	sio:Human	chear:Child		??mother		
??birth	chear:Birth			??child		
??preg	chear:Pregnancy			??child		
??sample1	S				??mother	
??sample2	S				??mother	
??pb_1	Pb		sio:isPartOf	??sample1		
??pb_2	Pb		sio:isPartOf	??sample2		

Table 4. SDD Example Codebook

Column	Code	Label	Class
race	0		chear:White
race	1		chear:BlackOrAfricanAmerican
race	2		chear:OtherRace
edu	0	high school degree or less	chear:HighSchoolOrLess
edu	1	technical college or some college	chear:SomeCollegeorTechnicalSchool
edu	2	college graduate	chear:CollegeGraduate
smoke	0	no smoking in pregnancy	chear:NonSmoker
smoke	1	some smoking in pregnancy	chear:Smoker

of abbreviated terms or units to their corresponding concepts. The set of code mappings used in CHEAR can be found on GitHub[4].

3.2 OWL Generation

Each cell of data from a dataset is used to create an instantiation of an attribute, based on the description of the column in the SDD. The value in the cell is used to assign a *sio:hasValue* property to the attribute instantiation. If the attributeOf column is filled out in the SDD, the *sio:isAttributeOf* property is used to link to the corresponding entity instantiation.If a unit is specified, the *sio:hasUnit* property is assigned the corresponding unit from the Units Ontology, which is determined by using the Code Mappings table. If a timepoint for the corresponding variable is specified, it is included in the OWL using the *sio:existsAt* property. The timepoint may also have an associated value, unit, and relation, as shown in the example OWL below.

```
:birthweight a chear:Weight;
    sio:isAttributeOf :joe;
    sio:hasValue 3;
    sio:hasUnit uo:kilogram;
    sio:existsAt [ a  sio:TimeInterval, chear:BirthTime;
```

[4] https://github.com/tetherless-world/chear-ontology/blob/master/code_mappings.csv

```
        sio:hasValue 0;
        sio:hasUnit sio:Day;
        sio:inRelationTo :birth  ];
    sio:existsAt [ a sio:TimeInterval, chear:GregorianTime;
        sio:hasValue "2016-03-12"^^xsd:dateTime;
        sio:hasUnit sio:Day;
        sio:inRelationTo :birth  ].
```

4 Evaluation

The SDD specification approach was applied to the National Health and Nutrition Survey (NHANES) data from 2013-2014. In a manner specifically tailored to the NHANES website structure, values for columns in the SDD specification were populated through a web scraping script that used the Python Beautiful Soup package. In order to assign attributes and entities, a look-up approach was used to compare NHANES entries with terms in SIO or CHEAR. Using this approach, we were able to generate SDD starting points and Codebooks for 150 documents in 6 categories (Questionnaire, Demographics, Dietary, Laboratory, Examination, and Limited Access) corresponding to roughly 4818 SDD rows and over 17000 codebook entries. Of the 4818 SDD rows, 1148 or 23.83% were mapped to existing concepts in SIO such as Age, Height, Race and Ethnicity, as well as terms from CHEAR, including Weight, Education Level, Language, and Income. The remaining rows were not mapped to any concepts due to limitations in the extraction algorithm, which used pattern matching in the labels and comments to search for the above SIO and CHEAR terms, rather than more advanced natural language processing techniques. Therefore, while the current process reduces the amount of time required, human input is still necessary to complete the annotation. It is an ongoing effort to manually annotate the remaining NHANES concepts. Furthermore, the SDD specification is being applied to additional publicly available datasets, including the Genomic Data Commons[5], the Surveillance, Epidemiology, and End Results Program[6], and the Medical Information Mart for Intensive Care [6]. Additionally, by using a script to convert from SDDs, Codebooks, and the corresponding data into the Resource Description Framework (RDF), Knowledge Graphs have been created for the subset of NHANES that had been annotated. These graphs are being actively used in a Data Analytics course at Rensselaer Polytechnic Institute to demonstrate to students how semantics can be leveraged to perform analytics.

5 Discussion

Concentrating on mapping many data sets to one single conceptual structure serves the semantic web goal of interoperability: by mapping to the SIO conceptualization datasets can be compared to any other dataset that has also

[5] https://gdc.cancer.gov
[6] https://seer.cancer.gov

been mapped. A Semantic Data Dictionary provides a formal means to map dataset columns into a compositional structure in a way that allows us to produce OWL-based metadata for those datasets, creating explicitly defined classes that dataset columns map to. For some studies, like NHANES, tools for web scraping can be used, such as the Python library Beautiful Soup [14], allowing for a semi-automatic population of variable names, labels, and definitions. Nevertheless, automating the population of entities, roles or relations that correspond to the variable cannot be accomplished simply by using web scraping techniques, requiring the collaboration with domain experts.

6 Conclusions

The Semantic Data Dictionary (SDD) standard allows for extension and integration of data from multiple public health and biomedical domains through a common metadata standard, and is convertible to OWL-based metadata that can be used to query for relevant datasets without knowledge of the structure of any one dataset. The CHEAR project uses the SDD specification to describe data related to demographics, anthropometry, birth outcomes, pregnancy characteristics, biological responses and targeted analytes. The Center for Architecture Science and Ecology (CASE) is using Semantic Data Dictionaries to annotate data related to biological and physical environments, human demographics and physiology, and cognition. The Healthy Birth, Growth, and Development (HBGD) is using the SDD specification to capture data summary statistics, such as mean, standard deviation, minimum and maximum confidence interval values, counts, and time information. As demonstrated by its applicability in the above projects, the SDD specification is an approach for semantic annotation that can be used to represent attributes described by data elements to allow for the integration of data from multiple sources.

7 Acknowledgements

This work was funded by the National Institute of Environmental Health Sciences (NIEHS) Award 0255-0236-4609 / 1U2CES026555-01. We would like to thank Susan Teitelbaum at the Icahn School of Medicine at Mount Sinai for her leadership on the overall CHEAR data resource project, as well as her guidance in exposure and health domains.

References

1. Buneman, P., Davidson, S., and Kosky, A. Theoretical aspects of schema merging. In *Advances in Database TechnologyEDBT'92* (1992), Springer, pp. 152–167.
2. Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A., and Zien, J. Y.

Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the 12th International Conference on World Wide Web* (New York, NY, USA, 2003), WWW '03, ACM, pp. 178–186.

3. Dingli, A., Ciravegna, F., and Wilks, Y. Automatic semantic annotation using unsupervised information extraction and integration. In *Proceedings of SemAnnot 2003 Workshop* (2003).
4. Dumontier, M. The Semanticscience Integrated Ontology (SIO). `http://sio.semanticscience.org`.
5. Dumontier, M., Baker, C. J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N. R., Duck, G., Furlong, L. I., Keath, N., Klassen, D., McCusker, J. P., Queralt-Rosinach, N., Samwald, M., Villanueva-Rosales, N., Wilkinson, M. D., and Hoehndorf, R. The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics 5*, 1 (2014), 14.
6. Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data 3* (2016).
7. Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., and Goranov, M. *Semantic Annotation, Indexing, and Retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 484–499.
8. Lenzerini, M. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (2002), ACM, pp. 233–246.
9. Maynard, D. Multi-source and multilingual information extraction. *Expert Update 6*, 3 (2003), 11–16.
10. McBrien, P., and Poulovassilis, A. A formalisation of semantic schema integration. *Information Systems 23*, 5 (1998), 307 – 334.
11. McCusker, J. P., Rashid, S. M., Liang, Z., Liu, Y., Chastain, K., Pinheiro, P., Stingone, J. A., and McGuinness, D. L. Broad, interdisciplinary science in tela: An exposure and child health ontology.
12. Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., and Goranov, M. Kim–semantic annotation platform. In *International Semantic Web Conference* (2003), Springer, pp. 834–849.
13. Reeve, L., and Han, H. Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM Symposium on Applied Computing* (New York, NY, USA, 2005), SAC '05, ACM, pp. 1634–1638.
14. Richardson, L. Beautiful soup documentation, 2007.
15. Sarasua, C., Simperl, E., and Noy, N. F. Crowdmap: Crowdsourcing ontology alignment with microtasks. In *International Semantic Web Conference* (2012), Springer, pp. 525–541.
16. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology 25*, 11 (2007), 1251.
17. Staab, S., Maedche, A., and Handschuh, S. *An annotation framework for the semantic web*. Inst. AIFB, Univ., 2001.
18. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., and Ciravegna, F. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web 4*, 1 (2006), 14 – 28.