The Semantic Data Dictionary Approach to Data Annotation & Integration

Sabbir Rashid
Katherine Chastain
Jeanette Stingone
Deborah McGuinness
Jim McCusker

Tetherless World Constellation

Rensselaer Polytechnic Institute

October 20, 2017

Overview

1. Motivation
2. Introduction
3. Related Work
4. Methods
   - The Semantic Data Dictionary Specification
   - Semantic Data Dictionary Examples
   - Knowledge Representation
5. Evaluation
6. Discussion
7. Conclusion

Motivation

- Standard Data Dictionary
  - A controlled vocabulary [Linnarsson and Wigertz, 1989]
  - Human readable
  - Difficult for a machine to understand
  - Integration tasks are not easily automated

- Semantic Data Dictionary (SDD) Specification
  - Allows for integration of data from multiple domains
  - Uses a common metadata standard
  - Can leverage domain-relevant terminology
  - Helps with the specification, curation and search of data
  - "Compositional" descriptions, rather than 1:1 mappings

Introduction

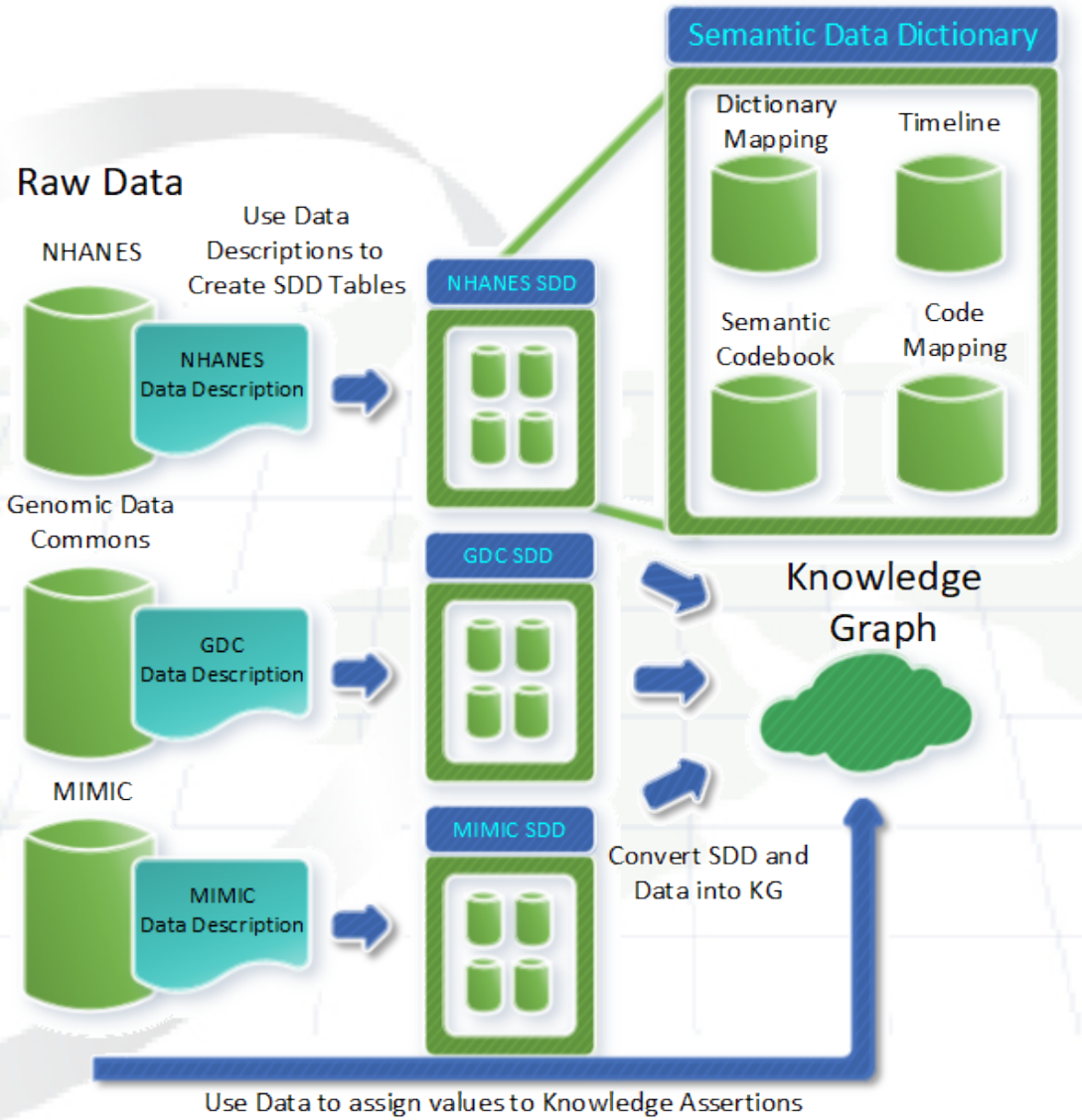
Some challenges a scientist may face include

- Finding data across data resources with the same semantic meaning
- Annotation of data without a vast knowledge of ontologies
- A single row in a dataset may contain reference on multiple entities

The Semantic Data Dictionary Specification

- Facilitates finding data that are relevant for comparison
- Creates a more accessible workflow

SDD Annotation Flow Diagram



Related Work

Data Integration - The ability to unite data from multiple sources in such a way that results in a unified view of the combined data [Lenzerini, 2002]

- Use of ontologies to annotate data is increasing
- Adhering to a common foundational model such as SIO or the OBO Foundry ontologies facilitates data integration

Schema Merging - An approach to integrate information from multiple datasets

- Two general methods commonly used [Buneman et al., 1992]
- Considerations when performing schema merging [McBrien and Poulovassilis, 1998]
- Methods, like CrowdMap [Sarasua et al., 2012], may take advantage of crowd sourcing

Semantic Annotation - The practice of assigning metadata descriptions about entities in a database or text

The Semantic Data Dictionary Specification

The Semantic Data Dictionary Specification - A way to represent implicit entities and their relationships using a general ontology

- Semanticscience Integrated Ontology (SIO)
  - High-level conceptualization of data
  - Three entity world view: attributes, objects, and processes [Dumontier et al., 2014]
  - General properties to describe the relationships
- Can use domain-specific ontologies to allow more fine-grained and dataset-specific annotations
- Contains information about the entity types represented in a tabular dataset
- Conveys relationships in both machine-readable and unambiguous manners

Dictionary Mapping Example (Actual)

Table 1: Example Dictionary Mapping (Actual Entry)

Column	Attribute	attributeOf	Unit	Time	inRelationTo
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	chear:Weight	??mother	kg	??visit1	
height	sio:Height	??mother	cm	??visit1	
smoker	chear:SmokingStatus	??mother		??pregn	
pb_1	sio:Concentration	??pb_1	mgL	??visit1	??sample1
pb_2	sio:Concentration	??pb_2	mgL	??visit2	??sample2
ga	chear:GestationalAge	??child	sio:Week	??birth	
birthwt	chear:Weight	??child	kg	??birth	

Dictionary Mapping Example (virtual)

Table 2: Example Dictionary Mapping (Virtual Entry)

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

The SDD Specification Table

Table 3: Semantic Data Dictionary Mappings

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

Codebook Code Mapping Tables

Table 4: Codebook and Code Mapping Tables

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

code	uri	label
Pb	chebi:25016	Lead
S	uberon:0001977	Serum
cm	obo:UO_0000015	centimeter
kg	obo:UO_0000009	kilogram
kgm2	obo:UO_0000086	kilogram per square meter
mgL	obo:UO_0000273	milligrams per liter

RDF Knowledge Entry (CHEAR)

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

Smoking Status Query

```
SELECT DISTINCT * WHERE {
?s rdf:type ?attribute .
OPTIONAL {?s sio:existsAt ?timepoint . }
?s sio:isAttributeOf ?attrOf .
?s sio:hasValue ?attributeVal .
OPTIONAL { ?cohort rdf:type hasco:Cohort .
      ?cohort sio:isAttributeOf ?attrOf .
      ?cohort sio:hasValue ?cohortVal . }
OPTIONAL {?s sio:hasUnit ?unit . }
FILTER (?attributeVal != 'NA') .
FILTER (?attribute = chear:SmokingStatus ) .
}
LIMIT 10000
```

Evaluation

- Approach was applied to the National Health and Nutrition Survey (NHANES) data from 2013-2014
  - Able to generate SDD and Codebook starting points
    - 150 documents
    - 6 categories
    - 4818 Dictionary Mapping rows
    - Over 17000 codebook entries
  - Of the 4818 SDD rows, 1148 or 23.83% were mapped to existing concepts in SIO or CHEAR
  - The remaining rows were not mapped to any concepts due to limitations in the extraction algorithm
    - Values for columns were populated through a web scraping script
    - A look-up approach was used to compare NHANES labels with terms in SIO or CHEAR

Discussion

- Human input is still necessary to complete the annotation
- Ongoing effort to manually annotate the remaining NHANES concepts
- The SDD specification is being applied to additional publicly available datasets
- Knowledge Graphs have been created for the subset of NHANES that had been annotated
- Actively used in a Data Analytics course at RPI to demonstrate to students how semantics can be leveraged to perform analytics

Conclusion

- Provides a formal means to annotate dataset columns
- Allows production of OWL-based metadata
- Helps address Semantic Web goal of Interoperability
  - Single conceptual structure
  - Comparison to any other dataset that has also been mapped
- Domain scientists can produce high quality integrated data

Projects -

- Children's Health Exposure Analysis Resource (CHEAR)
- The Center for Architecture Science and Ecology (CASE)
- The Healthy Birth, Growth, and Development (HBGD)
- The Center for Health Empowerment by Analytics, Learning, and Semantics (HEALS)

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.
- The members of the Center for Heath Empowerment by Analytics, Learning, and Semantics (HEALS) for their contributions to this project.
- The researchers working on the Human-Aware Data Acquisition Framework (HADatAc), including Paulo Pinheiro, Zhicheng Liang, and Yue Liu.

References I

Buneman, P., Davidson, S., and Kosky, A. (1992).
Theoretical aspects of schema merging.
In *Advances in Database TechnologyEDBT'92*, pages 152–167. Springer.

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

Dingli, A., Ciravegna, F., and Wilks, Y. (2003).
Automatic semantic annotation using unsupervised information extraction and integration.
In *Proceedings of SemAnnot 2003 Workshop*.

Dumontier, M.
The Semanticscience Integrated Ontology (SIO).
`http://sio.semanticscience.org`.

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. (2014).
The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery.
Journal of Biomedical Semantics, 5(1):14.

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. (2016).
Mimic-iii, a freely accessible critical care database.
Scientific data, 3.

References II

Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., and Goranov, M. (2003).
Semantic Annotation, Indexing, and Retrieval, pages 484–499.
Springer Berlin Heidelberg, Berlin, Heidelberg.

Lenzerini, M. (2002).
Data integration: A theoretical perspective.
In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246. ACM.

Linnarsson, R. and Wigertz, O. (1989).
The data dictionary–a controlled vocabulary for integrating clinical databases and medical knowledge bases.
Methods of information in medicine, 28(2):78–85.

Maynard, D. (2003).
Multi-source and multilingual information extraction.
Expert Update, 6(3):11–16.

McBrien, P. and Poulovassilis, A. (1998).
A formalisation of semantic schema integration.
Information Systems, 23(5):307 – 334.

McCusker, J. P., Rashid, S. M., Liang, Z., Liu, Y., Chastain, K., Pinheiro, P., Stingone, J. A., and McGuinness, D. L. (2017).
Broad, interdisciplinary science in tela: An exposure and child health ontology.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., and Goranov, M. (2003).
Kim–semantic annotation platform.
In *International Semantic Web Conference*, pages 834–849. Springer.

References III

Rashid, S., Chastain, K., Stingone, J., McGuinness, D., and McCusker, J. (2017).
The semantic data dictionary approach to data annotation & integration.
In *Proceedings of the First Workshop on Enabling Open Semantic Science (SemSci)*, pages 47–54.

Reeve, L. and Han, H. (2005).
Survey of semantic annotation platforms.
In *Proceedings of the 2005 ACM Symposium on Applied Computing*, SAC '05, pages 1634–1638, New York, NY, USA. ACM.

Richardson, L. (2007).
Beautiful soup documentation.

Sarasua, C., Simperl, E., and Noy, N. F. (2012).
Crowdmap: Crowdsourcing ontology alignment with microtasks.
In *International Semantic Web Conference*, pages 525–541. Springer.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. (2007).
The obo foundry: coordinated evolution of ontologies to support biomedical data integration.
Nature biotechnology, 25(11):1251.

Staab, S., Maedche, A., and Handschuh, S. (2001).
An annotation framework for the semantic web.
Inst. AIFB, Univ.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., and Ciravegna, F. (2006).
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):14 – 28.

Questions?