Aligning Diverse Data for a Personalized Health Knowledge Graph

Matthew Johnson, Meenu Ravi, Sabbir Rashid, Paulo Pinheiro and Deborah L. McGuinness
Rensselaer Polytechnic Institute, Troy, NY, 12180, USA

Introduction

Personalized health knowledge graphs (PHKGs) are data repositories that contain the personal and medical data associated with a user [1]. Contextualizing this information with medical guidelines can provide users with personalized health insights to prevent disease and live healthier lives. However, one of the primary challenges is to integrate diverse sets of data into a cohesive PHKG. This is a two-stage process that requires data providers to gather the necessary information and align that information to a vocabulary that a software agent can understand and use. For example, if an overweight user is trying to lose weight, they may consult the Clinical Practice Guidelines for Medical Care of Patients with Obesity [2] for advice or interact with a software agent like an Alexa device, which can use the guidelines to provide advice. To support such an agent, a user's PHKG will need to be populated with step and weight information from their smart devices, drug information from their pharmacy, and calorie intake from an app on their phone. This information must be aligned such that the data will link to concepts within the agent's operating vocabulary. This task is challenging because the data comes from multiple sources with a variety of schemas and templates. One solution to this problem is to use Semantic Data Dictionaries (SDD) [3] to normalize information fed into the PHKG, according to a well-defined set of ontologies. In this case, the SDDs are used to form links between user data (and their relevant data schema(s)) and concepts from the selected ontologies that are needed to enable inferencing and data alignment. In this paper, we present a new tool to generate SDDs along with two toolsets, sdd2rdf and HADatAc, which can be used in concert with SDDs to populate a PHKG. All of these tools were developed under the NIH's Children's Health Exposure Analysis Resource (CHEAR) program and expanded under the follow-on Human Health (HHEAR) program[1] to harmonize data from a variety of epidemiology studies and, as a result, can easily be utilized for data integration into a PHKG.

Semantic Data Dictionary Editor

A data dictionary (DD) is a collection of metadata that describes what a data element, e.g., a data column, means along with its relation to other dataset content. Similarly, SDDs formally describe the data entities using a domain ontology and their relationship to other entities, using ontologies such as the Semanticscience Integrated Ontology (SIO) [4]. To create an SDD, a user must generate a spreadsheet that maps each of the terms within the DD to terms within the domain ontologies. When a user is familiar with the DD and the ontology terminology, this process may be time-consuming but not difficult, however, when a new domain is encountered, this process becomes significantly more challenging. To lower the barrier to entry and to make the process more efficient, we have developed an SDD-Editor that simplifies the modeling process and enables domain experts to fully generate an SDD from a DD.

The SDD-Editor has been developed to reuse the visual metaphors common in spreadsheet editing software while integrating features needed to form links between medical concepts and ontologies. The SDD-Editor auto-populates rows from data dictionaries and provides the data description of each row when a domain expert clicks on a cell. SDD-Gen suggests potential class matches for various cells within the SDD; if no suitable suggestion exists from the selected ontologies, the domain expert can make

[1] https://www.niehs.nih.gov/research/supported/exposure/hhear/index.cfm

biomedical class searches in other ontologies using a search engine powered by Bioportal [5]. Once complete, the domain expert can run a validator which checks that all classes are valid and the SDD is correctly formatted. These features enable domain experts to transition from the role of class aggregator to class curator, and thus reduce their cognitive workload.

SDD-Gen is an inference engine that takes in a data dictionary description of a column of data, as well as a list of ontologies, i.e., the selected ontologies, and returns the most semantically similar ontology class. This is done by parsing both the data dictionary and the class descriptions from selected ontologies using an abstract meaning representation (AMR) parser [6]. Keywords are chosen from the AMR tree and mapped to a pre-trained GLoVe embedding space [7] where semantically similar words appear closer together. The keyword vectors are concatenated together and a lower-dimensional representation is learned [8]. In this new space, the closest matches between the data dictionary descriptions and the ontology class descriptions are returned as predictions.

Semantic Data Dictionary Integration

Once an SDD has been created, a data provider can use either sdd2rdf, a lightweight script that generates knowledge graph fragments as nanopublications [9], or HADatAc, a data acquisition infrastructure capable of integrating data for personalized health knowledge graphs. Nanopublications help to represent scientific statements and allow for the capture of the attribution, quality, and provenance surrounding a given assertion. In general, sdd2rdf is more useful for data providers who are looking for a simple data transformation without the need for more complex knowledge management solutions. On the other hand, HADatAc is designed for an end-user, providing an easy to use web interface that allows users to perform an in-depth faceted search of their personalized health knowledge graph.

sdd2rdf is a library written in Python for the interpretation of Semantic Data Dictionaries. It is written in a stand-alone manner, such that users who wish to convert data into a knowledge graph can do so by providing the SDD tables along with the data, without the need for additional documents. Furthermore, sdd2rdf is made openly available along with additional documentation, including examples, annotated SDDs, the resulting RDF, and tutorials.[2] Users that wish to generate the complete knowledge graph fragments are encouraged to run the sdd2rdf script for any of the datasets discussed in the documentation, using the provided Semantic Data Dictionaries[3] in conjecture with the publicly available data files. This includes data from the National Health and Nutrition Examination Survey (NHANES)[3], Genomic Cancer Atlas (TCGA)[4], and Clinical Interpretation of Variants in Cancer (CIViC)[5] datasets.

HADatAc[10] is a smart data acquisition infrastructure that ingests content from data files and live message streams into a queryable cross-study, cross-project knowledge graph. In the process of moving data from data sources into its knowledge graph, HADatAc normalizes, annotates, and integrates study data and metadata. This infrastructure has been designed to harmonize data across multiple studies in the area of environmental sciences, building sciences, engineering, global health, and more recently, to better understand how environmental factors impact human health. HADatAc provides a user interface, allowing users to load/explore domain ontologies, ingest datasets along with SDDs, and explore the generated knowledge graph using a faceted browser.

---

[2] https://tetherless-world.github.io/sdd/
[3] https://github.com/tetherless-world/sdd/tree/master/sdd_resources
[3] https://www.cdc.gov/nchs/nhanes/index.htm
[4] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
[5] https://civicdb.org/home/

Conclusion

We have presented our SDD-Editor that can align the diverse sets of data that are required by PHKGs, and thus make this new integrated knowledge graph a potential key knowledge source for PHKG-based applications, like the Alexa-based agent mentioned in the introduction. The SDD-Editor can be used by data providers or users to align schema information from different data sources, such as step count devices and applications, prescription information in health care systems, and calorie intake in personalized phone applications, to a given domain vocabulary. The generated SDD can then be used by either sdd2rdf or HADatAc, along with the data from the various sources, to populate a PHKG. Once populated, a weight loss agent could operate over the PHKG and provide the user with personalized weight loss guidance directly from medical guidelines. In addition, the user could perform their own analysis of the PHKG using HADatAc to further explore their data and generate additional insights.

References

[1] A. Gyrard, M. Gaur, S. Shekarpour, K. Thirunarayan, and A. Sheth, "Personalized health knowledge graph," in *ISWC 2018 Contextualized Knowledge Graph Workshop*, 2018.

[2] W. T. Garvey, J. I. Mechanick, E. M. Brett, A. J. Garber, D. L. Hurley, A. M. Jastreboff, K. Nadolsky, R. Pessah-Pollack, R. Plodkowski, and Reviewers of the AACE/ACE Obesity Clinical Practice Guidelines, "American association of clinical endocrinologists and american college of endocrinology comprehensive clinical practice guidelines for medical care of patients with obesity," *Endocrine Practice*, vol. 22, no. s3, pp. 1–203, 2016.

[3] M. Dumontier, et. al., "The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery," *Journal of biomedical semantics*, vol. 5, no. 1, p. 14, 2014.

[4] R. Sabbir M, K. Chastain, J. A. Stingone, D. L. McGuinness, and J. P. McCusker, "The semantic data dictionary – an approach for describing and annotating data," *Data Intelligence*, 2020.

[5] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jon-quet, D. L. Rubin, M. A. Storey, C. G. Chuteet al., "Bioportal: ontologies and integrated data resources at the click of a mouse," *Nucleic acids research*, vol. 37, pp. W170–W173, 2009.

[6] C. Wang, S. Pradhan, X. Pan, H. Ji, and N. Xue, "Camr at semeval-2016 task 8: An extended transition-based amr parser," in *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, 2016, pp. 1173–1178.

[7] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *2014 conference on empirical methods in natural language processing (EMNLP)*, 2014.

[8] L. Huang, J. May, X. Pan, and H. Ji, "Building a fine-grained entity typing system overnight for a new x (x= language, domain, genre)," *arXiv preprint arXiv:1603.03112*, 2016.

[9] P. Groth, A. Gibson, and J. Velterop, "The anatomy of a nanopublication," *Information Services & Use*, vol. 30, no. 1-2, pp. 51–56, 2010.

[10] P. Pinheiro, H. Santos, Z. Liang, Y. Liu, S. M. Rashid, D. L. McGuinness, and M. P. Bax, "Hadatac: A framework for scientific data integration using ontologies." in *International Semantic Web Conference*, 2018.

Acknowledgements: This work is partially supported through AFRL 88ABW-2020-0991 and NIH 2U2CES026555-02.