SciKG: Tutorial on Building Scientific Knowledge Graphs from Data, Data Dictionaries, and Codebooks*

Henrique Santos[1], Paulo Pinheiro[2], Jamie P. McCusker[1], Sabbir M. Rashid[1] and Deborah L. McGuinness[1]

[1]*Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy NY, United States*
[2]*Parcela Semântica Lda, Funchal, Portugal*

Abstract

Data from scientific studies are published in datasets, typically accompanied by data dictionaries and codebooks to support data understanding. To conduct rigorous analysis, data users need to leverage this documentation to correctly interpret the data. While this process can be burdensome for new data users, it is also prone to errors even for seasoned users. A computational formal model of the knowledge that was used to create the study can facilitate better understanding and thus improved usage of the study data. Knowledge graphs can be used effectively to capture this study knowledge. The SciKG tutorial aimed to introduce participants to the basics of knowledge graph construction using data, data dictionaries, and codebooks from scientific studies. It used the Center for Disease Control and Prevention's (CDC) National Health and Nutrition Examination Surveys (NHANES) data as a testbed and introduce standardized terminology, novel and established techniques, and resources such as scientific/biomedical ontologies, semantic data dictionaries, and knowledge graph frameworks in both lecture and practical sessions.
Website: https://tetherless-world.github.io/scikg-eswc-2023/

1. Tutorial Overview

The construction of knowledge graphs (KGs) for the biomedical domain (and, generally, the scientific domain) is a prominent field, with much attention from the community. Several venues have highlighted this, including the recently organized Personal Health Knowledge Graph workshop[1]. Scientific KGs have been deployed to support increasing automation in biomedical research, including for reproducible research [1].

SciKG was a **full-day** tutorial that introduced participants to the basics of knowledge graph construction with input from datasets from scientific studies and surveys, as well as the associated data dictionaries, codebooks, and documentation. To support this, the tutorial began with an overview of state-of-the-art scientific and biomedical ontologies that are commonly reused. Next, the participants were introduced to Semantic Data Dictionaries (SDDs) [2] and

ESWC'23: The 20th Extended Semantic Web Conference, May 28–June 2, 2023, Hersonissos, Greece
* Partially funded by the National Institute of Environmental Health Sciences (NIEHS) through the Human Health Exposure Analysis Resource (HHEAR) Data Center project number U2CES026555-02.
oliveh@rpi.edu (H. Santos); paulo@psemantica.com (P. Pinheiro); mccusj2@rpi.edu (J. P. McCusker); rashis2@rpi.edu (S. M. Rashid); dlm@cs.rpi.edu (D. L. McGuinness)
0000-0002-2110-6416 (H. Santos); 0000-0001-8469-4043 (P. Pinheiro); 0000-0003-1085-6059 (J. P. McCusker); 0000-0002-4162-8334 (S. M. Rashid); 0000-0001-7037-4567 (D. L. McGuinness)

[1]https://phkg.github.io

learned to create simple but functional SDDs to model some aspects of the publicly available NHANES data in practical sessions [3]. Using the modeling in SDDs, participants bootstrapped and interacted with the KG via established knowledge graph frameworks.

The SciKG tutorial was based on simple instructive (but insightful) examples. At the end of the tutorial, participants were be able to:

- Identify and reuse relevant scientific and biomedical ontologies
- Develop minimally working semantic data dictionaries that capture domain modeling from scientific data and documentation
- Use knowledge graph frameworks to bootstrap and manage scientific knowledge graphs
- Interact with the graph to retrieve data based on analysis-driven questions

This tutorial covered methods and tools that are established and being used in production environments at several institutions, including National Institute of Health-funded projects at the Icahn School of Medicine at Mount Sinai[2], McGill University's Peter Guo-hua Fu School of Architecture[3], Rensselaer Polytechnic Institute's Tetherless World Constellation[4], and Escola de Ciência de Informação at Universidade Federal de Minas Gerais[5]. The SciKG tutorial was inspired by previous tutorials on knowledge graph construction, including the Knowledge Graph Construction Tutorial[6] held at ESWC 2022, and the Tools for Creating and Exploiting Large Knowledge Graphs (KGTK)[7] held at ISWC 2021. These tutorials were focused on general tools for KG building, management, and exploration. SciKG, in its turn, focused on the construction of KGs from scientific data, with rigor in scientific knowledge maintenance and representation, covering not only techniques, but standardized scientific terminology and best practices. According to the conference organizers, SciKG had an average attendance of 25 people.

SciKG was primarily targeted at Semantic Web researchers working (or willing to work) with biomedical data. However, the acquired knowledge applies to studies beyond the biomedical domain, as the aspects of the scientific methods for data acquisition are common. The target audience included students and researchers interested in bioinformatics settings, as well as anyone interested in conducting research using scientific data, often from multiple sources.

2. Topics

The SciKG tutorial was divided into four sections. It started with an overview of how scientific study data is usually acquired, organized and published, and the current challenges (and opportunities for semantic web) involving the use of this data. Next, we introduced methods for scientific data annotation and terminology reuse. Following, we gave an overview of the state-of-the-art scientific and biomedical ontologies and provided real-world examples of their successful adoption. Finally, the tutorial introduced knowledge graph frameworks and demonstrate how they can be used to bootstrap and manage scientific KGs.

[2] https://hhearprogram.org/data-center
[3] https://www.mcgill.ca/architecture/
[4] https://tw.rpi.edu/project/human-health-exposure-analysis-repository-hhear
[5] http://eci.ufmg.br
[6] https://kg-construct.github.io/eswc-dkg-tutorial-2022/
[7] https://usc-isi-i2.github.io/kgtk-tutorial-iswc-2021/

Studies, Data, and Documentation

This section presented a common scientific methodology used in scientific studies and demonstrate how study data is usually acquired, organized, and published. Using the NHANES survey as an example, this section demonstrated how study documentation is usually used to describe the contents of data files using data dictionaries and codebooks, and how it is problematic in capturing all the semantics associated with the variable. Data dictionaries document the study variables contained in data files, providing a natural language description of the variable. For instance, the RIDAGEYR variable in NHANES is defined as *Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age*, and we can infer that it is measuring the age of the survey participant (not of other persons included in the data), using years as the unit of measurement (not months which is used for infants), recorded during screening time (instead of examination time).

Scientific and Biomedical Ontologies

This section introduced participants to some well-used state-of-the-art scientific and biomedical ontologies, including the Semanticscience Integrated Ontology (SIO) [4], the Human Aware Science Ontology (HAScO) [5], the Disease Ontology (DOID) [6], and the Chemical Elements of Biological Interest (ChEBI) [7] ontology, among others. We demonstrated how these terminologies can be and have been used to comprehensively model scientific knowledge in successful use cases.

Semantic Data Dictionaries

This topic demonstrated techniques for capturing study knowledge from data, and documentation. We introduced the Semantic Data Dictionary method for aligning and integrating data and provide a sufficient set of examples. A Semantic Data Dictionary (SDD) is a model for representing metadata from ontologies and structured vocabularies through a set of specifications that allows the assignment of a semantic representation of the data [2]. By the end of the practical sessions, participants were be able to produce simple SDDs using the NHANES data.

Knowledge Graph Frameworks

This topic introduced knowledge graph frameworks, and presented two that have built-in capabilities of working with SDDs, including the Human-Aware Data Acquisition Framework (HADatAc) [8] and Whyis [9]. HADatAc is an open-source infrastructure that enables combined acquisitions of data and metadata in a way that metadata is properly and logically connected to data, interacting with these sources to move the data from their transient state into a persistent repository, and enabling the data to be retrieved from their persistent repositories through the use of queries. Whyis is an open-source framework for creating custom provenance-driven knowledge graphs, supporting three principal tasks: knowledge curation, inference, and interaction. All knowledge in Whyis graphs is encapsulated in nanopublications, which simplify and standardize the production of qualified knowledge in knowledge graphs.

3. Tutorial Resources

Links to all utilized resources can be found in Table 3.

Tutorial website	https://tetherless-world.github.io/scikg-eswc-2023/
NHANES SDDs	https://github.com/tetherless-world/nhanes-hadatac
HADatAc	https://hadatac.org
Whyis	https://github.com/tetherless-world/whyis

Table 1
Tutorial resources.

References

[1] D. N. Nicholson, C. S. Greene, Constructing knowledge graphs and their biomedical applications, Computational and Structural Biotechnology Journal 18 (2020) 1414–1428.

[2] S. M. Rashid, J. P. McCusker, P. Pinheiro, M. P. Bax, H. Santos, J. A. Stingone, A. K. Das, D. L. McGuinness, The Semantic Data Dictionary – An Approach for Describing and Annotating Data, Data Intelligence 2 (2020) 443–486.

[3] H. Santos, P. Pinheiro, D. L. McGuinness, Knowledge Graph Construction from Data, Data Dictionaries, and Codebooks: the National Health and Nutrition Examination Surveys Use Case, 2022. URL: https://us2ts.org.

[4] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath, D. Klassen, J. P. McCusker, N. Queralt-Rosinach, M. Samwald, N. Villanueva-Rosales, M. D. Wilkinson, R. Hoehndorf, The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery, Journal of Biomedical Semantics 5 (2014) 14.

[5] P. Pinheiro, M. Bax, H. Santos, S. M. Rashid, Z. Liang, Y. Liu, J. P. McCusker, D. L. McGuinness, Annotating Diverse Scientific Data with HAScO, in: Proceedings of the Seminar on Ontology Research in Brazil 2018 (ONTOBRAS 2018). São Paulo, SP, Brazil, 2018.

[6] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein, K. Bisordi, N. Campion, B. Hyman, D. Kurland, C. P. Oates, S. Kibbey, P. Sreekumar, C. Le, M. Giglio, C. Greene, Human Disease Ontology 2018 update: classification, content and workflow expansion, Nucleic Acids Research 47 (2019) D955–D962.

[7] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, C. Steinbeck, The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013, Nucleic Acids Research 41 (2013) D456–D463.

[8] P. Pinheiro, H. Santos, Z. Liang, Y. Liu, S. M. Rashid, D. L. McGuinness, M. P. Bax, HADatAc: A Framework for Scientific Data Integration using Ontologies, in: Proceedings of the ISWC Posters & Demonstrations Track, 2018.

[9] J. McCusker, D. L. McGuinness, Whyis 2: An Open Source Framework for Knowledge Graph Development and Research, in: The Semantic Web, Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, 2023, pp. 538–554.