CI Compass
TECH NOTES
ci-compass.org

Tracking community access to Data Lifecycle data using Knowledge Graphs

Authors: Don Brower and Rodney Ewing
Date Published: August 23, 2022
DOI: 10.5281/zenodo.7007868

What is a Knowledge Graph?

A *knowledge graph* is a way to represent data and information that emphasizes the relationships between things. Conceptually, we think of them as a mathematical graph with labeled nodes and labeled edges connecting the nodes. One familiar example is the "citation graph" where the nodes represent individual published articles and the edges represent the relationship "cited by". Figure 1 shows an example of a knowledge graph and the information's associated relationships and linkages.

Why are Knowledge Graphs Used?

While there are many ways to work with relationship data, knowledge graphs provide both a conceptual model and a standardized way of working with the information. Knowledge graphs can help improve data discovery, access, integration, and analysis (Lane et al., 2020). Additionally, the information within a knowledge graph can be collected, extracted, and integrated to find new sources of knowledge (Ehrlinger et al., 2016). Through entity linking, link prediction, Artificial Intelligence (AI), and machine learning, humans and their computational agents can help improve the rich context and refine the relationships among the contextual entities that is enabled through knowledge graphs (Lane et al., 2020).

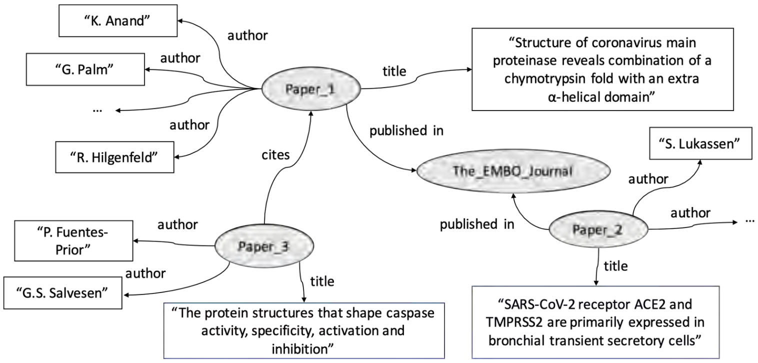
"K. Anand"
author
"G. Palm"
author
...
author
"R. Hilgenfeld"
Paper_1
title
"Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain"
published in
cites
The_EMBO_Journal
"S. Lukassen"
author
author
...
published in
Paper_2
title
"SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells"
"P. Fuentes-Prior"
author
Paper_3
author
"G.S. Salvesen"
title
"The protein structures that shape caspase activity, specificity, activation and inhibition"

Figure 1 A knowledge graph fragment from the citation/academic domain (Kejriwal, 2020).

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.

Processing and understanding bibliographic metadata (e.g., the citation graph) is a rich and useful application of knowledge graphs, but they are also useful in other contexts, such as pharmaceutical drug discovery or internet search engines (Singhal, 2012; Noy, 2018). These situations involve many diverse items, many types of relationships between items, and a need to find information based on this structure.

The building blocks of a knowledge graph are ontologies and identifiers. An *ontology* describes the classes of entities found in a knowledge graph and their possible relationships—akin to a database schema. For example, the conceptual "citation graph" has entities that are published articles and books, and the relation "cited by". The OpenAlex graph, the open source continuation of the Microsoft Academic Graph, extends the citation graph by having a broader definition of "scholarly work" that includes more items, such as datasets; having more types of entities, such as Authors and Institutions; and more kinds of relations (Priem, 2022).

Identifiers for entities and relationships are important, especially for facilitating the sharing and linking to other knowledge graphs. Shared ways of identifying people, organizations, and concepts include ORCID, the Research Organization Registry, and PubMed. There are ongoing efforts to standardize and unify the identity of things in other aspects of science, such as the persistent identification of scientific instruments (Krahl, 2022). Tracking the identity of articles and datasets is also a central concern of the FAIR (Findable, Accessible, Interoperable, and Reusable)data principles (Wilkinson et al., 2016). Many efforts to improve repositories and data with respect to the FAIR data principles also help with knowledge graphs. In addition to identification and citation of data, the machine-actionable metadata also helps computational agents to build and maintain knowledge graphs.

Since the relationships in a knowledge graph are labeled edges, it is possible to combine more than one ontology in a single knowledge graph. This lets the ontologies developed for different aspects of the scientific workflow to work together. One example ontology is PROV-O (https://www.w3.org/TR/prov-o/), a W3C standard for representing provenance information. Provenance is a key aspect to workflows and FAIR data, and PROV-O is general enough to be applied to any kind of entity, including intermediate data, published datasets, and articles. Another developed ontology is DINGO (https://dcodings.github.io/DINGO/) which is used for linking grants and funding with projects and outputs (Chialva, 2020).

A large, well adopted, public knowledge graph is *Wikidata* (https://www.wikidata.org/). Wikidata provides a public place where anyone can add concepts, things, or properties and get a unique identifier for them. This is very useful for sharing information and for enriching local knowledge graphs.

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.

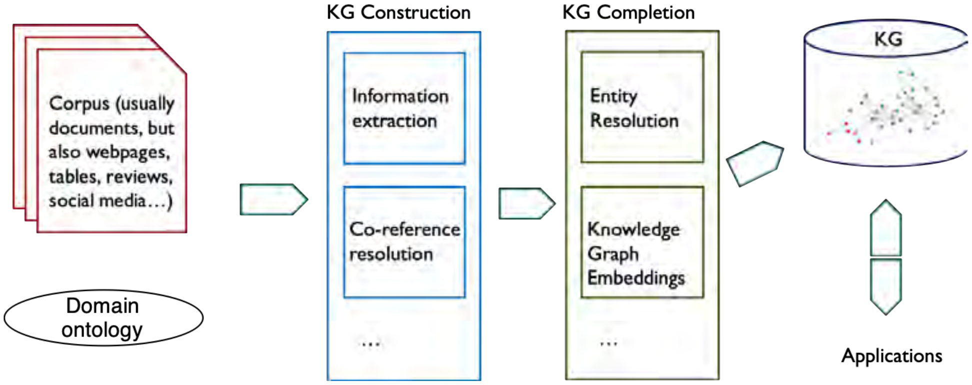

Figure 2: Schematic workflow for constructing a knowledge graph. (Kejriwal, 2020).

There are identifiers for many existing entities in the world, such as the Earth (identifier Q2), Science (the concept—Q336), Science (the journal—Q192864), the National Science Foundation (Q304878), and IceCube (the observatory—Q1569514).

Construction Process

A knowledge graph is useful only to the extent that it contains correct and relevant information. Thus a knowledge graph is not a static entity that is built once, but rather requires a process to update the data in it. This process from raw data to knowledge graph starts with identifying a corpus as a source; performing entity extraction from the source data; resolving duplicate entities; and then insertion into the knowledge graph. Figure 2 shows a schematic of this process.

Entity resolution and disambiguation is an important concern with knowledge graph construction. This is because knowledge graphs recognize that the same person, thing, or concept may have multiple ways of being referenced and we desire them to resolve to a unique identifier. The slogan "Things not strings," meaning the item of interest is the thing or person itself and not the name (Singhal 2012), captures this sentiment. This process, sometimes called information or knowledge extraction, can become complicated as the source text becomes more unstructured. Machine learning and NLP techniques can be applied to this process, which allows for scale and automation of the process (Noy, 2018).

Once established, the appropriate technologies for extraction, fusion, storage, and retrieval and visualization can be enabled to enhance search and discovery (Zhao et al., 2018). This offers the opportunity for the reuse of scientific data and research and establishes interoperability in the scientific research community with the potential to connect a network of scientists, who may or may not be connected, and their associated scientific research.

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.

What does it mean for NSF Major Facilities?

The NSF Major Facilities collect and make available a large number of datasets and other materials. There are many potential applications of knowledge graphs to facilitate the search and retrieval of this data, as well as for reporting on data use and impact. There is also the possibility of applying knowledge graphs to the science itself, such as understanding COVID-19 (Kejriwal, 2020).

Researchers and staff can use knowledge graphs to enable the connection of research and datasets through link prediction, entity linking, graph embedding, transitive inference, iterative improvement of machine learning models, and axiom configuration refinement. This rich context can also leverage open standards, open source, and the Administrative Data Research Facility (ADRF) framework to collect metadata on a specific scientific research topic. These methods can enable datasets and research to be connected that would not have otherwise been possible and in some cases multiple areas of research and datasets can become more useful when connected (Lane et al., 2020).

The additional structure of a knowledge graph allows for richer reporting that can supplement download counts and other measures used for showing impact and use of data. Citation counts for cited datasets and articles comes from counting specific links in a bibliometric knowledge graph. But such a knowledge graph would have many other useful relationships: citation and other links between papers, datasets, grants, and organizations. And the structure would allow for more useful breakdowns, such as counts by organization, funder, or subject area. The indirect links, such as following a grant to an awardee, to a dataset, and then a paper are a query especially suited for knowledge graphs.

References

Auer, S., Oelen, A., Haris, M., Stocker, M., D’Souza, J., Farfar, K. E., ... & Jaradeh, M. Y. (2020). Improving access to scientific literature with knowledge graphs. *Bibliothek Forschung und Praxis*, *44*(3), 516-529.

Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., & Vidal, M. E. (2018, June). Towards a knowledge graph for science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics* (pp. 1-6).

Berg-Cross, G. [Gary Berg-Cross]. (2022, February 2). *Wikidata: A knowledge graph for the earth sciences?* [Video]. YouTube. https://www.youtube.com/watch?v=3oN67CfirDI

Chialva, D., Mugabushaka, AM. (2020). DINGO: An Ontology for Projects and Grants Linked Data. In: , et al. ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium. TPDL ADBIS 2020 2020. Communications in Computer and Information Science, vol 1260. Springer, Cham. https://doi.org/10.1007/978-3-030-55814-7_15

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.

Ehrlinger, L., & Wöß, W. (2016). Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, *48*(1-4), 2.

Kejriwal, M. (2020). Knowledge Graphs and COVID-19: Opportunities, Challenges, and Implementation. Harvard Data Science Review, Special Issue 1. https://doi.org/10.1162/99608f92.e45650b8

Krahl, R., Darroch, L., Huber, R., Devaraju, A., Klump, J., Habermann, T., Stocker, M., & The Research Data Alliance Persistent Identification of Instruments Working Group members (2022). Metadata Schema for the Persistent Identification of Instruments (1.0). Research Data Alliance. https://doi.org/10.15497/RDA00070

Lane, J., Mulvany, I., & Nathan, P. (2020). Rich search and discovery for research datasets: Building the next generation of scholarly infrastructure.

Noy, Natasha, Gao, Yuqing, Jain, Anshu, Narayanan, Anant, Patterson, Alan, Taylor, Jamie. (2019) "Industry-scale Knowledge Graphs: Lessons and Challenges." ACM Queue, 17(2). https://queue.acm.org/detail.cfm?id=3332266

Priem, Jason, Piwowar, Heather, and Orr, Richard. "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts". (2022). Arxiv: https://arxiv.org/abs/2205.01833.

A. Singhal. “Introducing the Knowledge Graph: things, not strings”, Official Google Blog, May 2012. http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9. https://doi.org/10.1038/sdata.2016.18

Zhao, Z., Han, S. K., & So, I. M. (2018). Architecture of knowledge graph construction techniques. *International Journal of Pure and Applied Mathematics*, *118*(19), 1869-1883.

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.