Modeling with Knowledge Graphs

Kenneth Baclawski*, Gary Berg-Cross*, Cassiopeia Miles***, Douglas Miles***, Todd Schneider^, Ravi Sharma^^

*Northeastern University, **RDA/US Advisory Group, ***Logicmoo Co., ^Engineering Semantics, ^^Senior Enterprise Architect

Abstract

Knowledge graphs represent the confluence of many historical threads that have resulted in a popular new model for managing and retrieving large amounts of data. While a great deal of progress has been made on developing and deploying knowledge graph systems, many challenges remain. This special issue attempts to convey some of the flavor and excitement of this field.

Introduction

KNOWLEDGE GRAPHS (KGs) are lightweight versions of semantic networks that potentially scale to massive data repositories. This special issue expands some of the topics presented in an Ontology Summit on KGs held in 2020 (Baclawski *et al*, 2020; 2020b).

The modern notion of knowledge graph represents the confluence of several ideas and concepts, each of which has its own long history. All the ideas and concepts originate from the development of symbolic reasoning, physically manifested in the creation of symbols in artwork by humans to convey a story. Humans have been depicting non-figurative (i.e., abstract) artwork for at least 40,000 years and possibly much longer (Aubert et al, 2018). We now delve into three of these streams of thought; namely, Mathematics, Ontology and Linguistics.

Mathematics

Thirty thousand years ago, humans kept track of numerical quantities materially by carving slashes on fragments of bone (Cantlon, 2012). The prehistory of the mode of thought called mathematics includes understanding the first thing that may come to mind, which are the symbols we call "numerical terms." This led to the development of numeracy, the earliest forms of which were variables equated to lengths or measurements. Where formerly the concepts were sketched in material substances for tasks such as keeping track of a flock of sheep, numbers were later expressed as symbols in more durable media. As with many cultural developments, their first occurrence was qualitative rather than quantitative (Struik, 2012). As Adam Smith pointed out, numbers are "the most abstract ideas which the human mind is capable of forming." This development, like many others in mathematics to follow, came only slowly into use.

The evolution of symbolic representations could be understood as exploiting a basic human cognitive ability to make correspondences and use analogies from a concrete experience to a more abstract, modeled domain (Cooke, 2011). By 5,000 BCE there are examples of rudimentary abstraction in the first iconic written numerals (using the cuneiform script) as part of Babylonian culture. Clay tablets from 1800 to 1600 BCE, show that the Babylonians had a knowledge of fractions, algebra, quadratic and cubic equations. There is even evidence of a knowledge of trigonometric functions. (Aaboe, 1991; Robson, 2001)

A thousand years later, there was an abundance of knowledge captured in documents. One of these was the idea of geometry, and the general idea of an abstract proof using logical inferences. Theorems were demonstrably proved by Greek mathematicians starting at around 600 BCE (Boyer, 1968). This work had a major influence on the development of mathematics as a field of study. While the Greeks developed abstractions for geometry and numbers, and depicted geometric notions using graph-like structures, as in Figure 1, the abstract notion of a graph itself is more recent.

The paper written by Leonhard Euler on the Seven Bridges of Königsberg (published in the eighteenth century) is regarded as the first paper in the history of graph theory (Euler, 1736). Since that time, graphs have been used in many domains. One domain that was especially well suited to graphs is chemistry, and it was in the context of chemistry that Sylvester first introduced the word "graph" to mathematics (Sylvester, 1878). Graph theory is now an important branch of mathematics.

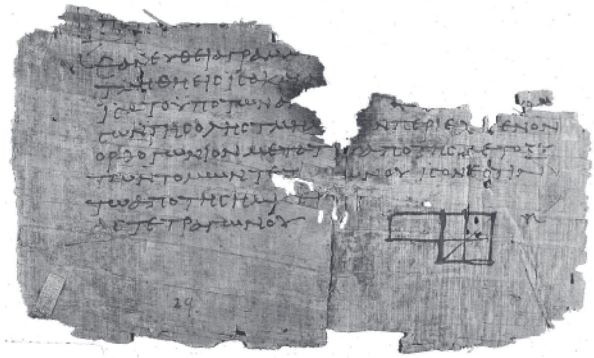

Figure 1: A fragment of Euclid's *Elements*, found at Oxyrhynchus and dated to circa 100 CE

James Sylvester was a colleague of Harvard mathematician Benjamin Peirce, whose son Charles Sanders Peirce developed Existential Graphs (EGs) in 1897. Charles Sanders Peirce and Gottlob Frege independently developed the foundations of what is now known as first-order logic, a notion that has come to dominate modern logic. Peirce's notion of an EG was designed to be fully capable of representing his algebraic notations of first-order and higher-order predicate calculus. This work established that graphs could play a role in the representation of knowledge. The term “knowledge graph” was coined in 1972 (Schneider, 1973).

Ontology

The second stream of ideas is the notion of ontology. In philosophy ontology is the study of concepts such as existence, being, becoming, and reality. The philosophical field of ontology goes back to ancient Indian philosophy in the first millennium BCE (Lochtefeld, 2002; Klostermaier, 2014; Larson, Bhattacharya, and Potter, 2014), and was systematically documented by Aristotle as a major philosophical topic. Aristotle’s theories set the standard for logic and ontology with: stable objects as composites of form and matter; ten categories for analyzing, describing, and classifying anything; and logic for specifying patterns and reasoning about them (Sowa, 2016).

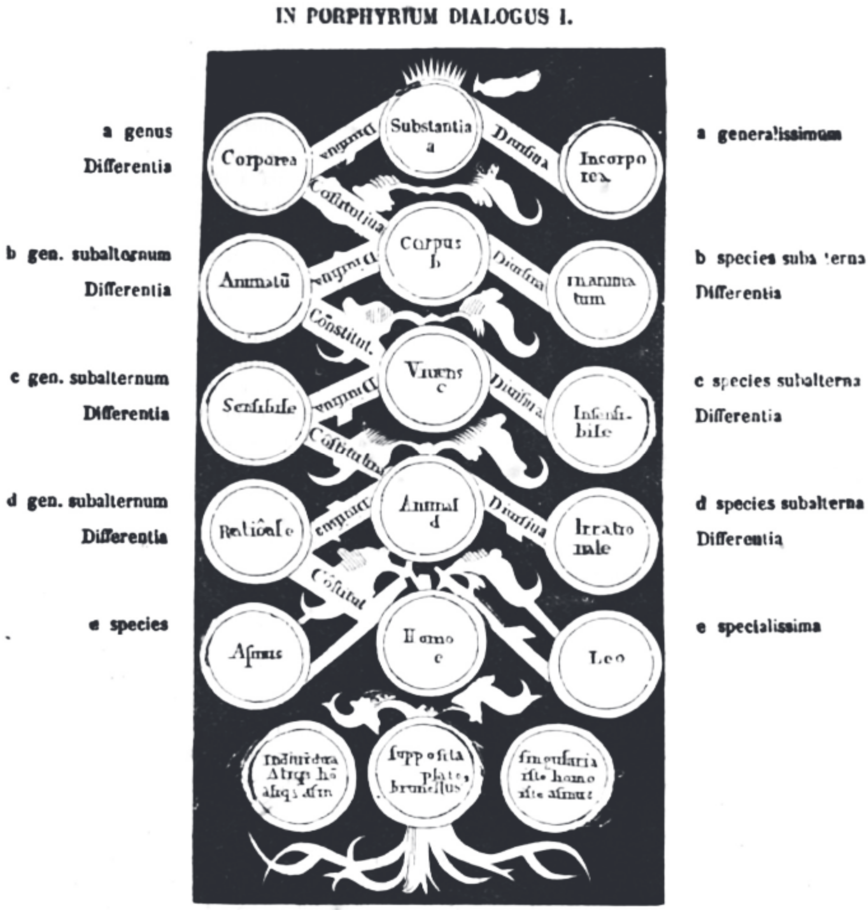

Figure 2: Porphyrian Tree by Boethius

Aristotle's work on ontology led to developments such as the Porphyrian Tree from Porphyry's “Introduction” to Aristotle’s categories in the third century CE. Figure 2 shows the Porphyrian Tree as drawn by Boethius from his translation of Porphyry's treatise in the sixth century. Ramon Llull further developed the Porphyrian Tree into a forest of sixteen trees in the thirteenth century, which set the stage for medieval philosophical-theological developments of logic and the problem of universals. The mathematical and philosophical discussions served as an inspiration to the pioneer in symbolic logic, Gottfried Wilhelm Leibniz, and the concepts in the Porphyrian Tree continue to benefit the classification of living organisms.

However, the use of ontology in modern information processing systems only emerged in the mid-1970s. It was at this time that AI researchers began to recognize that knowledge engineering was necessary for building large and powerful AI systems. By the 1980s, the AI community began to use the term “ontology” for a theory of a modeled world as a component of knowledge-based systems. Such ontologies were (and still are) usually based on classes and relationships between them, such as type hierarchies, part-whole relationships, and many others. Although classes and relationships were visualized using graph-like diagrams to help people to understand the ontology, the knowledge-based systems of the time did not describe themselves as being graphs, and the knowledge that was encoded was not regarded as being a “knowledge graph” per se.

Linguistics and Natural Language Processing

A third stream of thought came from the field of linguistics. The interplay of how humans communicate and naturally represent entities and relationships in the world can provide inputs to a knowledge graph, whether the inputs arise from text, images, sound, video, or any other medium. The small sizes and relative simplicity of the earliest data processing activities could be managed using fixed structures and traditional database technology because the linguistic needs were relatively simple. The advent of tools for entity extraction and Natural Language Processing (NLP) that are now easy to obtain and to deploy in applications has resulted in much larger amounts of information with a more complex interplay among the natural language terms. A more flexible data model is needed to deal with this complexity, which is one possible reason why KGs have become so popular.

While much of the work on NLP has been focused on English, support for other languages is now emerging. Recent efforts at generating KGs and answering queries have been made in Sanskrit (Terdalkar and Bhattacharya, 2019), and there are also efforts in other languages, some of which have been very advanced for some time now (Wang, 2013; Wu, 2018).

Confluence

The confluence of the three streams of ideas discussed above was not a single event. The work of Ramon Llull could be regarded as an early example of a notion of knowledge graph that combines aspects of mathematics and ontology. Some other examples are Peirce's work on triadic logic from the 1860s, which influenced the upper-level category system of the KBPedia Knowledge Ontology (KKO) (KBPedia, 2021), and John Sowa's conceptual graph notion that was inspired by semantic networks in Artificial Intelligence (Sowa, 1976).

One of the most significant developments in the history of knowledge graphs was the development of

the Resource Description Framework (RDF), which was first published in 1997 (Guha and Bray, 1997). Although RDF is a graph-based data model, it was not originally designed for KGs; indeed, RDF was not originally intended for data at all, but rather for metadata annotations (RDF, 2014). There are many query languages for RDF, but one of the most commonly used is SPARQL, whose standard was first released in 2008 (SPARQL, 2013).

The publication of Google's Knowledge Graph in 2012 was a major event in the history of KGs (Singhal, 2012). While Google didn't invent KGs, it was the main popularizer, and there has been steady growth of interest in KG research and development in recent years. The papers in this special issue provide an overview of KG system research and development, as well as some of the many challenges of this field.

Introducing the Papers

In this special issue, the first article discusses the issue of why people and organizations should devote resources to capturing and organizing information at all. Matthew West has had a distinguished career in information management and applied ontology. He is the Technical Lead for the UK Digital Twin programme and was awarded an OBE for services to information management in the 2021 New Years Honours List (UK). Matthew West's article proposes that information matters because it is used to support organizational decisions and, when critical information is sharable and structured, it is the key to enabling automation, thereby improving productivity. Knowledge graph systems, in particular, can contribute to business and information processes.

Having discussed why knowledge graphs should be developed and maintained, we then consider how to develop them. In practice, knowledge graphs, especially very large ones, are not constructed all at once but rather incrementally by an iterative process, adding both new data and new metadata, including deeper semantics. In “Issues in incrementally adding better semantics to Knowledge Graphs,” Gary Berg-Cross discusses the advantages and issues that arise when knowledge graphs are developed incrementally. Gary Berg-Cross is a Cognitive Psychologist who has published extensively in both Cognitive Psychology and Artificial Intelligence. Most recently he has been involved in various aspects of the Semantic Web, including participating in the Spatial Ontology Community of Practice to provide better semantics to support data sharing as well as vertical and horizontal integration, with special emphasis on the Earth Sciences.

Knowledge graphs are a popular technique for data management in spite of a lack of agreement about what knowledge graphs are. One of the accomplishments of the Ontology Summit 2020 was to reach a community consensus by major contributors to the field of knowledge graphs on a precise mathematical definition of a knowledge graph. In “A Knowledge Graph Data Model and Query Language,” Kenneth Baclawski has proposed that the definition of a knowledge graph can be the basis for a new model for data management and retrieval, called the *knowledge graph model*, and he also presents an example of a data language, KGSQL, for this new model. The knowledge graph model has a number of advantages compared with existing graph data models such as RDF and property graphs. KGSQL is syntactically similar to SPARQL, but allows one to take advantage of the new capabilities of the knowledge graph model. Kenneth Baclawski is an Emeritus Associate Professor of Computer Science at Northeastern University.

The special issue ends with a review of the major challenges facing the field of knowledge graph systems in the article “Challenges in the Design, Implementation, Operation and Maintenance of Knowledge Graphs” by Gary Berg-Cross. This article reviews both internal challenges within the KG system lifecycle and external challenges on KG systems. The internal challenges are for designing, populating, refining, deconflicting and maintaining a KG. The external challenges include capturing context, completeness issues, and domain-specific knowledge.

Acknowledgements

We would like to thank Janet Singer, Ram D. Sriram, and Andrea Westerinen for their invaluable help, without which this special issue would not have been possible.

References

Aaboe, A. (1991). "The culture of Babylonia: Babylonian mathematics, astrology, and astronomy". In Boardman, John; Edwards, I. E. S.; Hammond, N. G. L.; Sollberger, E.; Walker, C. B. F. (eds.). *The Assyrian and Babylonian Empires and other States of the Near East, from the Eighth to the Sixth Centuries B.C.* Cambridge University Press. ISBN 0-521-22717-8.

Aubert, M. *et al*. (2018). "Pleistocene cave art from Sulawesi, Indonesia," *Nature* volume 514, pages 223–227.

Baclawski, K., Bennett, M., Berg-Cross, G., Fritzsche, D., Sharma, R., Singer, J., Sowa, J., Sriram, R.D., Whitten, D. (2020). Ontology Summit 2020 Communiqué: Knowledge Graphs. In *Applied Ontology*, 16(2), 229–247.

Baclawski, K., Sowa, J., Sharma, R., Berg-Cross, G., Sriram, R., Singer, J. (2020b). Introduction and Overview. In Ontology Summit 2020: Knowledge Graphs. Retrieved 22 October 2021 from https://go.aws/38Sv50z.

Boyer, C. (1968). *A history of mathematics*. John Wiley & Sons, Inc., New York.

Cantlon, Jessica F. (2012). "Math, monkeys, and the developing brain." In Proceedings *of the National Academy of Sciences* 109.Supplement 1: 10725-10732.

Cooke, Roger L. (2011). The history of mathematics: A brief course. John Wiley & Sons.

Euler, L. (1736). Solutio problematis ad geometriam situs pertinentis. In *Comment. Acad. Sci. U. Petrop* 8, pp. 128-140.

Guha, R.V. and Bray, T. (1997). “Meta Content Framework using XML,” June 6, 1997. Retrieved 22 October 2021 from https://bit.ly/3b4pxTy.

KBpedia (2021). Open Standards, KBpedia. Retrieved 22 October 2021 from http://kbpedia.com/standards/

Klostermaier, K.K. (2014). A Concise Encyclopedia of Hinduism. One-world Publications, Oxford. p. 112. ISBN 978-1-78074-672-2.

Larson, G. J., R. S. Bhattacharya, and K. Potter, eds. (2014). "Samkhya." Pp. 3-11 in The Encyclopedia of Indian Philosophies 4. Princeton University Press. ISBN 978-0-691-60441-1.

Lochtefeld, James (2002), “Nirukta” in The Illustrated Encyclopedia of Hinduism, Vol. 2: N-Z, Rosen Publishing, ISBN 0-8239-2287-1, page 476.

Robson, E. (2001) “Neither Sherlock Holmes nor Babylon: a reassessment of Plimpton 322,” *Historia Math.* 28 (3), p. 202.

RDF (2014). RDF 1.1 Concepts and Abstract Syntax. Retrieved 22 October 2021 from www.w3.org/TR/rdf11-concepts/

Schneider, E.W. (1973). Course Modularization Applied: The Interface System and Its Implications For Sequence Control and Data Analysis. In *Association for the Development of Instructional Systems (ADIS)*, Chicago, Illinois, April 1972.

Singhal, Amit (2012). Introducing the Knowledge Graph: things, not strings. *Official Google Blog.* Retrieved 22 October 2021 from https://bit.ly/3Cjjsi0.

Sowa, J. (1976). “Conceptual Graphs for a Data Base Interface.” In *IBM* Journal of Research and Development 20(4): 336-357.

Sowa. J. (2016) Aristotle’s Patterns Of Logic and Ontology. Retrieved 22 October 2021 from https://bit.ly/2XHn4vu.

SPARQL (2013). SPARQL 1.1 Query Language. Retrieved 22 October 2021 from https://bit.ly/3ppnB0t.

Struik, Dirk J. (2012). A concise history of mathematics. Courier Corporation.

Sylvester, J.J. (1878) Chemistry and Algebra, *Nature*. 17 (432): 284.

Terdalkar, H. and Bhattacharya, A. (2019). Framework for Question-Answering in Sanskrit through Automated Construction of Knowledge Graphs. Retrieved 22 October 2021 from https://bit.ly/3m2clFc.

Wang, Z., Li, J., Wang, Z., Li, S., Li, M., Zhang, D., Si, Y., Liu, Y., Zhang, P., Tang, J. (2013). XLore: A Large-scale English-Chinese Bilingual Knowledge Graph. Retrieved 22 October 2021 from https://bit.ly/3Gbemab.

Wu, T., Qi, G., Li, C., Wang, M. (2018). A Survey of Techniques for Constructing Chinese Knowledge Graphs and Their Applications. In *Sustainability* 10(9). ISSN 2071-1050. Retrieved 22 October from https://bit.ly/2ZdjVDU.