




Toward an Open Knowledge Research Graph

Sören Auer^a and Sanjeet Mann ^b

^aPresenter; ^bRecorder

ABSTRACT

Knowledge graphs facilitate the discovery of information by organizing it into entities and describing the relationships of those entities to each other and to established ontologies. They are popular with search and e-commerce companies and could address the biggest problems in scientific communication, according to Sören Auer of the Technische Informationsbibliothek and Leibniz University of Hannover. In his NASIG vision session, Auer introduced attendees to knowledge graphs and explained how they could make scientific research more discoverable, efficient, and collaborative. Challenges include incentivizing researchers to participate and creating the training data needed to automate the generation of knowledge graphs in all fields of research.

KEYWORDS

Knowledge graph; scholarly communication; Semantic Web; linked data; scientific research; machine learning

Change in the digital world

Thank you to Violeta Ilik and the NASIG Program Planning Committee for inviting me to this conference. I would like to show you where I come from. Leibniz University of Hannover has a castle that belonged to a prince, and next to the castle is the Technische Informationsbibliothek (TIB), responsible for supporting the scientific and technology community in Germany with publications, access, licenses, and digital information services.

Figure 1 is an example of a knowledge graph about TIB. The basic ingredients of a knowledge graph are entities and relationships. We are the library of Leibniz University of Hannover and we are a member of Leibniz Association (a German research association). The university and the association are named after Gottfried Wilhelm Leibniz. He was a philosopher and mathematician, living at maybe one of the last times it was possible to cover all areas of science. He invented how to encode numbers in binary, and he was the namesake of a cookie which appeared at the end of the nineteenth century! At that time it was popular to name cookies after famous people like Mozart. And actually the Leibniz cookies are produced by the Bahlsen company, which is in a lawsuit with Leibniz University because Bahlsen owns the trademark to the name for e-commerce and Leibniz University recently opened an online shop for university merchandise.

I am new to the library world and happy to be here to learn. I had to do a bit of research on serials, and one example I found was mail order catalogs. In Germany the *Quelle* catalog was very popular, and this may be one of the reasons why the Iron Curtain fell. In East Germany it was very difficult to get this catalog, and once you got one, you were excited to see all the products sold in West Germany by mail order: hundreds of watches, for example, whereas in East Germany you would only have one or two watches to order. Another thing I found were road maps; it was difficult to find your way in a car without a map. Phone books were also popular, and every year a new phone book would come out. It looks like many of these serials disappeared. Today you buy products on Amazon and e-commerce websites, and you use Google Maps or navigation software. You do not have to look at thousands of pages of a mail order catalog to search for products that fulfill your interests. The world of publishing and commerce has become more dynamic, now that you can drill down to information in a map or database and access it at

CONTACT Sanjeet Mann  Sanjeet_Mann@redlands.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/wser

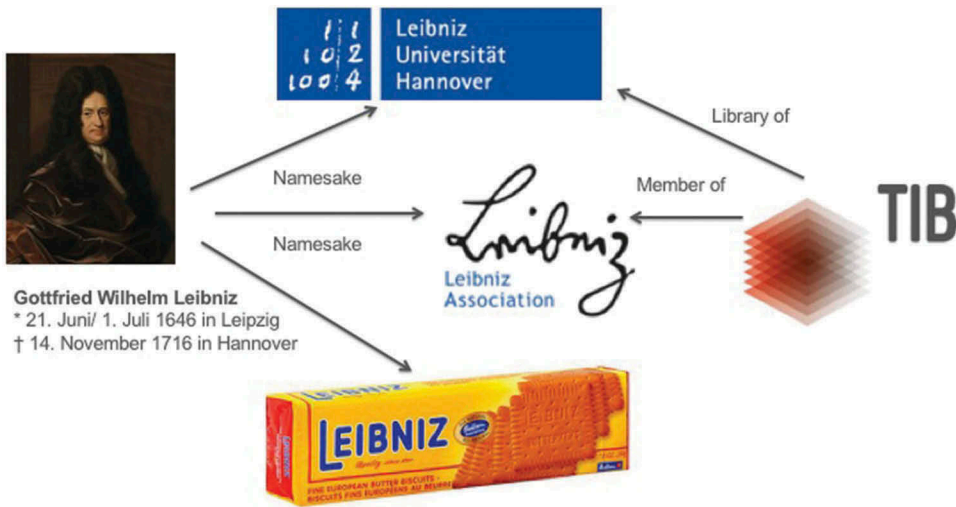


Figure 1. Small knowledge graph illustrating the relationships between Gottfried Wilhelm Leibniz and the organizations and cookies named for him.

your fingertips. Business models have completely changed and new players have appeared. The focus is on interlinking and integrating data from different sources, including crowdsourcing.

What about scientific publishing? It started in 1665 with *Philosophical Transactions of the Royal Society* in the United Kingdom, and publications in the 1860s and 1960s look much the same. Of course, today we have Open Access, identifiers like Open Researcher and Contributor ID (ORCID) and repositories like arXiv.org. But the change is minor compared to other industries. Articles are primarily based on Portable Document Format (PDF) and only partially machine readable. They do not allow embedding of semantics or facilitate interaction and repurposing of information.

Problems with scientific communication

We need more change in scientific communication because science has some serious flaws: the proliferation of scientific literature, a reproducibility crisis, duplication and inefficiency, and a lack of transparency. A report from the National Science Foundation says that scientific literature almost doubled between 2004 and 2014.¹ Last year, China outnumbered the United States as the largest publisher of scientific literature. As researchers we are looking for a needle in a haystack; it is increasingly difficult to make sense of the huge amount of papers that are out there.

This is one of the reasons we have a reproducibility crisis. In an article published in *Nature*, 1,500 scientists were asked to reproduce experiments. Seventy percent failed to reproduce other researchers' experiments, and 50% failed at reproducing their own experiment.² Of course, it varies between disciplines. Computer science is increasing reproducibility by using open source software, publishing the source code, and allowing others to run it.

There is a lot of duplication and inefficiency in science. We do the same thing over and over again and phrase it differently. Sometimes we do not know that someone else has done the same thing. Sometimes we do know, but write it in a different way so it will look different and we will get accepted to the conference!

The root cause of these problems is a lack of transparency. Information is hidden in texts and it's hard to make sense of them or integrate information from different publications. Machine assistance is hardly possible in unstructured PDF documents. You can keyword search through them, but it's not like a product database where you can drill down and find exactly what you are interested in. We identify papers by their Digital Object Identifiers, but you cannot identify the concepts, terminology, and research methods inside

papers because there is not metadata. You need to spend a lot of effort as a researcher reading, building up your own knowledge graph in your mind and then interacting with it.

We also have a collaboration problem, the “one brain barrier.” A small set of researchers can work together well, but beyond that it’s hard to put the pieces together and get an overview of what is already done in a field. You cannot build an engine or a house without exact specifications of the parts and how they fit together. I think you need to do the same with science and research; if you want to build a house of wisdom and knowledge, you need to identify these ingredients so that they fit together like a puzzle. My impression is they do not fit together and it is difficult to make sense of them.

Introduction to knowledge graphs

So how can we fix this? In 1945, Vannevar Bush had an early vision of the Memex, a machine that allows us to store knowledge and interact with it at our fingertips.³ His ideas of how this would be implemented were esoteric at the time, but nowadays we can do something like this using linked data. The basic principles are to use Universal Resource Indicators to identify things with a unique identifier that people and machines can look up on the web. You can return a standardized description for what the identifier means in Resource Description Framework (RDF), including links to related things.⁴

I want to show you how this RDF data model works, because it’s the basis for building a knowledge graph (Figure 2). RDF organizes information in subject–predicate–object triples, similar to how information is organized in sentences: NASIG–organizes–this conference. You can use the subject of one sentence as the object of another: the conference starts at a certain date (June 9) and takes place in a certain place (Atlanta). You can link information from different sources, using namespace prefixes from different vocabularies to make the knowledge graph grow. We see here Atlanta linking to DBpedia, a knowledge base extracted from Wikipedia that contains geocoordinates for pinning it on a map. This is a way of representing and linking knowledge that follows very simple principles. That is its power. Other systems like Extensible Markup Language or relational databases are more complex and restrictive. RDF is simpler for machines: you can throw together triples from different sources to make an integrated knowledge base.

So knowledge graphs allow us to represent information in terms of concepts, classes, properties, relationships, and entities. They use some knowledge representation formalism, such as RDF or a web ontology language. They allow knowledge to be understood holistically from multiple sources or domains.

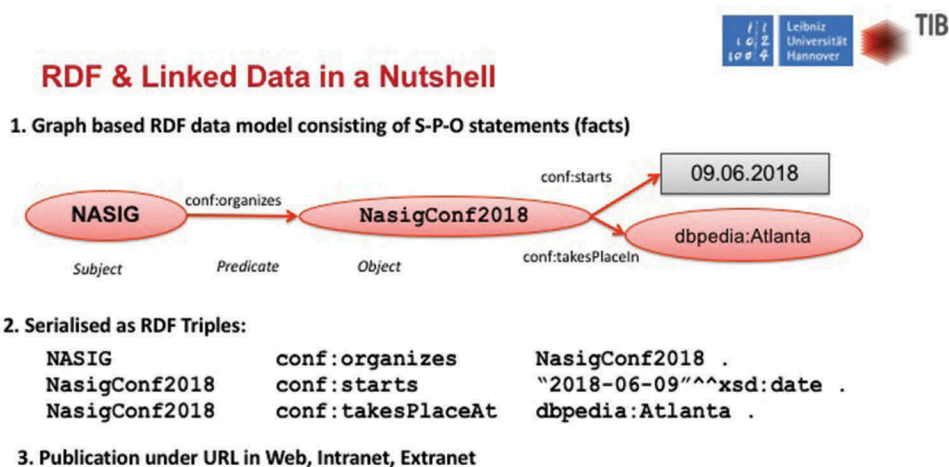


Figure 2. RDF & linked data in a nutshell.

They include metadata (description of data), vocabularies (structure of data), and ground truth (the data themselves). They are flexible, using any open or closed data, aggregated data, schemas, vocabularies, ontologies, or taxonomies. You can even link to external data outside your dataset. A large family of World Wide Web Consortium (W3C) standards supports representing data in RDF and connects to other language ecosystems like Javascript Object Notation System or Hypertext Markup Language, making RDF a kind of lingua franca for data.

Many players in the search and e-commerce area are already using this concept of representing information in a structured way. Google builds a large knowledge graph that it uses to integrate information from different services in the background of search results. When you do a Google search on pandas, for example, there's a fact box on the side that gives you information on the panda from Google's knowledge graph. The schema.org taxonomy used by more than 20% of websites contains markup as RDF triples and integrated metadata, which helps for example in product searching. The life sciences have a lot of structured vocabularies, ontologies, and knowledge bases, such as the National Center for Biomedical Oncology BioPortal and OpenPHACTS. And digital libraries use vocabularies a lot for metadata.

Knowledge graphs applied to scientific communication

My vision is that we build a knowledge graph to represent scientific information, linking data that have been traditionally represented in text to open up research. It would identify granular elements in the document like the research problem, definitions of terms, research methods used, and domain-specific concepts. In mathematics, there are theories, definitions, proofs and methods; in physics, you have experiments, data, and models; in chemistry, substances, structures, and reactions, and so on. Each domain has specific concepts and terminology and we need to make them identifiable and linkable. We need to integrate not only information that is contained in the document, but also other artifacts like videos, abstracts, software code, or research data published in a repository.

I want to illustrate this with an example. Clustered Regularly Interspaced Short Palindromic Repeats is a biochemistry method for genome editing mentioned in over 100,000 documents in Google Scholar. Maybe you are interested in how this method was applied to a particular insect. This is really finding the needle in the haystack! You could add the term "insect," but the publication might not use that term. It will be difficult to find a specific application here. The idea is that we add twenty or thirty RDF statements describing the publication, such as what research problem was addressed, as structured information that can be represented in a knowledge graph and linked between many different publications. If you are interested in Lepidoptera (butterflies), this term can be linked with "insects" and with many different species, and you can quickly identify who published research related to which applied method for which species.

One of the advantages of representing information in a knowledge graph is to get a better overview of different approaches to a problem. Currently what you see here would cost a few weeks or even months of PhD students' time to find, read, and compare hundreds of publications. Once you have a structured representation you can query knowledge graphs with a mouse click. There's a query language for knowledge graphs called SPARQL Protocol and RDF Query Language (SPARQL), standardized by the W3C.

Building a knowledge graph of science also increases the efficiency of scientific research. People use different terminology meaning the same things or use the same term for different things. This is currently very difficult to discover but will become clearer when we represent scientific information in a knowledge graph. The machine readability allows completely new search retrieval assistance techniques. Many of you have used Alexa or Google Now. Imagine if it was possible to ask about research problems in an intuitive way, interacting with the wisdom of science?

We can also reduce duplication and improve interdisciplinarity by making research more visible and accessible. Hopefully there will be less duplication because you can more easily discover what has been done before, with comments from other researchers and peer review by collaborators. Making work more transparent supports open science by giving laypersons and non-academics entry into the research field.

The Open Research Knowledge Graph project

There is a lot of work to do to implement this vision. TIB held a workshop earlier this year to talk about this challenging problem. We will not be able to do it in a fully automated way. There is a lot of talk about machine learning, but for machine learning you need training data, and all we have are unstructured documents. So we need to bring humans into the loop and create good user interfaces for creating training data. You need some way of crowdsourcing this, allowing researchers themselves to add information to the knowledge graph, and link their approaches to other ones. Of course we as librarians need to help them to create this knowledge base. Once we have built up more of this knowledge, maybe it can be used as training data to automate the process in the future.

What's really important is collaboration between scientists, librarians, knowledge engineers, and machines. Computer scientists often think machines can do the job, but then they leave you with 50–60% precision, which does not help. This needs to be an effort of the whole scholarly communication community. It would be great to stay in touch and get some of you involved with this. We want to start an open source software project to create the prototype for a research knowledge graph. We set up a mailing list and soon more information about the Open Research Knowledge Graph will be available at the domain orkg.org. It should be a community effort in the spirit of open source, Open Access, and open knowledge, because there are already efforts in this direction from commercial players like ResearchGate who do their knowledge graphs for organizing scholarly communications. I think it is important for us as librarians and researchers to offer an open alternative and a little bit of competition in this field. Thank you very much for your attention and I am happy to discuss any questions you might have.

Questions and answers

Question: I am with a vendor, and we are increasingly receiving requests from librarians and patrons of our databases to provide tools and datasets for Text and Data Mining (TDM). As I listened to your talk today, I heard many similarities between your knowledge graph and the goals of TDM. Could you describe the potential synergies of what you are doing?

Answer: This is a very good question. TDM is important, but I am skeptical that it will help build the knowledge graph because the precision is not high enough. Entity recognition has maybe reached 70 to 90% for metadata, but inside the document it will be much lower. And even 90% might not be high enough. If 10% of the knowledge graph is wrong, people might be quite annoyed using it. Machines are only smart when they have good training data created by humans. In translation software, the training data were created by human translators, and only after we had a lot of data did the machine translation start to work well. We need a lot of transcribed audio content to train a machine to do voice recognition. That's why I think we need crowdsourcing similar to Open Street Maps, which created a map of the world from volunteer contributions.

Question: How is this financially possible? Human labor is expensive, and libraries are already financially strapped. We do not have enough staff to handle the volume of information we have now, and, as you showed, information is increasing rapidly. Crowdsourcing might be one solution, but I have seen crowdsourcing projects that failed.

Answer: Yes, that is a difficult question. Maybe we need to do things differently. We spend a lot of effort on cataloging, on basically assigning keywords to publications. Like text mining, assigning keywords does not give you a big boost in usability and machine assistance. Maybe we need to collaborate more on services, pooling resources we have and involving researchers and scientists in some way. ORCID, ArXiv, and DataCite are good examples of collaborative services. If crowdsourcing does not work in the first attempt, maybe we need to try several times until we find the right way of doing it.

Question: How does the involvement of the commercial sector in some of these open projects help, or maybe disrupt, them? Over the years, we saw Mendeley bought by a commercial entity and GitHub has just announced that a commercial entity is going to buy it.

Answer: I think it can work. Some open initiatives are structured so they cannot be taken over easily by a commercial entity. Maybe we need that here, not because we are against commercial use, but because “open” can also be used by commercial players. There is a commercial ecosystem of apps and services around Open Street Maps. Almost every mobile phone, whether Android or iPhone, uses an open source core. Ninety percent of web servers are driven by Linux and companies like IBM are building on top of it. Both worlds can work together, and this competition will be healthy. We need more competition; the commercial players are dominating scholarly communication.

Question: Do you think with a knowledge graph that there’s a risk to creating higher barriers to scholarly communication? The system we have now is inefficient and redundant, but it’s easier to get an “out there” idea. Wouldn’t this system create more benefits for something that already has more identifiers established?

Answer: Knowledge graphs cover factual knowledge well, but in research it takes time for facts to emerge as the scientific discourse evolves. How do we represent that research communities do not always agree on things? We need to find the right balance, to make it on the one hand simple to contribute to and access the research graph, and on the other hand represent the emergence of concepts and the possibility of disagreement. I want to use an iterative approach starting with small examples in one domain. You have to start somewhere.

Question: There are parallels between your project and digitizing back issues of print journals, which is a tremendous amount of work. Will you start with newly formed texts and then work backwards?

Answer: I think so. Maybe in the future you can represent your contributions right away in a knowledge graph or publish smaller amounts of content and link them together. One area of difficulty is that incentive systems are currently based on citations of publications. How can you measure the impact a researcher made on his field with his knowledge graph contributions? Only then can you shift from traditional publication to knowledge graph contributions.

Question: You are talking primarily about scientific information. How do you see moving to social sciences or humanities research, which is very different?

Answer: You are right, I would not start with humanities! Social sciences have a lot of data and methodologies that can be represented as structured information. Different domains will have a different pace of using knowledge graphs, but even in the humanities it could be interesting to describe and link the terminology that is used.

Question: In the Open Access and open data movements, funder and institution-level mandates place some of the labor on authors and researchers, and they have an Office of Research Support or library that ensures compliance. Do you think any of the labor for knowledge graphs can be pushed off to authors or an editorial process?

Answer: Yes, we want to build a widget so authors can create a snippet of the knowledge graph describing their research during the submission process. In order for researchers to adopt this, we need to give them a direct benefit. For example, you could find related works, identify who has done similar work, or increase your own visibility. Librarians can also help these communities to create their taxonomies.

Question: A lot of us in the commercial sector have our own internal knowledge graphing projects. We get excited about the technology and build up amazing numbers of triples, but I do not know if people have thought deeply about what kind of questions we need this giant corpus of knowledge to answer.

Answer: Your standard questions are to define a scientific problem, find out who else is working on it, and compare the approaches for addressing it. That is extremely cumbersome work, but we could change that if the knowledge graph could go deeper into the document. Currently we create hypotheses based on our intuition, but in a knowledge graph you could discover new relationships and analyze them in a more automated, more assisted way. This could be a huge benefit for science. Of course, checking would still be part of research, so we will not become obsolete, but we would have more variety of hypotheses to start from.

Question: When you spoke about Vannevar Bush, it reminded me of his contemporary Paul Otlet, who tried to organize the world's knowledge onto three-by-five inch cards. In many ways it seems like you are trying to achieve a similar goal.

Answer: We have the possibility with the Internet to truly collaborate on organizing knowledge. Knowledge graphs will not work with one or five or even a team of ten people, but only if thousands or millions of people work together, following the example of Wikidata, a knowledge graph that is built on Wikipedia. Look at Open Street Maps. It's not just a map, but hundreds of maps for accessibility, bicycles, catastrophe prevention. It works because it has a simple data model, just RDF key-value pairs that can be attached to any line, point, or geometric object with flexible annotations to create completely new representations. We need to do something similar in science in a crowdsourced way. I guess we are at the end of our time. Thank you very much, it was a really great discussion! [applause]

Notes

1. Karen E. White, Carol Robbins, Beethika Khan, and Christina Feynman, "Science and Engineering Publication Output Trends: 2014 Shows Rise of Developing Country Output while Developed Countries Dominate Highly Cited Publications," *NCSES InfoBrief*, October 2017. <https://www.nsf.gov/statistics/2018/nsf18300/nsf18300.pdf> (accessed June 28, 2018).
2. Mona Baker, "1,500 Scientists Lift the Lid on Reproducibility," *Nature* 533, no. 7604 (May 25, 2016). <https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970> (accessed June 28, 2018).
3. Vannevar Bush, "As We May Think," *Atlantic Monthly* (July 1945). <https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881> (accessed July 20, 2018).
4. Sören Auer, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, and Amrapali Zaveri, "Introduction to Linked Data and its Lifecycle on the Web," in *Reasoning Web. Semantic Technologies for Intelligent Data Analysis*, eds. Sebastian Rudolph, Georg Gottlob, Ian Horrocks and Frank van Harmelen (Berlin: Springer, 2013), 1-90.

Notes on contributors

Sören Auer is Director of the Technische Informationsbibliothek (TIB) and Professor of Data Science and Digital Libraries at Leibniz University of Hannover.

Sanjeet Mann is Interim Assistant Director and Arts & Systems Librarian at Armacost Library, University of Redlands.

ORCID

Sanjeet Mann  <http://orcid.org/0000-0003-4442-1053>