# The importance of knowledge bases for artificial intelligence in science

K. Forbus, Northwestern University, United States

## Introduction

For artificial intelligence (AI) systems to increase the productivity of science, they need to understand both the domains of science they are operating in, and the world in which that domain is embedded. In other words, they need knowledge bases that provide such information in explicit and verifiable forms to support reasoning that includes transparent explanations for their conclusions. This essay explains the idea of knowledge bases and knowledge graphs, summarising the state of the art and the improvements needed to support broader uses of AI in science. These improvements include commonsense knowledge to tie scientific concepts to the everyday world and to provide common ground for communication with human partners; expressive representations for encoding scientific knowledge; and robust reasoning techniques that go beyond simple retrieval. Research could work towards an open knowledge network to provide a community resource that supports re-use, replication and dissemination.

Knowledge is a hallmark of human intelligence, and a key goal of science is to generate replicable knowledge. AI systems with enough shared knowledge to reason with, and learn from, human partners could lead to revolutionary advances in science (Gil et al., 2018; Kitano, 2021). In AI, the term "knowledge base" is commonly used to refer to a system's knowledge.[1]

As this section explains, there are multiple kinds of knowledge. For some types, the commercial world has already deployed knowledge bases with billions of facts to support web search and simple forms of question answering. However, for several other kinds of knowledge – including some relevant for using AI to accelerate science – progress has been slow, despite the potential value. A US report arguing for the construction of an open knowledge network has this conclusion:

Artificial intelligence, machine learning, natural language technologies, and robotics are all driving innovation in information systems. Developing the knowledge bases, graphs, and networks that lie at the heart of these systems is expensive and tends to be domain-specific, and the largest currently are focused on consumer products (e.g. for web search, advertising placement, and question answering). An open and broad community effort to develop a national-scale data infrastructure – an Open Knowledge Network – would distribute the development expense, be accessible to a broad group of stakeholders, and be domain-agnostic. This infrastructure has the potential to drive innovation across medicine, science, engineering, and finance, and achieve a new round of explosive scientific and economic growth not seen since the adoption of the Internet. (NTSC, 2018)

This essay explains knowledge bases and knowledge graphs, what needs to be done and how it might be accomplished.

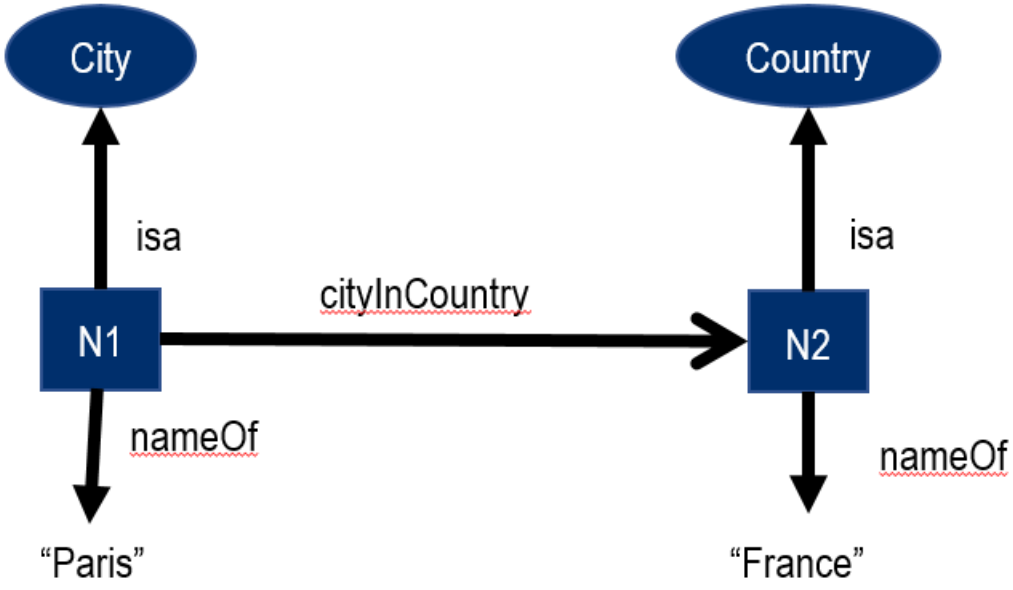
## Knowledge bases and knowledge graphs

Knowledge graphs – symbolic structures that express properties of entities and relationships among them – are the most common form of knowledge bases. Entities are represented by nodes in the graph, while labelled arcs specify their properties and relationships. For example, the sentence "Paris is a city in the country of France" might be represented inside a knowledge base via the following facts:

isa(N1, City)
isa(N2, Country)
nameOf(N1,"Paris")
nameOf(N2, "France")
cityInCountry(N1,N2)

Each of these statements is a logical form expressing that a relationship (the predicate) holds between its arguments. For example, in the first statement, the predicate "isa" means that the entity provided as the first argument (e.g. N1, where "N" refers to a node) is an instance of the concept provided by the second argument (e.g. City). Similarly, the predicate "nameOf" indicates that the string given as the second argument should be used as the name for the entity given as the first argument. Meanwhile, "cityInCountry" indicates the first argument is a city geographically located within the country given as the second argument. An equivalent graphical representation of these facts is given in Figure 1:

### Figure 1. Graphical representation of a knowledge base




Such symbolic representations are crucial in knowledge bases for three reasons. First, it is important to be able to vet a system's knowledge. Symbolic representations can be read by people, given the appropriate tools, as suggested by the example above. Second, such representations support reasoning. Third, they support transparent explanations, i.e. explanations that show how a system came to its conclusions.

While some researchers have tried to use "large language models" as knowledge bases (e.g. Petroni et al., 2019), such efforts have not been promising to date. Large language models fail to provide reliable ways to vet what the system knows, they do not provide accurate reasoning (Marcus and Davis, 2020), and they cannot provide transparency in their operations. Moreover, large language models have serious problems

with fairness and equity since the biases of the materials they are trained upon shine through in their application (Bender et al., 2021; Lin, Hilton and Evans, 2021). Consequently, the rest of the essay focuses on knowledge graphs.

## Knowledge graphs today

Commercial knowledge graphs used in web searches (like Microsoft's Satori and Google's Knowledge Graph) contain billions of nodes, each woven into a network by even more links connecting them (Noy et al., 2019). Similarly, Amazon and other companies use knowledge graphs to represent product information to make better recommendations to customers. These large-scale knowledge graphs are constructed by a combination of manual labour (including both highly trained professionals and crowdsourcing) and automatic techniques. A great deal of engineering goes into efficient large-scale retrieval and inference using such knowledge graphs.

Most commercial knowledge graphs focus on encoding specifics about entities in the world (such as product and customer information). Similarly, some knowledge graphs have been built by extracting knowledge from textual resources, such as Wikipedia. While these knowledge graphs are useful for some kinds of factual question answering, they lack several kinds of knowledge needed to build knowledge graphs for scientific research. For example, inferential knowledge, i.e. the kinds of rules used to infer things, is missing.

Some knowledge bases have such rules as well. For example, the Cyc knowledge base (e.g. Lenat et al., 2010) is designed to support a wider variety of reasoning, via rules and more complex types of statements. For example, when asked if the planet Earth can run a marathon, it can deduce that the Earth cannot because this requires being a living thing, and the Earth, as a planet, isn't a living thing. This example illustrates an insight that Cyc researchers came to long ago: a surprising amount of commonsense knowledge isn't in explicit resources like encyclopedias. Such tacit knowledge, discussed below, also lies in the things one must know to be able to read an encyclopedia.

There have been several efforts to build knowledge bases for particular areas of science. These have been driven by the need to improve literature searches and to archive community information and expertise (e.g. genomic data, workflows). Like commercial knowledge graphs, these efforts have been facilitated by the widespread use of Semantic Web protocols. These enable the same graphs to be used with different software implementation platforms. While use of AI in science is promising, research efforts lack the breadth of expressiveness and support for inference that will be needed to fully realise this potential.

## What is missing?

The impressive scale and utility of commercial knowledge graphs suggest that large-scale knowledge graphs for science are possible. However, additional research is needed in at least three areas: commonsense knowledge, professional knowledge and complex reasoning at scale. Each is discussed in turn.

### *Commonsense knowledge*

Why commonsense knowledge? Scientific theories rest on tacit knowledge shared by all scientists due to their experience as people in the physical, social and mental worlds. Examples include the billiard ball model of gases and the lava lamp model of convection. To understand their human partners, AI systems for science need to share this common ground to some reasonable degree. Some of this tacit knowledge can be captured by qualitative representations, which provide human-like descriptions of quantities, space, causality and processes (Forbus, 2019). Other aspects require multimodal grounding, e.g. what particular

objects and systems look like. The role of experiential knowledge in commonsense remains an open research question: is most commonsense knowledge encoded via general rules, or do we mostly reason by analogy from experience?

### *Professional knowledge*

Professional knowledge also raises important challenges. Highly expressive representations are needed to encode scientific theories. For instance, in addition to the theories themselves, how they are operationalised (by connecting the professional concepts to the everyday world) must be represented as well. In other words, knowledge must be represented to support the process of model formulation (Forbus, 2019).

Today's AI systems for science and engineering factor out model formulation by focusing on tasks and/or domains. The broader the scientific reasoning, the more tacit knowledge needs to be incorporated into the system. Representation schemes need to support explicit contexts, e.g. to represent competing theories and reason about the range of applicability for a theory. For example, when does one need to use classical, relativistic mechanics, versus quantum mechanics?

### *Complex reasoning at scale*

No AI system comes close to the flexibility of human reasoning. Consider, for example, constructing or even just understanding thought experiments. When Einstein imagined travelling on a beam of light, it required reasoning through novel (and sometimes impossible) situations. AI systems cannot yet do this in a general way.

In specific domains, such as software verification, reasoning systems can far outstrip human capabilities. However, such systems cannot be taught to operate in a new domain without reprogramming. For specific scientific tasks and domains, special-purpose high-performance reasoning systems could likely provide important benefits. In the longer term, improved AI reasoning flexibility could bring it closer to that of people. This will enable AI to move towards being a collaborator in science, as well as a tool.

## Towards knowledge bases for science: What might be done?

The commercial world is unlikely to build broad commonsense knowledge bases. Firms mostly do not care about scientific knowledge. While a large-scale high-quality commonsense knowledge graph would benefit everyone, the effort needed to build one is beyond the usual research horizons of the private sector.[2]

Partly for this reason, the idea of an open knowledge network is gaining traction in the research community. Open licensing, such as Creative Commons Attribution Only, matters. For example, the Suggested Upper Merged Ontology – intended as a foundation for a variety of computer information processing systems – provides some of the most abstract layers of an ontology (SUMO, 2022). However, it uses a licence antithetical to most commercial applications. Moreover, modules that extend it include a hodgepodge of licences that make it difficult to build on top of.[3] To maximise utility to the scientific community, in terms of reusability, replicability and dissemination, funding is needed for the construction of open knowledge graphs.

Aiming for generality is hard, and it is tempting to go straight to applications. To date, for example, the US National Science Foundation has only funded a handful of projects in open knowledge networks. All of them focused on specific domains and tasks (e.g. finding relevant legal documents, reasoning about flooding) (NSF, 2022).

A mixed approach might yield even better results. Teams of AI scientists and scientists from other domains might collaborate on knowledge graphs for fields that include both professional knowledge and relevant

commonsense knowledge. In biology, for example, efforts could focus beyond biochemistry or genetics to produce everyday knowledge about animals and plants that connects professional concepts to the everyday world. Other efforts should use community testbeds where commonsense reasoning is needed, e.g. robotics (including simulated worlds) for some kinds of commonsense knowledge and story understanding for others.

Each effort would be required to draw upon the growing shared knowledge graph. There will be competing and complementary approaches to concepts, but that is fine if the underlying graph infrastructure supports multiple contexts.[4] The result will be a federation of knowledge graphs, ideally continually updated as research progresses and eventually encompassing all scientific knowledge.

Cognitive science and social sciences should be included, along with physical and biological sciences, in building an open knowledge network. This is important because progress in these areas will help guide AI towards more trustable, human-like algorithms and conceptual structures.

## Conclusion

Progress in using AI to accelerate science will need to include efforts to build large-scale knowledge graphs of commonsense knowledge, professional knowledge and the bridges between them. This will not be done by the commercial world because it is not directly related to their everyday concerns. However, they too would benefit from aspects of it. Building an open knowledge network can improve the re-use, replicability and dissemination of scientific knowledge. This will require a long-term, large-scale community effort. However, with the right mix of projects, incremental results of value along the way will also serve to guide further efforts.

## References

Bender, E. et al. (2021), “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610-623, https://doi.org/10.1145/3442188.3445922.

Carroll, J.J. et al. (2005), “Named graphs”, *Journal of Web Semantics*, Vol. 3/4, pp. 247-267, https://doi.org/10.1016/j.websem.2005.09.001.

Forbus, K. (2019), *Qualitative Representations: How People Reason and Learn about the Continuous World,* MIT Press, Cambridge, MA.

Gil, Y. et al. (2018), “Intelligent systems for geosciences: An essential research agenda”, Communications *of the ACM,* Vol. 62/1, pp. 76-84, https://doi.org/10.1145/3192335.

Kitano, H. (2021), “Nobel Turing challenge: Creating the engine for scientific discovery”, *Nature Systems Biology and Applications*, Vol. 7/29, https://doi.org/10.1038/s41540-021-00189-3.

Lenat, D. et al. (2010), “Harnessing cyc to answer clinical researchers’ ad hoc queries”, *AI Magazine*, Vol. 31/3, pp. 13-32, http://dx.doi.org/10.1609/aimag.v31i3.2299.

Lin, S., J. Hilton and O. Evans (2021), “TruthfulQA: Measuring how models mimic human falsehoods”, *arXiv*, arXiv:2109.07958v1, https://doi.org/10.48550/arXiv.2109.07958.

Marcus, G. and E. Davis (2020), “GPT-3, Bloviator: OpenAI’s language generator has no idea what it is talking about”, 22 August, *MIT Technology Review*, https://www.technologyreview.com.

Noy, N. et al. (2019), “Industry-scale knowledge graphs: Lessons and challenges”, *ACM Queue*, Vol. 12/2, https://queue.acm.org/detail.cfm?id=3332266.

NSF (2022), “Convergence Accelerator Portfolio”, webpage, https://beta.nsf.gov/funding/initiatives/convergence-accelerator/portfolio (accessed 28 January 2022).

NSTC (2018), Open Knowledge Network: Summary of the Big Data IWG Workshop, October 4-5, 2017, National Science and Technology Council, Washington, DC.

Petroni, F. et al. (2019), "Language models as knowledge bases?", *arXiv,* arXiv:1909.01066v2, https://doi.org/10.48550/arXiv.1909.01066.

SUMO (2022), Suggested Upper Merged Ontology website, https://www.ontologyportal.org (accessed 23 November 2022).

## Notes

[1] This is quite different from a common informal use of the term knowledge base to mean a collection of documents for some purpose.

[2] Cycorp, founded to continue building the Cyc knowledge base, is an exception. It is funded by a mixture of commercial and government projects. Unfortunately, its proprietary nature makes it less useful for some purposes, e.g. researchers cannot freely distribute it along with their source code to ensure the replicability of their research.

[3] Personal communication. By contrast, the OpenCyc ontology, also by Cycorp, is available under a CC-Attribution Only licence. Thus, others can build up on it (e.g. www.qrg.northwestern.edu/nextkb/index.html).

[4] Any large-scale knowledge base needs to use contexts in any case, to handle culture-specific knowledge, works of fiction, and alternate explanations and theories. The OpenCyc ontology uses "microtheories" for this purpose, and a similar mechanism is found in the Resource Description Framework (RDF) semantic web representation in the form of "named graphs" (Carroll et al., 2005).