KULA
Knowledge Creation, Dissemination, and Preservation Studies

Semantic Encyclopedias and Boolean Dreams

Alexandra Provo

Volume 6, numéro 3, 2022

Metadata as Knowledge

URI : https://id.erudit.org/iderudit/1091370ar
DOI : https://doi.org/10.18357/kula.155

Aller au sommaire du numéro

Éditeur(s)

University of Victoria Libraries

ISSN

2398-4112 (numérique)

Découvrir la revue

Citer cet article

Provo, A. (2022). Semantic Encyclopedias and Boolean Dreams. *KULA*, *6*(3), 1–15. https://doi.org/10.18357/kula.155

Résumé de l'article

When metadata becomes knowledge, opportunities for multiplicity and risks of harm and exclusion arise. As GLAM institutions contribute to the Semantic Web, we must pay attention to the implications of participation. While the Semantic Web grew out of the flourishing of web technologies in the 1990s, recognizing its roots in classical/symbolic AI (referred to as Good Old Fashioned Artificial Intelligence, or GOFAI)—in particular, expert systems and knowledge representation—encourages critical questions like: which problems from knowledge representation and expert systems does the Semantic Web inherit? Are GOFAI failures really failures, or does the gap between rhetoric and practice point to generative possibilities (some of which can now be seen in Semantic Web initiatives)? What can we learn from AI critics, feminist approaches, and the unmasking of encyclopedic neutrality? This research article will explore how critiques of AI expert systems and Cyc, an ongoing project to create a common sense knowledge base, might apply to Semantic Web efforts like Wikipedia, Wikidata, DBpedia, and Schema.org.

KULA KNOWLEDGE CREATION, DISSEMINATION, AND PRESERVATION STUDIES

Provo, Alexandra. 2022. Semantic Encyclopedias and Boolean Dreams. *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 6(3). https://doi.org/10.18357/kula.155

RESEARCH ARTICLE

Semantic Encyclopedias and Boolean Dreams

Alexandra Provo
New York University

When metadata becomes knowledge, opportunities for multiplicity and risks of harm and exclusion arise. As GLAM institutions contribute to the Semantic Web, we must pay attention to the implications of participation. While the Semantic Web grew out of the flourishing of web technologies in the 1990s, recognizing its roots in classical/symbolic AI (referred to as Good Old Fashioned Artificial Intelligence, or GOFAI)—in particular, expert systems and knowledge representation—encourages critical questions like: which problems from knowledge representation and expert systems does the Semantic Web inherit? Are GOFAI failures really failures, or does the gap between rhetoric and practice point to generative possibilities (some of which can now be seen in Semantic Web initiatives)? What can we learn from AI critics, feminist approaches, and the unmasking of encyclopedic neutrality? This research article will explore how critiques of AI expert systems and Cyc, an ongoing project to create a common sense knowledge base, might apply to Semantic Web efforts like Wikipedia, Wikidata, DBpedia, and Schema.org.

Keywords: Semantic Web; knowledge representation; Cyc; Wikipedia; Wikidata; Schema.org; DBpedia; GOFAI; feminist epistemology

Introduction

In "How We Construct Subjects," Hope Olson (2008) critiques the Aristotelian inheritance and hierarchical structure of controlled vocabularies and library classification. Her feminist alternative vision is one of connected knowing that emphasizes non-dominant ways of knowing such as women's ways. When I first read Olson's article, I thought of linked data and its graph structure as a possible embodiment of networked/connected knowing. However, semantic networks and graph structures come with their own philosophical, socio-cultural, and historical baggage. Unpacking this baggage and applying critiques of knowledge representation in artificial intelligence (AI) can help illuminate some of both the advantages and the possible pitfalls of today's linked data and Semantic Web initiatives. Metadata and structured data in the Semantic Web shift library cataloging practice from resource description to entity management. With this comes a shift in emphasis from information deployed for specific search functions to the creation of encoded knowledge. When metadata becomes knowledge, opportunities for multiplicity in perspectives and risks of harmful representation arise. As galleries, libraries, archives, and museums (GLAM) contribute to the Semantic Web, we must pay attention to the implications of participation.

This article starts with the idea that linked data and the Semantic Web share formal and design characteristics, key players, and technical architects with classical/symbolic AI (Good Old Fashioned Artificial Intelligence, or GOFAI), expert systems, and the Cyc project. The lineage connecting GOFAI and the Semantic Web is well recognized in Semantic Web scholarship (Wilks 2008; Halpin 2013; Sparck Jones 2004; Verborgh and Vander Sande 2020), although it may be less obvious to practitioners (especially those newer to linked data and the Semantic Web).[1] I argue that some critiques of Cyc apply to the Semantic Web, although Semantic Web projects do not suffer from all of Cyc's issues: some issues are resolved in Semantic Web projects, and some do not apply because the Semantic Web's scope and purpose is different from that of GOFAI.

[1] I am grateful to Professor Joseph Lemelin for teaching me about symbolic AI (GOFAI), Cyc, and other aspects of the history and philosophy of AI in his NYU course, Computation and Critique. This article was developed based on the term paper I wrote for his course in fall 2020 and has benefited greatly from his generous and attentive comments.

I contend that by paying attention to these points of convergence and divergence, we can recognize the possibilities and limitations of linked data initiatives. Such recognition can help practitioners avoid slipping into what Douglas Hofstadter (cited in Ekbia 2008, 100) termed a "Boolean dream."

In the pages that follow, I write about Wikipedia and Wikidata as a beginner editor and edit-a-thon event organizer, a sometimes user of Schema.org, and a practitioner who has worked on small-scale cultural heritage linked open data projects. I am most concerned with the discourse surrounding Cyc and expert systems and will introduce technical details in service of illuminating this discourse.

Knowledge Representation's Roots in GOFAI

Classical or symbolic artificial intelligence—often referred to as "Good Old Fashioned Artificial Intelligence" or GOFAI (Haugeland 1985, 112)—makes use of "programmed instructions operating on formal symbolic representations" (Boden 2014, 89). This branch of AI was the prominent approach from the 1950s to the mid-1980s, in contrast to today's focus on the connectionist/neural network framework (seen in machine learning approaches). GOFAI grows out of the theory of physical symbol systems as defined by Allen Newell and Herbert Simon (1976). According to Newell and Simon, in a physical symbol system, physical patterns (symbols in a computer program) are arranged into structures called expressions. The symbols therein are related to one another in some way and are acted upon by processes: creation, modification, reproduction, and destruction. Key features of the physical symbol system are designation (i.e., an expression points to an object that can affect or be affected by the system) and interpretation (i.e., the system can carry out a process based on an expression that designates the process).

Though AI practitioners like Newell and Simon seemed focused on engineering problems and not philosophy, AI critics often remark on the rationalist philosophical underpinnings of GOFAI. For example, when he coined the term "GOFAI," philosopher John Haugeland asserted that it adopts "Hobbes's idea that ratiocination is computation" (1985, 112). Longstanding AI critic Hubert Dreyfus traces GOFAI's deep philosophical heritage in his reflections on his time evaluating Newell and Simon's work at RAND in 1963:

> As I studied the RAND papers and memos, I found to my surprise that, far from replacing philosophy, the pioneers in CS [Cognitive Simulation] had learned a lot, directly and indirectly from the philosophers. They had taken over Hobbes' claim that reasoning was calculating, Descartes' idea that the mind manipulated mental representations, Leibniz's search for of a ''universal characteristic''—a set of primitives in which all knowledge could be expressed—, Kant's claim that concepts were rules, Frege's claim that we could formalize those rules, and Russell's postulation of logical atoms as the building blocks of reality. In short, without realizing it, AI researchers were hard at work finding the features, rules, and representations needed for turning rationalist philosophy into a research program. (Dreyfus 2012, 89)

Brian Cantwell Smith points to four "vaguely Cartesian assumptions" underlying GOFAI: 1) that "the essence of intelligence is thought," 2) that "ideal model of thought is logical inference," 3) that "perception is at a lower level than thought," and 4) that the "ontology of the world is . . . formal: discrete, well-defined, mesoscale objects exemplifying properties and standing in unambiguous relations" (Smith 2019, 7). According to Smith, these assumptions are not inherently computational but instead connect to computational approaches by way of insights from Boole, Peirce, and Frege, explained by Smith as four core principles of a system: 1) it "works, mechanically, in ways explicable by science," 2) it "supports semantic interpretation—that is, can be taken to be about (mean, represent, etc.) facts and situations in the outside world," 3) it is "normatively assessed or governed in terms of the semantic interpretation," or can be evaluated as to "whether they are right or wrong, true or false, useful or useless," and 4) its "semantic relations to the world (including reference) are not effective" (2019, 9–12). Smith writes that these principles underlie formal logic and computing and that "in the GOFAI era, it was assumed that the evident way to implement [these principles] was to build an interconnected network of discrete symbols or data structures, organized in roughly propositional form, operating according to a regimen dictated by an equally symbolic program" (2019, 20). GOFAI can thus be understood as an approach to AI grounded in explicit symbolic representation drawing on rationalist philosophical traditions.

A physical symbol system makes use of explicit statements and rules to carry out processes. Such statements and rules might encode instructions for carrying out a task, or they might encode knowledge. The latter role is of interest for this article. In the 1970s and 1980s, expert systems encoded knowledge related to specific domains such as medicine or chemistry. As anthropologist Diana Forsythe defines them, expert systems are

> intended to automate decision-making processes normally undertaken by a given human "expert" by capturing and coding in machine-readable form the background knowledge and rules of thumb

> ("heuristics") used by the expert to make decisions in a particular subject area (or "domain"). This information is encoded in the system's "knowledge base", which is then manipulated by the system's "inference engine" in order to reach conclusions relating to the tasks at hand. (1993, 451)

As Harry Halpin explains, a knowledge representation language is "a language whose primary purpose is the representation of non-digital content in a digital encoding" (2013, 51). Earlier GOFAI attempts at knowledge representation focused on first-order logic/predicate calculus, while later efforts attempted to address some of the limitations of first-order logic by exploring semantic networks and frame-based systems with slots and values (Halpin 2013, 53–55).

Cyc and the Semantic Web

Cyc (a play on the word *encyclopedia*) is an ongoing large-scale knowledge base project that aims to encode "common sense." Begun in 1984 by Douglas Lenat with key development by Ramanathan V. Guha and Edward Feigenbaum, among others, Cyc first developed a frame-based language called CycL and later made use of a higher-order logical language (Foxvog 2010, 259). In the late 1980s and early 1990s, in the midst of the so-called "AI Winter," Cyc was thought of as a flagship knowledge representation project carrying the standard of GOFAI (Adam 1998, 67). A database of explicit representations connected to one another by axioms and other kinds of relationships (in other words, a machine-actionable ontology), Cyc belongs to the tradition of physical symbol systems as defined by Newell and Simon. Originally founded as a research project within the Microelectronics and Computer Technology Corporation (MCC) and projected to take thirty years (with its first phase concluding in 1994), Cyc continues on today and is now a private commercial project of Cycorp.

In their late 1980s/early 1990s paper "On the Thresholds of Knowledge,"[2] Douglas Lenat and Edward Feigenbaum follow in the rhetorical footsteps of AI pioneers Newell and Simon by laying out several grand and general AI principles and hypotheses, which can be understood as the basic presumptions underlying Cyc. After establishing their definition of intelligence as "the power to rapidly find an adequate solution in what appears a priori (to observers) to be an immense search space" (Lenat and Feigenbaum 1991, 186), the authors present the first hypothesis, the Knowledge Principle, which in essence attributes a system's intelligent action to its knowledge. The second generalization they outline is the Breadth Hypothesis, which states that intelligent actors draw on general knowledge as they solve problems, or in other words they analogize "specific knowledge from far-flung domains" (Lenat and Feigenbaum 1991, 186).

The Knowledge Principle in particular comes out of the tradition of expert systems, which were the kinds of programs Lenat had worked on in the 1970s and early 1980s. Expert systems run into trouble when confronted with a task or question that strays outside of their narrow domains. As Margaret Boden writes, this is often called brittleness, "in the sense that missing and/or contradictory data would result in a nonsensical response from the computer" (2014, 93). For example, the MYCIN expert system, designed to help diagnose blood and bacterial diseases, once suggested prior amniocentesis as a diagnosis for a male patient since the system did not "know" that people without uteruses do not get pregnant (Forsythe's example). As another example, a hypothetical system might not catch that a teenager cannot have worked for a company for twenty years (Adam's example). Through the Breadth Hypothesis and by shifting focus from domain-specific expertise to common sense knowledge, Cyc aims to address this so-called brittleness of AI systems. As outlined in "On the Thresholds of Knowledge," the day-to-day routine of encoding Cyc's knowledge base at that time involved humans reading through text (encyclopedia entries and newspaper articles) and entering facts stated in these sources as well as "what the writer of that sentence assumed the reader already knew about the world" (Lenat and Feigenbaum 1991, 219). Today, Cycorp continues to hire for "ontology engineer" roles.

Around the same time that Cyc was coming into being, the World Wide Web exploded onto the scene; subsequently, in the late 1990s and early 2000s, the Semantic Web emerged. The Semantic Web refers to the shift, envisioned by World Wide Web founder Tim Berners-Lee, from a web of HTML documents to a web of data (also referred to as linked data). For example, instead of using "meaningless" HTML tags such as <h1> or , in the Semantic Web one encodes the information on a web page in a more meaningful way (e.g., by indicating that <h1> contains a person's name or that the page describes a video with a title, run time, etc.). This principle is sometimes referred to as "things not strings." The provision of "semantic" linked data (as opposed to blobs of text with a bit of HTML structure) is supported by markup and syntax languages created in the late 1990s, such as Resource Description Framework (RDF), RDF schema (RDFS), and Web Ontology Language (OWL). These were informed by the development of eXtensible Markup Language (XML), which, unlike HTML, allows tags to be customized and defined in document definitions or schemas.

[2] Written in 1987, revised and published in 1989, and republished in 1991 along with their response to Brian Cantwell Smith's critical review.

Practitioners sometimes think of the Semantic Web's origins only in terms of the development of the World Wide Web, to the obfuscation of its AI heritage. Although it is not a secret that the Semantic Web has connections to knowledge representation efforts in AI practice,[3] this inheritance may not always be visible front and center. For example, the recently published and highly accessible primer *Linked Data for the Perplexed Librarian* (Carlson et al. 2020) introduces the Semantic Web in relation only to the World Wide Web. In contrast, Halpin (2013) remarks on the parallels between GOFAI's knowledge representation tradition and the Semantic Web, both the common goal to find a universal encoding and the resemblance of GOFAI semantic networks and Semantic Web resources and links.[4]

Furthermore, Cyc and the Semantic Web share both vision and a key architect. In the mid-1990s, Ramanathan V. Guha split from Cyc and moved on to work on emerging web technologies, and Lenat established Cycorp. Trailing Guha leads directly to some of the foundational technologies underpinning the Semantic Web, which he helped architect. Guha helped create both CycL and RDF (Halpin 2013, 56–57). Following his departure from Cyc, he worked for Apple and Netscape, playing a key role in developing RDF (the core linked data language) based on his Meta Content Framework (MCF) (Andreessen 2008). This development led to RSS (which originally stood for RDF Site Summary). Ben Hammersley (2003, 2) positions CycL and other knowledge representation languages from AI research traditions as the roots of RDF and, indeed, RDF's structure consists of binary statements in the form of subject-predicate-object that remind one of a simplified CycL. RDF also has a similar ability to define classes and subclasses, properties and subproperties. Later, languages like OWL emerged to enable axioms and other logical structures, again reminiscent of CycL.

As noted above, however, there are fundamental differences between Cyc and the Semantic Web. Unlike AI, the Semantic Web does not make claims about intelligence writ large or give a theory of mind. Instead, its focus is on encoding metadata, facilitating information retrieval, and connecting resources to one another. As Yorick Wilks relates, the Semantic Web is grounded in web document technologies and document markup/annotation (2008, 42), which means that in addition to its roots in AI and knowledge representation, the Semantic Web also grows out of text structuring traditions like HTML and XML (and its predecessor SGML) as well as internet protocols. The Semantic Web's goal is generally scoped to the task of information retrieval within the realm of the World Wide Web and navigating its connections. In a 2013 talk given at the 12th International Semantic Web Conference (ISWC), Guha (2013) stated that a key difference between Cyc and the Semantic Web was adoption: he characterized Cyc as a project born out of academia that expected its product to "wake up" and which was not focused on trying to get many people to contribute to it, whereas the focus of the Semantic Web was to get web developers to insert structured, somewhat meaningful markup on their web pages. In other words, Cyc took a more closed, top-down approach, while the Semantic Web's vision could be construed as a distributed, bottom-up approach.

In its early years, the Semantic Web was somewhat intangible and not widely adopted (although interestingly, Cyc participated by releasing a linked data subset of the knowledge base as OpenCyc, which has since disappeared). There is still an acknowledged gap between Semantic Web research and practice (Verborgh and Vander Sande 2020); however, between 2007 and 2012, several projects emerged and pushed the Semantic Web and linked data from dreamlike vision to something more tangible: DBpedia (2007), Schema.org (2011), and Wikidata (2012).

Here is another area of resemblance, reflection, and/or convergence with Cyc's legacy: Schema.org, which represents a pivotal project in the history of the Semantic Web, was created by Guha. An ontology coordinated by search engine companies, it makes use of the RDF subject-predicate-object syntax to structure metadata about web pages and the things described therein. Like Cyc, it, too, positions Thing as its universal class under which all others fall. However, unlike Cyc, Schema.org is not about common sense but rather about describing objects and their attributes. DBpedia and Wikidata reflect Cyc in their combination of RDF or RDF-inspired syntax as well as their fascination and connection with encyclopedias. Released in 2007, DBpedia converted some structured infoboxes on Wikipedia, a global online encyclopedia editable by anyone, into RDF. Wikidata, launched in 2012, is a database of linked data derived in part from Wikipedia (in that each Wikipedia page has a corresponding Wikidata page) but which contributors can also add to independently. Put simply, Wikidata is composed of items (identified by numbers prefixed with the letter "Q" and consisting of classes and instances) and properties (identified by numbers prefixed by the letter "P" and constituting the project's relationships). Beyond statements consisting of subjects, predicates, and objects, Wikidata can also capture statement qualifiers and references.

[3] In fact, this connection is made in the 2001 *Scientific American* article that coined the term *Semantic Web* (Berners-Lee, Hendler, and Lassila 2001). Articles like Yorick Wilks's "The Semantic Web: Apotheosis of Annotation, but What Are Its Semantics?" position Cyc as "one of the predecessors of the SW as a universal repository of formalized knowledge" (2008, 46).

[4] Although crucially, RDF and web protocols address semantic networks' issue of ambiguity and lack of interoperability by incorporating URIs.

Initially created to support linking between Wikipedia articles in different languages, Wikidata has since grown into a vast ontology and knowledge base. In an eerie way, Wikidata and DBpedia hearken back to Cyc and GOFAI in their resemblance to one of Smith's alternative suggestions for Cyc:

> L&F [Lenat and Feigenbaum], on the current design, retain only the formal data structures they generate, discarding the natural language articles, digests, etc., used in its preparation. Suppose, instead, they were to retain all those English entries, thick with connotation and ineffable significance, and use their data structures and inference engines as an active indexing scheme. Forget intelligence completely, in other words; take the project as one of constructing the world's largest hypertext system, with CYC functioning as a radically improved (and active) counterpart for the Dewey decimal system. (Smith 1991, 282)

Leaving aside that Dewey Decimal Classification can be problematic, Halpin points out that this passage from Smith's article "strangely prefigures not only search engines, but the revitalization of knowledge representation languages due to the Semantic Web" (2013, 56). Even the Cyc founders themselves wrote about its anticipated first practical application being "information management" in light of "vastly greater amounts of information being available on-line," where unfortunately "most of the task of knowing where relevant information resides, and accessing it, has been left to users" (Guha and Lenat 1994, 131–32). In sum, while Semantic Web technologies such as RDF and ontologies/knowledge bases like Schema.org, DBpedia, and Wikidata are not simply Cyc in another guise, they share history, influences, formalization strategies, and encyclopedic inspiration.

Critiques of Cyc

Now that I have established the parallels and common inheritances of GOFAI projects like Cyc and the Semantic Web, we can begin to investigate whether critiques of Cyc and related GOFAI knowledge representation projects such as expert systems might apply to Semantic Web initiatives. The following section draws on critiques by Hamid Reza Ekbia, Harry M. Collins, Alison Adam, Hubert Dreyfus, Diana Forsythe, and Brian Cantwell Smith.

It must be noted that Cyc is something of a specter since it is proprietary and therefore it is difficult to know how it is actually set up and functioning. Details are necessarily secondhand unless one can gain access to a training or a demonstration (as Ekbia chronicled in his 2008 book *Artificial Dreams*). Therefore, I focus less on Cyc's technical details and more on the critiques that have been made, in order to read those critiques into Semantic Web examples to see how they apply.

Formal Representation: Knowing That Versus Knowing How

One of the fundamental lines of critique relating to Cyc consists of a dispute about what constitutes knowledge and whether it can be represented formally. Cyc assumes that all knowledge can be expressed explicitly and that it can be extracted. Adam identifies the roots of this assumption in traditional epistemology. As she notes, this position emphasizes propositional knowledge, or "'knowing that' knowledge over other types of knowledge, particularly the 'knowing how' type of knowledge described by phenomenologists and those who emphasize the role of the body in making knowledge" (Adam 2000, 237). Similarly, according to Hubert Dreyfus, the understanding of common sense as the kind of knowledge that can be represented digitally comes from a line of thought he calls representationalism, which stems from Descartes's and Leibniz's rationalist emphasis on propositional knowledge. According to Dreyfus, the drive to represent "all that we know . . . in *formal* rules and features only arises after one has already assumed that common sense derives from *a vast data base of propositional knowledge*" (1992, xvii).

"Knowing how" is a kind of knowledge not amenable to formal representation. In "Why Expert Systems Do Not Exhibit Expertise," Hubert Dreyfus and Stuart Dreyfus argue that "we possess something called 'know-how,' which we have acquired from practice and sometimes painful experience. That know-how is not accessible to us in the form of facts and rules" (1986, 86). In his introduction to *What Computers Still Can't Do*, Hubert Dreyfus posits that "background knowledge consists largely of skills for dealing with things and people rather than facts about them" (1992, xii). According to Ekbia, Hofstadter thought that common sense was not a kind of expertise but, rather, a more general ability and cognitive style and that trying to formalize this kind of thought was a "Boolean dream" (Ekbia 2008, 100). Dreyfus and Dreyfus (1986) emphasize the role of intuition and image-based thinking in expert skill and argue that holistic understanding cannot be replicated in represented facts and heuristic rules.

Critics make the point that computers and human brains function differently and represent the world differently. For example, Smith notes that "not all of what matters about a situation need be captured, at

least in the traditional sense, in the meanings of its constituent representations" and claims that "the full significance of an intentional action can outstrip its content" (1991, 273). Ultimately, Smith advocates for accepting that "our systems represent the world differently from us" (1991, 284). These critics thus point out that not all knowledge is amenable to digital representation.

Other Critical Points

Critics have given various reasons why not all knowledge can be formalized. Ekbia summarizes the main points about knowledge that critics of expert systems have raised to counter the view that (all) knowledge can be formalized as follows:

- knowledge is socially and culturally constituted;
- knowledge is not self-evident but must be interpreted;
- people are not aware of everything they know and a good deal of knowledge is tacit;
- much knowledge is not in people's heads at all but, rather, is distributed in their bodily skills and social know-how;
- the relation between what people think, say, and are observed to do is highly complex. (Ekbia 2008, 101)

The critical threads I am interested in overlap with Ekbia's points and will be grouped as follows:

- role of culture and community in knowing
- context and the Frame Problem
- consensus reality
- view from nowhere
- issues with the sources of knowledge and extractive nature of formalization
- lack of complex examples
- closed development versus transparency
- grandiosity of rhetoric versus practical feasibility.

Role of Culture and Community in Knowing

Numerous critics have pointed out that knowing does not happen in a vacuum. For example, Collins pushes back against what he calls the Rules Model, in which the interpretive abilities needed to understand "the formal and statable content of human knowledge" themselves are "made of endlessly ramifying rules that remain unexpressed" (1990, 94). To Collins, this view does not represent reality. Rather,

> it is our cultural skills that enable us to make the world of concerted behavior. We do this by agreeing that a certain object is, say, a Rembrandt or a certain symbol is an s. That is how we digitize the world. It is our common culture that makes it possible to come to these agreements, and it is our means of making these agreements that comprises our culture. (Collins 1990, 109)

In other words, learning happens through acculturation, even though we think it is conveyed by rules. Similarly, Dreyfus (1992, xxii) draws on Heidegger, Merleau-Ponty, and Bordieu to conclude that "what counts as the facts" is tied to everyday know-how and skill. Based on her anthropological fieldwork among AI practitioners in the 1980s and 1990s, Forsythe notes that "knowledge" tends to be understood to mean formal or codified knowledge, such as that contained in encyclopedias and textbooks. "What everybody knows" is less likely to be treated as knowledge. This contrasts with the anthropological view of knowledge (including common sense) as cultural. According to Forsythe, anthropologists view knowledge as cultural and thus local, meaning "the notion of universally applicable common sense is an oxymoron" (2001, 21).

Forsythe actually critiqued Collins's critique, noting that his invocation of "culture" was an underdeveloped gloss and pointing out that it matters "which cultural tradition defines what will be taken as skill in a particular context" (1993, 473n20). She says that Collins's generalization of culture parallels the knowledge engineers' practice of "deleting" the social and cultural. These deletions involve eliding questions such as "when an expert system is built, whose interpretation of reality should it encode? Whose 'knowledge' should constitute the 'knowledge base'; whose practice should be enshrined as 'expertise'?" (Forsythe 1993, 469) and rendering invisible their own role in selecting and interpreting that which is encoded as knowledge. While Cyc departed from an expert systems approach by focusing on common sense and "what everybody knows," the process relied on codified knowledge as a source and it presumed agreement about what everybody knows.

Context and the Frame Problem

Issues of cultural context and purported universality in Cyc also point toward the classic Frame Problem in AI. Boden explains the Frame Problem as two problems: "first, knowing which aspects of a situation would be changed by a particular action, and which would not; second, reasoning with incomplete knowledge, due to our inevitable ignorance about the facts of the real world and to the vagueness of ordinary-language concepts" (2014, 93). Murray Shanahan distinguishes between an epistemological Frame Problem ("how is it possible for holistic, open-ended, context-sensitive relevance to be captured by a set of propositional, language-like representations of the sort used in classical AI?") and a computational Frame Problem: "how could an inference process tractably be confined to just what is relevant, given that relevance is holistic, open-ended, and context-sensitive?" (Shanahan 2016). For Boden, "the frame problem arises whenever implications tacitly assumed by human thinkers are ignored by the computer because they haven't been made explicit" (2018, 43).

Ekbia summarizes Smith's point that Cyc's way of making knowledge explicit fixed meaning independently of context and applauds Cyc's later development of multidimensional context "spaces," giving the example of time's dimensions of truth and relevance in Cyc and how this leads to understanding of inference as process rather than something static (2008, 124). Even with the development of microtheories and contexts, though, Ekbia notes that "the most serious problem, however, is that of 'relevance' – that is, in what context an assertion belongs" (2008, 123).

Consensus Reality

Context also comes into play in Cyc's presumption of the existence of "consensus reality," which critics have pointed out as problematic. Smith posits that "what the structure means can't be separated from the whole complex of inferential, conversational, social, and other purposes to which it is put" (1991, 264). He distinguishes between ambiguity and indexicality, challenging the idea that meaning could (or should) be systematically catalogued as properties.

Adam notes that AI developers believe there is a real world that can be accessed through perception about which we all agree (2000, 241). Drawing on Forsythe's observations, Adam posits that Cyc is built on the assumption that one can access or observe a "real" world, a notion challenged by the argument that "all our observations are mediated by our theories of the world" (1998, 71). Cyc's contexts and microtheories, Adam (1998) argues, do not slice parts of the world the way humans would because common sense is more seamless than Cyc's discrete contexts.

View from Nowhere

Critics of Cyc have also pointed out that it is built based on a purportedly objective "view from nowhere" (Nagel 1986) and a centering of the "non-weird" (a term Adam [1998] borrows from epistemologist Richard Foley, and which refers to a dominant or "normal" perspective). The problems with consensus reality emerge most clearly in Cyc's glossing of an amorphous "we"—an example of a "view from nowhere." When assessing Cyc by asking, "is representational content contextual (situated)?," Smith (1991, 262) remarks on traditional logic's disregard of context and its consequent "view from nowhere." Locating Cyc's approach of the early 1990s in this vein, he concludes that they, too, ignore context in the same way. In the Cyc project's partitioning of knowledge and belief, in which some contexts are tagged as "beliefs," Adam finds evidence that what the Cyc knowledge base encodes is very much a view from "TheWorldAsTheBuildersOfCycBelieveItToBe," belying not in fact a view from nowhere, but the adoption of an "epistemologically authoritative 'non-weird' perspective" (1998, 74). The "we" is Cyc's creators and a hypothetical group of healthy, sane "non-babies" (Adam 1998, 74). As Adam writes, "Cyc is an example of what Code (1993) has described as the universal knowing subject, or the view from nowhere being used potentially to discount views which are 'crazy' or 'weird' in Foley's terms or one of Lenat's minority beliefs" (2000, 244).

As Adam explains, the unstated knowing subject responsible for this view can be glimpsed in Cyc's multiple contexts, which encode different worldviews (such as capitalism and Marxism). The existence of these contexts—which emerged later on, after Smith's review of the project—implies a judgement about which worldviews get represented or modeled as contexts. Inclusion at the level of model, according to Adam, confers epistemic status and is only granted to phenomena that are "exact enough," meaning some are left out (1998, 72). For example, Adam suggests that while Cyc might regard Marxism and capitalism as contexts, it might not elevate feminist views on economics that take women's unpaid labor into account (1998, 71). Using the lens of feminist epistemology, Adam builds on Forsythe's idea of deleting the social and says that GOFAI systems delete the subject, arguing that "what you know depends crucially on who you are" (Adam 2000, 232).

Source of Knowledge and Extractive Nature of Formalization

A related invisible assumption among AI practitioners that critics have revealed is that knowledge is "concrete" and, in the words of Forsythe, "a straightforward, bounded entity that can intentionally be acquired" (1993, 458).

According to Forsythe, words like *extract*, *acquired*, *uncovered*, *revealed*, *cloned*, *elicited*, and *stored* have been used in knowledge engineering literature and belie an understanding of the process as one of transfer as opposed to the more interpretive or constructive concept of translation (1993, 459). Forsythe argues that "knowledge engineers' conception of their task as knowledge transfer represents a significant case of deletion, since the notion of straightforward transfer completely obscures their own role in the selection and interpretation of what goes into a knowledge base" (1993, 463).

Forsythe (1993, 452) also observes that knowledge engineers privilege textbook knowledge. Similarly, describing a knowledge elicitation session observed during his fieldwork among knowledge engineers and crystal growers at the University of Bath, Collins remarks on the dominance of textbook information as opposed to experience during the interaction between the expert and the knowledge engineer. According to Dreyfus (1992, xxiv), Cyc likewise envisioned its future automated knowledge acquisition process to be one derived from independently reading written sources.

Closed Development and Lack of Transparency

Although not one of the most common critiques of the project, the closed and proprietary nature of Cyc has also been a subject of criticism. Ekbia notes that the project's drift from AI research project to proprietary business "makes it difficult to judge Cyc on purely theoretical grounds" (2008, 106) and that its secretive nature makes it hard to ascertain whether the project has actually achieved its stated goals. Cyc's approach was to create baseline data to be used widely, but to do so by employing a small group of specialists rather than inviting broad participation and collaboration. In his analysis of Cyc, Ekbia relied on a 1994 report of a question and answer session with Cyc by Vaughan Pratt, along with Ekbia's own participation in a training course in 2002. According to Ekbia, trainings are aimed at current or potential customers, and it can take years of waiting to land a spot in a training session (2008, 361n17). Although Cyc did make a small part of its ontology available on the (Semantic) Web as OpenCyc and a larger portion of the database available on request as ResearchCyc, both of these initiatives have since been discontinued ("KBpedia - OpenCyc Not Online" n.d.).

Complex Examples

Piling on top of the problem of Cyc's lack of transparency, critics such as Adam have taken issue with the examples provided by the developers. Adam argues that examples used to explain Cyc are selective and problematically simple. For Adam (1998, 59), "simple examples neatly sidestep issues of collective responsibility" and hold up "the role of the individual, rationalist, universal knower." More complex examples, in contrast, can reveal contradictions or more clearly expose the worldview of the "knowing subject." Adam takes issue with the prevalence of simple examples in both epistemology and AI discourse. For example, Lenat and Feigenbaum discuss the example of water and encoding facts about it in "On the Thresholds of Knowledge." Other examples they provide of consensus reality include "the number of tires an auto has; who Ronald Reagan is; what happens if you fall asleep when driving—what we called consensus reality" (Lenat and Feigenbaum 1991, 201). Adam relays Cyc's examples of "what they call 'Americana' as in 'you are not likely to get a speeding ticket in mid- or late-twentieth century America if you're driving less than 5 m.p.h. over the speed limit'" (1998, 73), bringing up the point that though this example seems simple, it hides nuanced questions about how aspects such gender, race, and age intersect with speeding ticket issuance. According to Adam, simple examples reinforce a presumption of agreement; she contends that "were they to choose epistemologically more complex examples, it would be much more difficult to maintain a stance with which 'we' all agree" (1998, 73).

Grandiosity of Claims

Finally, a somewhat implicit critique of Cyc (and many other GOFAI projects) is the grandiosity of its claims. Lenat and Feigenbaum write in "On the Thresholds of Knowledge" that "in the case of the Cyc project, the goal is to capture the full breadth of human knowledge" (1991, 219). Dreyfus was famously critical of AI's optimism, beginning with his 1963 book, *Alchemy in Artificial Intelligence*, and in his later published works on AI. In a 2012 article, Dreyfus takes issue with Cyc's contention that it could encode all of common sense within ten to thirty years, calling out Cyc's receding goalposts as evidence of its "first step fallacy," or the assumption that early success would lead to subsequent success (2012, 95). Smith (1991) is also highly critical of the Cyc team's sweeping claims regarding both the qualities of intelligence and the approach to attaining it. As Smith (1991, 282) points out, the comparatively modest task of hypertext indexing is more reasonable and attainable.

Applying Critiques of Cyc to the Semantic Web

Given the similarities and inheritances of Cyc and the Semantic Web, can we apply some of the same critiques? Ekbia would probably say that yes, much of what is said about Cyc could apply to the Semantic

Web. As he writes, "any result about Cyc may apply to all systems that fall into this category because of the formality of their approach . . . their adherence to the logicist framework, their limitation to explicit forms of knowledge, and their incapability of dealing with context- and use-dependent meaning" (Ekbia 2008, 128).

At the same time, as noted above, the Semantic Web projects I discuss have a different scope and aim than intelligence writ large. For example, unlike Cyc, Schema.org is not about common sense but rather about describing objects and their attributes. Wikipedia and Wikidata are also more about things and attributes and less about common sense "facts" of the Cyc tradition—such as "'people have two arms and two legs'" (Adam 1998, 73)—although any ontology gets into "common sense" territory in its uppermost layer, where "primitive" statements and structures are laid out (such as making all classes subclasses of a highest-level class, such as Thing, or in the case of Schema.org, making Product a high-level class). By having a different scope and aim than Cyc and other GOFAI projects, the Semantic Web avoids some of their pitfalls.

Still, I argue that we urgently need to (continue to) critique Semantic Web projects. As Adam reports, "Lenat and Guha's hope was that, by the turn of the century, it would be commonplace to expect a new computer to come equipped with Cyc" (1998, 69); now, it is commonplace for systems to use content from Wikimedia projects as a basis upon which to build. As Dario Taraborelli notes, content in Wikipedia gets propagated to linked data systems and the "entire ecosystem of AI platforms" (ITU 2017). In other words, Wikimedia content is ubiquitous as a source of "facts" for other platforms, including those using machine learning/AI. What assumptions and potential issues come along with using Semantic Web linked data as the "facts" underlying AI systems of today?

Formal Representation of Knowledge

By virtue of their technical architecture, Semantic Web projects like Schema.org, DBpedia, and Wikidata operate under the same premise as Cyc: that knowledge can be represented formally. The exception is Wikipedia's largely unstructured narrative text, which privileges the written word but is comparatively more free-form and flexible than an ontology. Notably, however, in some recent initiatives structured data and/or code is used to generate narrative encyclopedia entries. For example, the MBABEL Wikipedia template generates draft Wikipedia articles based on Wikidata statements. The new Abstract Wikipedia project (spearheaded by the founder of Wikidata, Denny Vrandečić), aims to create a system in which natural language encyclopedia articles can be drafted in a variety of languages using a knowledge representation architecture based on the representation of lexical and grammatical features of languages. Abstract Wikipedia is an effort to align Wikipedias in different languages and enable Wikipedias with fewer articles or less comprehensive coverage of a topic to easily generate text (Vrandečić 2018). Abstract Wikipedia's goal to encode "common knowledge" across languages hearkens back to Cyc's attempts to capture common sense.

Here, the lesson to be learned from Cyc is not to operate under the presumption that it is possible to encode *all* kinds of knowledge formally in a database (graph or otherwise). We should expect and recognize that some knowledge, such as tacit knowledge and experiential, embodied knowledge, will simply be absent from Semantic Web projects or will have undergone significant transformation as opposed to transfer. This point is a slightly different take than critiques of representation in Wikipedia and Wikidata, which focus on crucial problems such as the lack of articles by and on topics relating to the Global South, BIPOC individuals, women, and gender nonconforming individuals (discussed further below).[5] Critiques of Cyc regarding varieties of knowledge that resist formalization question the very notion of knowledge in these systems. Underneath the concept of representation as presence, this line of critique challenges whether the underlying representational structures themselves do justice to different kinds of knowledge, pointing to the structural issue of what can even be represented in the first place in systems like Wikipedia and Wikidata and using ontologies like Schema.org.

Role of Culture and Community in Knowing

As many critics of Cyc point out, meaning is contextual and requires an interpreter with cultural know-how and background. Statements mean different things in different situations and cultural milieus.

One way that culture and community are apparent in Wikipedia and Wikidata is in their multilingual approach. Each language Wikipedia has its own community of editors and content varies between language editions, which means that each is more culturally specific and less universal in approach than a single encyclopedia—although, as Vrandečić has pointed out, language is not a neat proxy for culture and there can be disagreement on representation of topics within a given language encyclopedia (2019, 8). Wikidata engages in flattening by designating one entity page to connect articles/concepts across Wikipedias, attaching labels in multiple languages to the same Wikidata entity page.

[5] Initiatives doing work in this area include Art + Feminism, AfroCROWD, and Whose Knowledge?

While this flattening is a move toward universalization, due to Wikidata's multilingualism, points of friction between concepts in different languages and cultures have the potential to become more visible in Wikidata. Though perhaps not as epistemologically complex as the examples Adam advocates for, the representation of gelato in Wikidata points to the culturally specific nature of some concepts. In Wikidata, the item known as *ice cream* in English is labeled *gelato* in Italian, but there is also a separate item labeled *gelato* in English and *gelato italiano* in Italian. This second gelato features a very specific description in Italian, which indicates the specific percentage of milk fat and air that is present in the dessert.

Another area where culture and community come into play is in relationships in an ontology between identity terms, which are culturally and personally specific, political, and nuanced. Language around identities is constructed and negotiated, and not easily reduced to classification schemes and blanket statements. While in this article I will not analyze specific examples of race, ethnicity, or identity terms from Wikidata that are not related to my lived experience as a white, cisgender, heterosexual woman, since I do not feel it would be appropriate, I encourage qualified researchers to undertake critical readings of identity terms in Wikidata, specifically how they relate to one another through properties. Such readings have the potential to reveal how particular interpretations of the world come to be codified, and how assertions either open or foreclose multiplicity.

Finally, in line with Smith's imagined use of Cyc, the direct links between Wikipedia (which consists of narrative text) and ontological data structures like Wikidata and DBpedia address the need for context and more nuanced interpretation. By juxtaposing both narrative and structured data, such linked data has the potential to be more embedded and contextualized. However, the relationship between narrative resources and structured data itself needs to be properly contextualized and transparent, especially as the relationship between Wikidata and narrative Wikipedia articles begins to flow from Wikidata to Wikipedia(s) instead of from Wikipedias and other source texts to Wikidata or DBpedia. As an example, where before Wikipedia articles spawned corresponding Wikidata items, the MBABEL Wikipedia template generates "structured narratives" based on Wikidata statements (Azzellini 2019). Crucially, however, the MBABEL template is intended to generate text that is later refined and edited by human editors before publication ("Wikipedia:Mbabel" 2022). Similarly, Abstract Wikipedia envisions individual language Wikipedia communities choosing which text to incorporate from the project and what to overlay or suppress (Vrandečić 2019, 11).

Context and the Frame Problem

Examining issues of context is especially important as linked data gets reused (for example, in corporate black box applications like Google's knowledge graph and rich snippets). The way that Semantic Web data is presented to users for interpretation relates to the Frame Problem, which is about deciding what information is relevant in certain situations. Wikidata's architecture provides the ability to qualify statements—for example, with temporal dimensions (though fewer than possible in Cyc)—and it accommodates contradictory statements about the same item thanks to its reference/citation qualifiers. However, as Ekbia (2008) points out about Cyc, relevance is a problem. In other words, how does a system using linked data know which claim to use? Here in both Cyc and in the Semantic Web could be a looming danger of an infinite regress of rules.

The Semantic Web may sidestep the Frame Problem, however, by not claiming or aiming to accomplish tasks as an autonomous intelligence. The most visible and concrete use of Semantic Web data is by search engine knowledge graphs and cards (and to a lesser extent emerging library technologies that seek to present similar knowledge cards). When a user searches for ambiguous terms, multiple knowledge cards are presented, leaving the final determination of relevance up to the human user. When Guha and Lenat (1994) noted that Cyc could be applied to information management problems, they may have been thinking that Cyc would do all of this work for the user; yet, knowledge graphs and knowledge cards meet the user halfway. Indeed, Smith writes, common sense projects like Cyc "live on in such structures as Google's 'Knowledge Graph,' which is used to organize and provide short answers to queries in search engines. Tellingly, though, the results of such searches are snippets for interpretation by human intelligence, not the basis of the machine's own intelligence" (2019, 37).

Consensus Reality

Semantic Web projects approach consensus reality differently than Cyc. The Semantic Web's "anyone can say anything" principle means that, at least in theory, it is possible to encode multiple perspectives. Rather than emphasize a body of facts upon which everyone agrees that refer to a universal world "out there," Wikipedia and Wikidata emphasize verifiability and citation. In Wikidata, multiple (including conflicting) claims can be made under the same property, which are differentiated using ranking and qualifiers.

In addition, Wikipedia and Wikidata generally take a bottom-up approach to data modeling rather than a rigid and prescriptive approach. For example, property proposal discussions often involve discussion about

alternate ways to model or express the same thing. When creating new items, editors often look to robustly describe items to use as models. Well-described items can even be explicitly stated to be model items by using the *model item* property and following the guidance at Wikidata:Model items ("Wikidata:Model Items" n.d.). Model items are also sometimes identified by WikiProjects—affinity groups focused on certain topics or types of Wikidata items or Wikipedia articles. While WikiProjects do often provide application profiles or data models on their WikiProject pages (for example, "Wikidata:WikiProject Books" n.d.), there is no enforcement of compliance to these data models. Rather, they are suggestions and are often in progress or under revision and development.

While Schema.org has fewer properties than Wikidata—1,453 ("Organization of Schemas" n.d.) versus 10,000 ("Wikidata:List of Properties - Wikidata" n.d.)—and has a different process for adding new ones, its approach to RDF range restrictions demonstrates an intentional flexibility. For example, regarding the *gender* property, the inclusion of the Text datatype in addition to the GenderType controlled list (Enumeration in Schema.org terminology) allows for free text values, and the definition of the property states, "as with the gender of individuals, we do not try to enumerate all possibilities" ("Gender: A Schema.Org Property" n.d.).

Although a bottom-up approach to modeling makes Wikidata much less usable because it means it is inconsistent, the platform's flexible approach means multiple data models and worldviews can be represented. Similarly, while in Schema.org a bottom-up approach to values in the object position of the triple statement may hinder retrieval due to inconsistency, it has the potential to facilitate self-description and more nuanced data.

View from Nowhere

Like Cyc, Semantic Web projects have the potential to fall into the trap of the view from nowhere. For example, Wikimedia or Wikipedia-based projects operate under policies prioritizing neutrality. However, the "we" of Wiki platforms is easier to find than the "we" in Cyc. For one, Wikimedia platforms trace edits to either named accounts (often pseudonymous) or IP addresses. While editor demographic statistics such as gender or race are not collected (nor should they be), these measures mean that there is some level of attribution and visibility of contribution to the platforms.

Adam's quest to find the knowing subject in Cyc is salient for the Semantic Web and is addressed by projects undertaking representational interventions such as Art + Feminism and Whose Knowledge?, among others. These initiatives ask questions like "who is editing Wikipedia articles and contributing to Wikidata?," raising awareness and providing training opportunities to increase participation by "weird" people (as a positive inverse of Foley's "non-weird" referred to by Adam [1998]). In this area, projects like Wikipedia and Wikidata do better than Cyc. Unlike proprietary and secretive Cyc, Wikipedia and Wikidata are (at least nominally) participatory and open for editing by anyone with internet access (although this does not mean there are not barriers to participation, both technical and social). Anyone can add an item to Wikidata right away, and anyone can suggest properties, which go through a public review and approval process. Wiki platforms also feature a robust edit history for each page and "talk" pages, which function as forums for discussion among editors.

Source of Knowledge

Even if we can begin to better know who is creating the knowledge, we still run into the problem that structured data privileges certain sources that perpetuate dominant, "non-weird" perspectives. For example, Schema.org statements based on web pages derive knowledge from what is represented on the internet, which is certainly not neutral nor reflective of the wider universe. As well, given that among Schema.org's creators and maintainers are large search engine companies, there is a certain emphasis in the ontology on products and commercial exchange.

The kinds of sources preferred in Wikipedia and Wikidata mirror the emphasis in expert systems on textbook knowledge and secondary sources rather than direct experience. For example, Wikipedia has a policy of no original research and secondary and tertiary sources are emphasized while primary sources are discouraged ("Wikipedia:No Original Research" 2022). Oral traditions also do not fit into Wikipedia's conception of a reliable source because they are considered primary sources, at least according to some editors ("Wikipedia:Oral History" 2020). While the emphasis on verifiability and citation in Wikimedia projects is laudable and an essential component of their transparency, it means that the kind of knowledge allowed is particular and does not encompass all varieties of sources of knowledge and that published, written sources are privileged over others.

Closed Development and Lack of Transparency

Semantic Web projects are generally much more public and transparent than Cyc. As Guha (2013) notes, the focus of Schema.org and other Semantic Web ontologies was to encourage website developers and admins

to insert structured data in their own web pages, in contrast with Cyc, where the encoding of structured data was/is done by a small group of employees.

DBpedia is operated by a consortium association and derived from completely public information on Wikipedia ("About DBpedia" n.d.). DBpedia does welcome contribution and collaboration, although one must understand Semantic Web technologies in order to contribute (for example, by improving mappings between Wikipedia infoboxes and DBpedia's ontology). Wikipedia and Wikidata are themselves by their nature open and collaborative. Wikipedia and Wikidata provide talk pages and discussion spaces such as forums within which editors discuss and debate content and policies. Though the process of adding properties and classes to Schema.org is not as immediate as it is on Wikidata, Schema.org governance and discussion is public on Github and public mailing lists.

Transparency alone is not enough to ensure robust participation and easy access, however. There are still barriers to participation in Semantic Web projects, even though systems are open to any user and documentation is openly available, because there is still a level of technical inscrutability that intimidates participants. One has to know the terminology and where to go in order to read and participate in the behind the scenes processes. Furthermore, participation in the governance of Semantic Web projects can be time-consuming.

Complex Examples

In Wikidata, each property page's documentation includes examples. When it comes to epistemologically complex examples, Wikidata's proposal discussion process and the presence of an ongoing discussion page on approved properties provide a platform for the introduction of complex examples. For example, the property proposal for P2562 (English label *married name*) features a robust discussion, including "exceptional" or complicated cases ("Wikidata:Property Proposal/Archive/45 - Married Name" n.d.). The approved property's discussion page shows ongoing discourse about and modification of property constraints based on arguments made by Wikidata editors ("Property Talk:P2562" n.d.).

Schema.org includes examples of the properties in use in various markup syntaxes on property pages, although examples are not present on all property pages. For example, many of the properties related to the *person* class do not seem to include examples. Similar to Wikidata, discussion about properties and classes can be accessed by clicking "Check for open issues" under the "more" section of the page, which searches open Schema.org Github issues. For example, the property *honorificSuffix* leads to an open issue (at the time of this writing) regarding the semantics of the property, which includes examples that complicate the existing definition of "An honorific suffix following a Person's name such as M.D. /PhD/MSCSW" ("HonorificSuffix: A Schema.Org Property" n.d.).

Grandiosity of Claims

The rhetoric of Wikipedia has historically been that it will "freely share the sum of all knowledge" ("Wikipedia:About" 2022a). The vision statement of the Wikimedia Foundation, which supports projects like Wikipedia and Wikidata, is to "imagine a world in which every single human being can freely share in the sum of all knowledge" ("Vision" 2022). This language parallels Cyc's stated aim to "capture the full breadth of human knowledge." It is worth being cautious about moonshot discourse and the language of totality and universality. In fact, in May 2022 Wikipedia contributors shifted the language on the "About" page away from the phrase "sum of all knowledge" to refer to it as "a widely accessible and free encyclopedia that contains information on all branches of knowledge" ("Wikipedia:About" 2022b). With its tight scope, Schema.org's mission avoids the trap of grandiosity by declaring as its purpose "to create, maintain, and promote schemas for structured data on the Internet" ("About" n.d.).

Conclusion

This article has shown that Semantic Web projects share history and architecture with GOFAI's expert systems and specific projects like Cyc. I have argued that Semantic Web projects inherit some of the same issues critics pointed out about their predecessors, such as assumptions relating to the formal representation of knowledge, privileging a "view from nowhere," and certain kinds of sources of knowledge. However, the Semantic Web's more modest scope when compared to AI and its quest for autonomous artificial general intelligence, along with the Semantic Web's more transparent approach to discussion and development, make it less susceptible to some of the critiques of Cyc.

Participation in linked data and Semantic Web initiatives should be construed as participation in artificial intelligence development. I hope that foregrounding the Semantic Web's connections to AI will open up space for future scholarship to apply aspects of the robust discourse happening now on AI

ethics to Semantic Web knowledge representation projects. Currently, a major focus of AI criticism is machine learning and algorithmic bias (for example, Safiya Noble's *Algorithms of Oppression*, Meredith Broussard's *Artificial Unintelligence*, and Shalini Kantayya's film *Coded Bias*). Two recent reports about AI in libraries focus on machine learning and natural language processing of unstructured text (Padilla 2020; Cordell 2020). The newly formed AI4LAM group ("About" n.d.a) is an example of library, archive, and museum practitioners coming together to learn about and conduct research into how natural language processing and related machine learning techniques could be leveraged to accomplish tasks such as metadata creation. However, though it is not usually talked about under the umbrella of AI, GLAM involvement in the Semantic Web should *also* be thought of as a form of AI practice: it is a foray into the "forgotten" branch of AI, Good Old Fashioned AI. Understanding the Semantic Web's connections to (GOF)AI foregrounds the viewpoint that machine learning is not the only AI work going on in libraries, and suggests that today's broader critiques of AI ethics might also be applicable to Semantic Web projects.

Autonomous artificial general intelligence has not been achieved and may never be, but this does not mean that projects like Cyc are necessarily failures. Instead, Cyc's Semantic Web cousins accomplish something else by encoding knowledge the way that they do. Berners-Lee, Hendler, and Lassila (2001, 37) defined the Semantic Web as "not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation." Rather than acting autonomously, it functions as part of a collaborative relationship between human and machine. As it is practiced in projects like Wikidata and Wikipedia, DBpedia, and Schema.org, the Semantic Web seems to have become what Smith suggested for Cyc: an interconnected index that links source texts and structured knowledge.

At the same time, knowledge bases such as Wikidata, DBpedia, and Wikipedia itself are used within AI systems whose scope goes beyond indexing, and this reuse should be approached critically. The Semantic Web is not reflective of the entire world (not everything is on the internet!), and yet if it is used as a representation of the world, we must recognize that it comes from a particular strategy (knowledge representation) and source (the World Wide Web and Wikimedia editing communities). We can avoid slipping into a new Boolean dream by approaching participation in the Semantic Web with awareness gained from applying the critical lenses that have been aimed at Cyc. Such awareness can help us recognize the limits and assumptions of knowledge representation and ontologies. We can also learn from the Semantic Web's comparative successes: the transparency of the Semantic Web projects discussed here is a point of hope and encouragement and a potential path toward responsible knowledge curation that leans more toward collaboration than extraction.

Competing Interests

The author has no competing interests to declare.

References

"About." n.d.a. AI4LAM. Accessed May 30, 2022. https://sites.google.com/view/ai4lam/about.

"About." n.d.b. Schema.org. Accessed May 30, 2022. https://schema.org/docs/about.html.

"About DBpedia." n.d. DBpedia Association. Accessed May 27, 2022. https://www.dbpedia.org/about/.

Adam, Alison. 1998. *Artificial Knowing: Gender and the Thinking Machine*. London: Taylor & Francis. https://doi.org/10.4324/9780203005057.

Adam, Alison. 2000. "Deleting the Subject: A Feminist Reading of Epistemology in Artificial Intelligence." *Minds and Machines* 10 (2): 231–53. https://doi.org/10.1023/A:1008306015799.

Andreessen, Marc. 2008. "Innovators of the Net: R.V. Guha and RDF." *TechVision*, February 5, 2008. https://web.archive.org/web/20080205163659/http://wp.netscape.com/columns/techvision/innovators_rg.html.

Azzellini, Érica. 2019. "Presentation for WikidataCon 2019 on Mbabel Tool: Automatic Articles on Wikipedia with Wikidata." Presented at the WikidataCon, Berlin, October 25, 2019. https://commons.wikimedia.org/wiki/File:WikidataCon_-_Lightning_Talk_-2_(1).pdf. Archived at: https://perma.cc/9NS2-SPH7.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284 (5): 34.

Boden, Margaret A. 2014. "GOFAI." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 89–107. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.007.

Boden, Margaret A. 2016. "General Intelligence as the Holy Grail." In *AI: Its Nature and Future*. Oxford, United Kingdom: Oxford University Press.

Carlson, Scott, Cory Lampert, Darnelle Melvin, and Anne Washington. 2020. *Linked Data for the Perplexed Librarian*. Chicago: ALA Editions.

Collins, Harry M. 1990. *Artificial Experts: Social Knowledge and Intelligent Machines*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/1416.001.0001.

Cordell, Ryan. 2020. "Machine Learning + Libraries: A Report on the State of the Field." LC Labs. Library of Congress. https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf.

Dreyfus, Hubert, and Stuart Dreyfus. 1986. "Why Expert Systems Do Not Exhibit Expertise." *IEEE Expert* 1 (2): 86–90. https://doi.org/10.1109/MEX.1986.4306957.

Dreyfus, Hubert L. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press.

Dreyfus, Hubert L. 2012. "A History of First Step Fallacies." *Minds and Machines* 22 (2): 87–99. https://doi.org/10.1007/s11023-012-9276-0.

Ekbia, Hamid Reza. 2008. *Artificial Dreams: The Quest for Non-Biological Intelligence*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511802126.

Forsythe, Diana E. 1993. "Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence." *Social Studies of Science* 23 (3): 445–77. https://www.jstor.org/stable/370256.

Forsythe, Diana E. 2001. "The Construction of Work in Artificial Intelligence." In *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*, edited by David J. Hess. Stanford, CA: Stanford University Press, 16–34.

Foxvog, Douglas. 2010. "Cyc." In *Theory and Applications of Ontology: Computer Applications*, edited by Roberto Poli, Michael Healy, and Achilles Kameas, 259–78. Dordrecht: Springer. https://doi.org/10.1007/978-90-481-8847-5_12.

Guha, Ramanathan V., and Douglas Bruce Lenat. 1994. "Enabling Agents to Work Together." *Communications of the ACM* 37 (7): 126–42. https://doi.org/10.1145/176789.176804.

Guha, Ramanathan V. 2013. "Light at the End of the Tunnel." Presented at the 12th International Semantic Web Conference (ISWC), Sydney, Australia, October 2013. http://videolectures.net/iswc2013_guha_tunnel/.

Halpin, Harry. 2013. *Social Semantics: The Search for Meaning on the Web*. Semantic Web and Beyond. Boston, MA: Springer. https://doi.org/10.1007/978-1-4614-1885-6.

Hammersley, Ben. 2003. *Content Syndication with RSS*. O'Reilly. http://archive.org/details/contentsyndicati00hamm.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

"HonorificSuffix: A Schema.org Property." n.d. Schema.org. Accessed May 27, 2022. https://schema.org/honorificSuffix.

ITU. 2017. "AI for Good Interviews: Dario Taraborelli, Head of Research, Wikimedia." YouTube video, 4:38. https://www.youtube.com/watch?v=J_OlBC04uQc&list=PLpoIPNlF8P2PFPZFhYVaUsZrlxcQr6Bhx&index=46.

"KBpedia – OpenCyc Not Online." n.d. KBpedia. Accessed May 27, 2022. https://kbpedia.org/resources/opencyc/.

Lenat, Douglas B., and Edward A. Feigenbaum. 1991. "On the Thresholds of Knowledge." *Artificial Intelligence* 47 (1–3): 185–250. https://doi.org/10.1016/0004-3702(91)90055-O.

Nagel, Thomas. 1986. *The View from Nowhere*. New York: Oxford University Press.

Newell, Allen, and Herbert A. Simon. 1976. "Computer Science as Empirical Inquiry: Symbols and Search." *Communications of the ACM* 19 (3): 113–26.

Olson, Hope A. 2008. "How We Construct Subjects: A Feminist Analysis." *Library Trends* 56 (2): 509–41. https://doi.org/10.1353/lib.2008.0007.

"Organization of Schemas." n.d. Schema.org. Accessed May 27, 2022. https://schema.org/docs/schemas.html.

Padilla, Thomas. 2020. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. OCLC Research. https://www.oclc.org/content/dam/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai-a4.pdf. Archived at: https://perma.cc/U8ZV-RWQS.

"Property Talk:P2562." n.d. Wikidata. Accessed May 27, 2022. https://www.wikidata.org/wiki/Property_talk:P2562.

Shanahan, Murray. 2016. "The Frame Problem." In *The Stanford Encyclopedia of Philosophy Archive*, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2016/entries/frame-problem/. Archived at: https://perma.cc/4WP3-Q6KC.

Smith, Brian Cantwell. 1991. "The Owl and the Electric Encyclopedia." *Artificial Intelligence* 47 (1–3): 251–88. https://doi.org/10.1016/0004-3702(91)90056-P.

Smith, Brian Cantwell. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/12385.001.0001.

Sparck Jones, Karen. 2004. "What's New about the Semantic Web? Some Questions." *ACM SIGIR Forum* 38 (2): 18–23. https://doi.org/10.1145/1041394.1041398.

Verborgh, Ruben, and Miel Vander Sande. 2020. "The Semantic Web Identity Crisis: In Search of the Trivialities That Never Were." *Semantic Web* 11 (1): 19–27.

"Vision." 2022. Wikimedia Meta-Wiki. April 29, 2022. https://meta.wikimedia.org/wiki/Vision.

Vrandečić, Denny. 2018. "Capturing Meaning: Toward an Abstract Wikipedia." In *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas,co-located with 17th International Semantic Web Conference*, edited by Marieke van Erp, Medha Atre, Vanessa Lopez, Kavitha Srinivas, and Carolina Fortuna. http://ceur-ws.org/Vol-2180/ISWC_2018_Outrageous_Ideas_paper_6.pdf. Archived at: https://perma.cc/59X3-BMX5.

Vrandečić, Denny. 2019. "Collaborating on the Sum of All Knowledge Across Languages." In *Wikipedia @ 20*, edited by Joseph Reagle and Jackie Koerner. https://wikipedia20.pubpub.org/pub/vyf7ksah/release/6. Archived at: https://perma.cc/6GKH-KS8N.

"Wikidata:List of Properties - Wikidata." n.d. Accessed May 27, 2022. https://www.wikidata.org/wiki/Wikidata:List_of_properties.

"Wikidata:Model Items." n.d. Wikidata. Accessed May 27, 2022. https://www.wikidata.org/wiki/Wikidata:Model_items.

"Wikidata:Property Proposal/Archive/45 - Married Name." n.d. Wikidata. Accessed May 27, 2022. https://www.wikidata.org/wiki/Wikidata:Property_proposal/Archive/45#P2562.

"Wikidata:WikiProject Books." n.d. Wikidata. Accessed May 27, 2022. https://www.wikidata.org/wiki/Wikidata:WikiProject_Books.

"Wikipedia:About." 2022a. Wikipedia. Updated March 8, 2022, 02:15. https://en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=1075857249.

"Wikipedia:About." 2022b. Wikipedia. Updated May 27, 2022, 11:09. https://en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=1090099519.

"Wikipedia:Mbabel." 2022. Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Mbabel&oldid=1079369722.

"Wikipedia:No Original Research." 2022. Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:No_original_research&oldid=1088999029.

"Wikipedia:Oral History." 2020. Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Oral_history&oldid=981574513.

Wilks, Yorick. 2008. "The Semantic Web: Apotheosis of Annotation, but What Are Its Semantics?" *IEEE Intelligent Systems* 23 (3): 41–49. https://doi.org/10.1109/MIS.2008.53.

How to cite this article: Provo, Alexandra. 2022. Semantic Encyclopedias and Boolean Dreams. *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 6(3). https://doi.org/10.18357/kula.155

Submitted: 4 August 2021 **Accepted:** 24 March 2022 **Published:** 27 July 2022

KULA: Knowledge Creation, Dissemination, and Preservation Studies is a peer-reviewed open access journal published by University of Victoria Libraries.

OPEN ACCESS