Special Issue: Semantic Media

SM+S
social media + society

Social Media + Society
July-September 2023: 1–13
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20563051231195552
journals.sagepub.com/home/sms
Sage

Wikidata as Semantic Infrastructure: Knowledge Representation, Data Labor, and Truth in a More-Than-Technical Project

Heather Ford[1] and Andrew Iliadis[2]

Abstract

Wikidata is a knowledge base (i.e., a database of facts) project of the Wikimedia Foundation and a sister project to Wikipedia that contains billions of facts that anyone can edit. Platform companies, such as Amazon, Google, and Microsoft use Wikidata to retrieve and transmit facts across the internet, while researchers rely on Wikidata as a data source for facts. Various Wikipedia researchers have commended Wikidata for its collaborative nature and liberatory potential, yet less attention has been paid to the social and political implications of Wikidata. This article aims to advance work in this context by introducing the concept of semantic infrastructure and outlining how Wikidata's role as semantic infrastructure is the primary vehicle by which Wikipedia has become infrastructural for digital platforms. We develop two key themes that build on questions of power that arise in infrastructure studies and apply to Wikidata: knowledge representation and data labor. We argue that considering these issues about Wikidata helps contextualize such infrastructural technologies within existing media and communication studies debates and highlights the contingencies upon which their outcomes depend.

Keywords

Wikidata, semantic infrastructure, knowledge representation, data labor, Wikipedia

Introduction

Wikidata was launched in 2012 as the newest project of the Wikimedia Foundation, the non-profit organization that hosts Wikipedia. The project was developed predominantly by Wikimedia Deutschland and funded with donations from the Allen Institute for Artificial Intelligence, the Gordon and Betty Moore Foundation, and Google, totaling 1.3M EUR. In a press release announcing the project, Wikidata was billed as "a collaboratively edited database of the world's knowledge" (Roth, 2012). Titled "The Wikipedia Data Revolution," the press release described the benefits that Wikidata promised for Wikipedia and other actors. Wikipedia editors would benefit from Wikidata's central storage of interwiki links and infobox data across multiple-language versions of Wikipedia. The project would result in "higher consistency and quality within Wikipedia articles," "decrease the maintenance effort" for Wikipedia editors and volunteers, and increase the "availability of information in the smaller language editions" of Wikipedia (Roth, 2012). The data also promised benefits for "numerous external applications, especially for annotating and connecting data in the sciences, in government, and for applications using data in very different ways" (Roth, 2012). In a *TechCrunch* article published at the time, Wikidata was described as a "new effort . . . (that) will bring all the localized versions of Wikipedia on par with each other in terms of the basic facts they house" (Pérez, 2012).

Core to the Wikidata project was a simple idea: instead of replicating the same information across multiple-language versions of Wikipedia and other Wikimedia projects (such as Wikimedia Commons, which hosts images, video, and other multimedia), this information could be held centrally in one common data store. For example, instead of having to update

[1]University of Technology Sydney, Australia
[2]Temple University, USA

Corresponding Author:
Andrew Iliadis, Department of Media Studies and Production, Klein College of Media and Communication, Temple University, Philadelphia, PA 19122, USA.
Email: andrew.iliadis@temple.edu
Twitter: @andrewiliadis

the new president of Brazil after a national election in more than 300 language versions of Wikipedia or replicating all language links to the same article about Brazil in the left-hand sidebar, such data could easily be updated on Wikidata and then "pushed" to all the pages on the Wikimedia platform that used it. According to its creators, such a feature would improve the efficiency and economy of wiki projects using a centralized data store from which any wiki project could retrieve data.

Yet, Wikidata would not only have efficiency implications. Wikidata was an entirely different project, built on an original platform and constructed using new software called Wikibase. Wikibase is an open database tool for working with a "knowledge base" of structured data in a central repository rather than the kind of unstructured natural language contained in MediaWiki (the software housing Wikipedia). As a result, volunteers needed new skills and expertise (in databases, structured data, and knowledge graphs, in particular), and new rules needed to be established. At first, Wikidata had few rules; the project was framed by its engineering founders simply as a "technical" project: it was "the largest technical project ever undertaken by one of the 40 international Wikimedia chapters," according to then-CEO of Wikimedia Deutschland, Pavel Richter (Roth, 2012). Yet, it soon became apparent that Wikidata was much more than a mere technical project (Tharani, 2021)—an interwiki link, for example, plays a functional role in moving users between one language version and another and is also a statement that two entities are semantically identical.

English Wikipedia, for example, distinguishes between Tibet (the geographic region) and the Tibet Autonomous Region (the political entity), but other-language Wikipedias distinguish them differently. How should Wikidata bring these concepts together in its centralized database? The wholesale extraction of data items by bots can introduce changes in meaning, and when data are stored outside of one community of practice in another where it has different implications, disputes, and debates are bound to arise (Ford & Graham, 2016). The repercussions are compounded when this information is scaled across other media products.

As a fact-facilitating infrastructure, Wikidata also has implications for large online platforms like search engines and virtual assistants. In a 2021 *Wired* article, Richard Cooke (2021) wrote, "As platforms like Google and Alexa work to provide instant answers to random questions, Wikidata will be one of the critical architectures that link the world's information together." An earlier piece in the magazine from 2019 titled "Inside the Alexa-Friendly World of Wikidata" stated that "Virtual assistants do their jobs better thanks to Wikidata, which aims to (eventually) represent everything in the universe in a way computers can understand" (Simonite, 2019). The ways that Wikidata classifies phenomena have a ripple effect on how media, such as search engines and virtual assistants retrieve and directly convey facts, answers, meanings, and knowledge about things in consumer products, resulting in those same products no longer leading people to other sources (Iliadis, 2022). Such platformized versions of fact-production (e.g., receiving Wikidata facts via a search engine or virtual assistant) may also include contexts where facts can be and often are contested, such as in health care and political decision-making (Tripodi, 2022). By facts, we mean truth claims that are autonomous (they have meaning and can exist on their own), short (usually consisting of a subject, object, and qualifier), and specific (relating to a particular area of knowledge) (Ford, 2022, p. 6). The present article theorizes Wikidata as a critical mediator of truth claims on the web today with significant social and political implications. We do this by framing Wikidata as a *semantic infrastructure* used to convey factual data and as the key means by which Wikipedia and its attendant meanings have become infrastructural.

Such semantic infrastructures comprise web data that have been formally organized and labeled as facts using some structured/regulated ontological classification system and its attendant tools (Allhutter, 2019; Iliadis, 2018, 2019; Poirier, 2019; Waller, 2016). While Wikipedia has had infrastructural characteristics from its early beginnings, Wikidata has expedited Wikipedia's transformation into a data infrastructure for conveying facts. The problem is that when Wikidata distributes its facts through the web in discreet bits of atomized information (what are known as "semantic triples" representing subjects, predicates, and objects as found in databases), such facts are no longer permanently anchored to the references to which they were initially connected (McMahon et al., 2017), the verifiability principles on which they were founded (Ford, 2020; Wikipedia, 2023), and narratives that lie beneath facts (Ford, 2022).

Drawing from work on infrastructure and platform studies (Helmond, 2015; Plantin et al., 2018), as well as from research on semantic platformization (Iliadis, 2022; Iliadis & Acker, 2022; Iliadis et al., 2023), we articulate Wikidata as a key actant in the remaking of social and political relations in the context of datafication and artificial intelligence (AI) across the broader web in media products like search engines and virtual assistants. While Wikidata is studied in fields, such as information and computer science (Erxleben et al., 2014; Pellissier Tanon et al., 2016), up to now, Wikidata and its politics are studied by only a few media and communications scholars relative to its central role in search (Ford, 2020, 2022; Iliadis, 2022; McDowell & Vetter, 2022, forthcoming), though some studies that reference Wikidata are beginning to appear in this domain, such as those seeking to analyze gender inequality in Wikipedia (Konieczny & Klein, 2018; Tripodi, 2023).

Our contribution to the field is to bring Wikidata into conversation with media, communication, and infrastructure/platform studies by articulating the numerous ways that semantic infrastructures matter for how mediation occurs across media platforms and, thus, how we communicate and relate to one another. Our explanations of how Wikidata operates as semantic infrastructure within a larger web ecosystem dominated by large, commercial platform companies that extract its data offer new interpretations of the fragility of data infrastructures and the importance of human-curated datasets in the context of AI debates.

Wikidata as Semantic Infrastructure

Infrastructures typically refer to shared public services like sewers, telephone poles, and electricity. According to Bowker et al. (2010, p. 98), information infrastructure refers to "digital facilities and services usually associated with the internet." Information infrastructures are thus enabling resources in network form, whose key role is that of a distributor. But rather than goods or services, information infrastructures distribute "knowledge, culture, and practice" (Bowker et al., 2010, p. 114). Such structures may do this by developing labeling or classification schemes that divide the world into categories that are then offered as an enormous, open store of data that others can query for various purposes, such as retrieving facts and sharing information. Recently, several scholars have elaborated on the political nature of such infrastructural processes of digitization and datafication, including in the domains of archiving and preservation (Thylstrup, 2019, 2022), governance and management (Flyverbom & Murray, 2018), metrics and sorting (Alaimo & Kallinikos, 2021), and the creation of global ontologies for things like web search (Iliadis et al., 2023) and surveillance services (Iliadis & Acker, 2022).

Wikidata constitutes infrastructure in terms of its significant scale, long-term sustainability, and democratic goals as a shared public good (Bowker et al., 2010; Plantin & Punathambekar, 2019), and it is an enabling resource for housing facts from which popular search products may distribute facts across the wider web. Wikidata (2022c) aims to cover "the diversity of knowledge" for "anyone in the world", and, at the time of writing, it describes over 100 million entities (things like people, places, and products described in the knowledge base, including their relationships). In this, Wikidata appears similar to other global structured data projects, such as Schema.org, which seek to turn the natural language of websites and applications into structured "factual" data that can be retrieved by search engines and virtual assistants (Iliadis et al., 2023). Yet, one of the key differences between the projects is that Wikidata is a database of facts, whereas structured data like Schema.org are used to annotate facts on the broader web. Sustainability is another characteristic of Wikidata relevant to infrastructures; Wikidata focuses on long-term and foundational resources rather than growth. Wikidata's introduction page describes the infrastructure as "an ongoing project that is under active development" (Wikidata, 2022c), and there is a vibrant Wikidata community involved in building and improving content, producing newsletters, discussing technical challenges, and meeting for annual Wikidata conferences.

Like many large-scale infrastructures (Bowker et al., 2010), Wikidata is connected to shared, sustainable infrastructure developed as a public good. One of Wikidata's guiding principles is openness; it uses the Creative Commons Zero, otherwise known as the CC0 license, which reserves no copyright, even for commercial purposes. This license offers additional freedom to downstream users and developers compared to Wikipedia since, unlike Wikipedia, Wikidata *does not require its data to be attributed*. Wikipedia uses the Creative Commons Attribution Share-Alike license that enables follow-on reuse even by commercial parties but requires those who use the content to license the resulting work under the same Attribution Share-Alike license. The share-alike provision is foundational to "copyleft," the legal technique for granting certain freedoms over copies of copyrighted works with the requirement that the same rights are preserved in derivative works. Wikidata, in contrast, decided early in its history to license its content under the CCZero Public Domain Dedication, which waives all rights to the work worldwide under copyright law. Thus, Wikidata is developed to enable reuse by external parties without the need for attribution or reciprocity. The service receives millions of visits and requests per year (Wikimedia, 2023) as it can be visualized via the Wikidata Query Service and connected to apps and platforms via application programming interfaces (APIs).

There is a curious tension concerning how we might categorize Wikidata as an infrastructure. For instance, Bowker et al. (2010) describe information infrastructures (with an emphasis on serving scientific fields), while Plantin and Punathambekar (2019) introduce the concept of "media infrastructures" as "the infrastructures that undergird and sustain media and communication networks and cultures across the world" (p. 165). Whereas Bowker et al.'s (2010) information infrastructures have included items like computational services, help desks, and data repositories, media infrastructures refer specifically to digital platforms characterized by programmability, generativity, and reliance on users' participation. Uniquely, Wikidata lies between these two types of infrastructure, serving scientific ventures (e.g., building metadata around diseases like COVID-19) and media and communication services (notably, Google). Wikidata's uses are thus vast; according to its developers, its data are used for "accessing basic information about a concept, machine learning, data cleaning and reconciliation, data exploration and visualization, tagging and entity recognition as well as internationalization of content" (Vrandečić et al., 2023, p. 617).

Wikidata is programmable in terms of its content and software, and its outputs are available to scientists who use its data and the public who access Wikidata facts via online platforms. One of Wikidata's key founding goals was to enable third parties to use its structured data freely—in other words, Wikidata is entirely accessible to third parties. Internally, the data are programmable by anyone, with a few restrictions reserved for administrations or long-term volunteers. These volunteers can build tools to extract or maintain data. Meanwhile, Wikidata is also generative, in that, the outcome of interactions is not necessarily known in advance; all significant platforms benefit from Wikidata's open infrastructure (excluding those in restrictive countries like China, where it is banned). According to its founders, Wikidata is "a hub in the Linked Data Web and beyond," connecting to over 7,500 other websites, catalogs, and other databases, including OpenStreetMap, MusicBrainz, and Scribe, Big Tech companies like Amazon, Apple, Google, IBM, and OpenAI, user-generated media sites like Reddit, Wolfram Alpha, and Twitter, and cultural/educational institutions, including the Internet Archive and the Smithsonian (Vrandečić et al., 2023, pp. 616–617).

Several scientific and academic projects use Wikidata, from constructing knowledge graphs for COVID-19 information (Turki et al., 2022) to the WikiGenomes project, a "freely open, editable, and centralized model organism database for the biological research community," powered by Wikidata (wikigenomes.org). Computer scientists have attempted to use knowledge bases such as Wikidata as a source of factual knowledge to train language models in machine learning (Safavi & Koutra, 2021). In the digital humanities, Thornton and Seals-Nutt's (2018) Science Stories project uses Wikidata to tell multimedia stories about scientists, including those from underrepresented groups. Wikidata relies on such users' participation to add, maintain, and translate structured data content, develop software tools, and educate users. Individuals are invited to contribute as "editors" (adding or editing data), "developers" (building Wikidata bots to automate tasks or contributing to Wikibase), or "ambassadors" to "spread the word about Wikidata to others, answer questions about the project, and serve as educational resources for Wikidata" (Wikidata, 2022b).

Wikidata, then, shares the characteristics of both media and information infrastructures. It is unique because it is a *semantic infrastructure* that produces facts using an ontological classification system for structured data, which then serves these facts to search engines and virtual assistants. Semantic infrastructures "allow for, enable, and afford in the creation and transmission of facts in semantic media products" that everyday consumers use (Iliadis, 2022, p. 22). Such infrastructures might include companies' "proprietary databases (databases of facts that companies own), web schemas (which administrators use to mark up their pages for retrieval by these companies), and open data repositories (free-to-use data collections of publicly sourced facts)," all of which can be used by products like Google Search or Amazon Alexa to present facts (Iliadis, 2022, p. 24). Wikidata, as a semantic infrastructure, can thus answer complex factual questions about the world. These statements of fact are described in terms of subjects, predicates, and objects (Wikidata, 2022a)—for example, in the sentence "The earth is round," "earth" is the subject, "round" is the object, and "is" is the predicate. On Wikidata, a subject is called an "item," the predicate is called a "property," and the object is called the "value." The subjects and items are things, the predicates and properties are relationships, and the objects and values are other things. These kinds of semantic infrastructures have been studied in contexts that include scientific research and open and linked data across the web (Allhutter, 2019; Iliadis, 2018, 2019; Iliadis et al., 2023; Poirier, 2019; Waller, 2016).

Semantic triples allow users to query Wikidata and ask questions about things like family relationships of notable individuals (e.g., how they are related to one another). Information about people on Wikidata often includes statements concerning their country of citizenship, where they were born, their occupation, where they went to school, their city of residence, their ethnic group, or to which political party they belong. These details represent the information stored as structured data on each person's Wikidata page. While some of this information also occurs in the everyday natural language of Wikipedia pages, the Wikidata pages that contain this structured data allow for complex semantic querying. Wikidata allows automated machines like search engines and digital assistants' APIs to look up semantically rich questions. Examples include "Names of 100 cities with a population larger than 1,000,000 in the native languages of their countries," "Current U.S. members of the Senate with district, party, and date they assumed office," and "People who lived in the same period as another person."

Furthermore, Wikidata is not only infrastructure but also the primary mechanism by which Wikipedia has become *infrastructural*. Wikipedia had infrastructural tendencies from its earliest days and quickly developed into a project that was vast in scale. The project fulfilled multiple infrastructure characteristics: it had a significant, long-term goal (representing "the sum of all human knowledge") and established itself as a not-for-profit without advertising and with global public education goals early on. DBpedia (another project that seeks to extract structured data from Wikipedia) began long before Wikidata. Yet, Wikidata's germination in Wikibase (the relatively user-friendly platform that "anyone can edit") and its storage of Wikipedia data established a "rocket fuel for facts" (Ford, 2022, p. 8). Wikidata was first populated by extracting and housing data from Wikipedia, such as interwiki links and infoboxes. Now, the links between the same article in multiple Wikipedia language versions and the facts in Wikipedia's infoboxes are centrally located in Wikidata and pushed to Wikipedia when a user queries an article. Wikidata effectively datafied Wikipedia content, so that, it can be pulled easily into other uses by third-party providers.

Wikidata facts are represented in popular media products, such as Google Search and Google's Knowledge Graph. For example, various search engine optimization (SEO) industry sources and academic research articles discuss Wikidata's importance for integrating factual data with such search engines and virtual assistants (Barnard, 2020; Barysevich, 2021; Cazier, 2016; Clark et al., 2022; Edward, 2015; Hogan et al., 2021; Kopp, 2022a, 2022b; Pecánek, 2020; Poddębniak, 2023), including Google research sources (Pellissier Tanon et al., 2016) and sections of the Google Knowledge Graph Search API page. Understanding how Wikidata creates, stores, and transmits these facts should be a key concern for media researchers trying to understand who can now shape our knowledge of the world and by which new tactics and processes. Data provided by search engines and virtual assistants are increasingly used to package expertise and to guide attention. Such techniques are apparent in semantic infrastructures like Wikidata, whose categories and logics scale across multiple sites and platforms that use its data.

Studying Wikidata as infrastructure thus helps recognize how mediation depends on shared organizing principles that connect seemingly unconnected entities. As a semantic infrastructure for facts shared across the web, Wikidata is also increasingly becoming a site of political struggle. In calling attention to Wikidata's politics, we refer to how its infrastructural relations affect outcomes in the world. Two fundamental power relations in theorizing semantic infrastructures are pertinent here. First, semantic infrastructures' power/knowledge nexus concerns the "new ways of knowing across information infrastructures" (Bowker et al., 2010, p. 113) because of the shared use of information services. Such is the ontological dimension of infrastructure, where "the nature of knowledge work is changing with the introduction of new information technologies, modes of representation, and the accompanying shifts in work practice and systems for the accreditation of knowledge" (Bowker et al., 2010, p. 105). The second is about labor, where studies have surfaced on the many new roles that have emerged in building semantic infrastructure and how the work and workers needed to maintain infrastructure are often undervalued. This undervaluing is the social dimension of semantic infrastructure where "new forms of sociality are being enabled/shaped by and shaping" new technologies along with new communities of knowledge workers, attendant studies of distributed collaborative practice, and the new relations that develop (Bowker et al., 2010, p. 105).

Wikidata is a valuable case for understanding the contingencies of labor and knowledge outcomes about semantic infrastructure. Like other platform infrastructures, Wikidata relies on largely invisible labor and maintenance. Yet, this invisibility and undervaluing are contingent on critical features of open data and the current web ecosystem relating to copyright, the web's organizational structure, and knowledge representation practices via machine learning. Similarly, Wikidata's introduction of new ways of knowing is contingent on how ontologies organize data, the design of participation mechanisms, and the pursued data collection, curation, and maintenance practices. In each case, epistemic principles, rules, and training are crucial to how power is exercised and experienced, ultimately affecting outcomes in the world (Abizadeh, 2023).

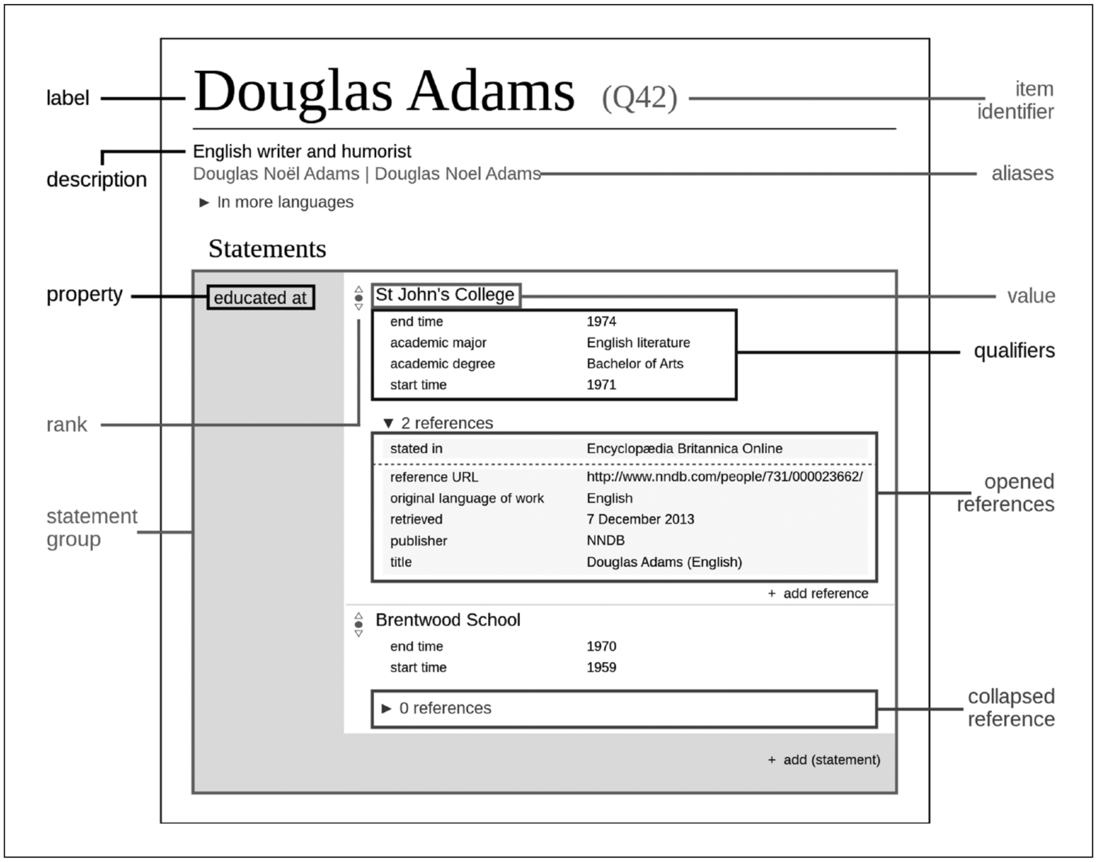
Knowledge Representation

Questions of knowledge representation have always been central to Wikipedia as an enterprise (McDowell & Vetter, 2021). Similarly, Wikidata classifications are influential because of their infrastructural function; they reverberate through the internet and have significant consequences for those they represent. Questions of representation involve who is producing data, who is silenced by data, and how meaning is closed through data production. Wikidata, for example, follows Wikipedia in its Western, male biases; Ahmed and Poulter (2022) find "just under four times as many statements about Western artists as non-Western artists, and nine times as many statements about Western as non-Western masterpieces" (p. 12). Furthermore, Wikidata's role in constructing knowledge lies beyond the number of statements about individual groups. Wikidata also has a material and symbolic impact on the foundational principles and rules by which truth is determined. As Monea (2016) has shown, other modern knowledge bases, such as Google's Knowledge Graph, articulate an underlying conservative logic of representation that prevents the feature from meaningfully discerning differences that do not conform to the subject-predicate-object structure of semantic triples while still other scholars, including Wikimedians, are likewise critically examining knowledge representation in Wikidata. Anasuya Sengupta is co-director (with Adele Vrana and Siko Bouterse) of the Whose Knowledge? (2022) project, which seeks to uncover sociocultural biases and decolonize Wikidata by focusing on knowledge justice for structured data. For example, Graham and Sengupta (2017) ask, "We're all connected now, so why is the internet so white and western?" In answering such questions, infrastructure studies have outlined how classification has social and political consequences because it creates boundaries for identities (Khan et al., 2022; Larkin, 2013). In this section, we outline how Wikidata's introduction of new ways of knowing is contingent on the design of Wikidata's interface, how data are organized in ontologies, and its policies and principles.

Interface Design

Each subject and object on Wikidata must have a structured data entity page; in this respect, like other databases, Wikidata is brittle in that the knowledge it contains must conform to its semantic data model (based on identifying subjects, predicates, and objects in semantic triples). Figure 1 shows the metadata associated with a typical Wikidata entity expressing the semantic data model.

Each entity in Wikidata is assigned a "Q" number; for example, in Figure 1, the "Q" identifier number (Q42) is for identifying the entity "Douglas Adams" (the famous author of *The Hitchhiker's Guide to the Galaxy*). There are also the relevant information terms around that entity that include the label ("Douglas Adams"), description ("English writer and



Douglas Adams (Q42)

English writer and humorist
Douglas Noël Adams | Douglas Noel Adams
► In more languages

Statements

educated at

St John's College

end time	1974
academic major	English literature
academic degree	Bachelor of Arts
start time	1971

▼ 2 references

stated in	Encyclopædia Britannica Online
reference URL	http://www.nndb.com/people/731/000023662/
original language of work	English
retrieved	7 December 2013
publisher	NNDB
title	Douglas Adams (English)

+ add reference

Brentwood School

end time	1970
start time	1959

► 0 references

+ add (statement)

Figure 1. Wikidata entity and its structured data representing facts (Wikipedia).

humorist"), property ("educated at"), value ("St John's College"), qualifiers ("academic major"), and other information. By way of another, more abstract example, the Wikidata object for "everyday life" is "Q1129653" and is described as "routine processes in humans' daily and weekly cycle." In Wikidata, "everyday life" is a subclass of "human condition," an instance of a "habit," part of a "personal life," has the quality of being "diurnal," and is said to be the same as a "daily routine." In these ways, Wikidata employs an ontological approach to break down the unstructured language found in Wikipedia articles and the entire knowledge base into discrete units of semantic triples. It is as if Wikidata has condensed the "bare facts" from the wealth of information in Wikipedia into a highly organized structure. As a result, narratives, detailed historical backgrounds, academic references, news stories, and in-depth descriptions of individuals' lives, including their upbringing and manner of death, are no longer present. What remains is a database of interconnected facts that can be queried, providing individuals with quick answers, concise knowledge, and limited data specific to their inquiries. In this manner, Wikidata serves as the foundational ontological and semantic infrastructure that supports Wikipedia's more elaborate and descriptive natural language content.

Rather than enabling any single language version to dominate the production of statements about a particular entity, Wikidata was designed to allow innumerable statements to sit equally in the database (Vrandečić & Krötzsch, 2014). Yet, as explained above, Wikidata is a central (single) store of facts from all language versions of Wikipedia. Even though editors from language versions other than English can create statements in Wikidata, they cannot equally discuss and debate more significant issues relating to how the project is designed, how disagreements can be resolved, and so forth. Although multiple statements and labels can exist on Wikidata, statements in languages like English dominate because they are the facts selected by third-party platforms. When Wikidata was first launched, Graham (2012) wrote in *The Atlantic* that it was significant that Wikidata provided only a single (English-language) space for the discussion of entities and that this represented a substantial move from Wikipedia's multiple-language discussion spaces:

> Look, for instance, at the Wikipedia pages about the Bronze Statue of Tallinn (a highly controversial moment in Estonia's history that sparked one of the world's first "cyberwars" between Russia and Estonia). The Estonian and Russian versions of that article present interestingly different versions of the very same place and events. The Arabic and Hebrew articles about Hezbollah offer perhaps an even starker contrast of the ways in which different communities of editors agree on different types of representation and truths.

Wikidata would, therefore, necessarily be obscure to those who can participate in their local language Wikipedias but cannot help curate the data and facts that are ultimately being controlled remotely on Wikidata.

Ontologies

Wikidata plays a central role in classification, employing multiple classification levels within the project. Scholars have developed modeling theories to examine the taxonomic hierarchies present in Wikidata, emphasizing the importance of comprehending its taxonomy and ontology (Brasileiro et al., 2016). This understanding is crucial because it contributes to categorizing and labeling factual information. Consequently, the sociopolitical ramifications are intertwined with the truthfulness and semantic meaning of facts conveyed through Wikidata, which serves as a contemporary philosopher, defining and describing the existence and interpretation of things through semantic media, such as knowledge panels and virtual assistants. Wikidata operates as a knowledge base with an overarching ontology known as an "upper-level" ontology, which serves as a repository for domain-specific information. For instance, the Blood Ontology in the medical field and the Financial Industry Business Ontology in commerce aim to organize metadata categories related to their respective subjects and the entities within those domains. At times, these diverse ontologies need to be interoperable to combine the referenced data. This is where upper-level ontologies, such as the one provided by Wikidata, prove valuable. These ontologies establish abstract methods of categorizing entities and relationships that are independent of specific domains. Only a select few upper ontologies have gained widespread usage in the field of information science. These include the Basic Formal Ontology (BFO), which is employed in bioinformatics and intelligence domains and has recently been approved as a standard by the International Organization for Standardization. Another notable upper-level ontology is the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), which was one of the earliest ontologies used to represent common-sense views of reality across domains like manufacturing, financial transactions, and cultural heritage. The Suggested Upper Merged Ontology (SUMO) is also worth mentioning, as it is the only upper-level ontology linked to the WordNet lexicon. While these ontologies have been utilized in various domains to facilitate interoperability among datasets, prominent internet platform companies primarily rely on the Wikidata ontology.

Wikidata is arguably one of the web's most extensive upper-level ontologies, and the ontology page on Wikidata indicates that it aspires or will attempt to reconcile other top-level ontologies (like BFO, DOLCE, etc.). The WikiProject Ontology page states that the Wikidata ontology is "mainly about reaching deep into the nature of being, becoming, existence, and reality, and applying those insights during Wikidata's maintenance tasks" (Wikidata, 2022d). The practical aim that it describes includes: "to support a broad semantic interoperability between notable ontologies like DOLCE, BFO, SUMO, Lemon, RDA, etc.," "to build consensus around the main branches of our core concept tree and how they relate to each other," and "to gain a deep understanding about the meaning of our upper ontology and to transfer this knowledge to others in practical terms" (Wikidata, 2022d). The description of Wikidata's ontology suggests it is the highest-ranking or "ultimate" ontology among all upper ontologies. The feasibility of achieving a universal upper-level ontology is debatable, as it is challenging, if not impossible, for everyone to reach a consensus on a single ontology and method of categorizing the world. However, it could be argued that the "winner" in the competition among upper-level ontologies is simply the ontology that gains the widest adoption in information science and on the web. In this context, Wikidata may be considered the winner, given that Wikipedia is one of the most significant collaborative endeavors on the internet.

Wikidata's top-level ontology includes multiple phenomena that are fundamentally subjective or at least contestable, depending on how they are applied as facts to entities in the knowledge base. Many of these vocabulary terms are ambiguous, and one can easily imagine disagreements concerning their definitions and applications. Below are a selection of these entities and their labels on Wikidata, which can be retrieved using Wikidata's SPARQL query service:

1. Meaning ("nature of meaning in the philosophy of language, semantics, metaphysics and metasemantics")
2. Hypothetical entity ("entity whose existence is possible, but not proven")
3. Pricelessness ("state or condition of being priceless; very high value")
4. Greatness ("concept of a state of superiority affecting a person or object in a particular place or area")
5. Acceptability ("characteristic of a thing being subject to acceptance for some purpose")
6. Bad ("item with negative value to the consumer")

Vocabulary terms, such as those described above are not located in an obscure lower area of the ontology. After the superclass root term "entity," they are included in the top two levels, and as such, these entities may be connected to facts and other entities in Wikidata. What some entities mean and how they might be applied to other entities and contexts where their meaning may change is not entirely clear. What qualifies for something to be connected to facts, such as "bad," "disability," "greatness," and "heritage?" Whereas Wikidata enables multiple statements to be created by editors to supposedly enable knowledge plurality, a more essential and undergirding knowledge creation happens at the level of the ontology itself, where relationships are instantiated and definitions restricted in ways that are hidden to the average

Wikidata user. There is no way to exist outside of the brittle definitions offered by the Wikidata ontology, which remains obscure to most everyday web users.

The knowledge representation of Wikidata is thus contingent on the meanings established in its ontologies and categories that express "facts," and these items serve as the underlying fact-transmitting infrastructure for online platforms. Wikidata's interface for interacting with these ontologies enables new kinds of infrastructure studies where scholars can examine the semantics of facts, how they are produced, and ultimately distributed in the broader web. In these ways, Wikidata acts as a critical mediator of knowledge and a gatekeeper of information, providing infrastructure for consumer products that answer people's questions.

Data Labor

The search engines, streaming services, and AI chatbots that people increasingly use to navigate everyday life rely on the essentially human labor of labeling, modeling, and screening information via projects like Wikidata. Data processing (e.g., labeling and modeling) of semantic infrastructure is a crucial feature for web platforms that organize information, including streaming services and search engines (Eriksson et al., 2019). Platforms require the labor of workers to build these data models, which act as mediated layers that signal/highlight content (e.g., Google surfacing facts from Wikidata in its knowledge panels).

Infrastructure studies have surfaced two critical features of labor that are worth exploring about Wikidata. First, infrastructures require the development of new functions and responsibilities that produce new types of expertise. Studies have exposed the data labor of "working ontologists" (Allhutter, 2019), from cleaning the datasets that feed machine learning to those curating data for science and information management, as well as social media (Arrieta-Ibarra et al., 2018; Denton et al., 2021; Geiger et al., 2020; Gray & Suri, 2019; Hutchinson et al., 2021; Iliadis, 2019; Irani & Silberman, 2013; Jones et al., 2022; Miceli et al., 2020; Plantin, 2019; Sambasivan et al., 2021). New categories of work that often remain hidden bring with them new labor relations within the Wikimedia ecosystem and beyond, between volunteer and paid labor, non-profit and for-profit information services, and between volunteers on different projects in the Wikimedia community. A closer look into the types of work performed by these laborers and the invisibility of this work across other relations highlights several contingencies for determining the types of processes performed and the obscurity of these connections.

Work to maintain Wikidata's knowledge base includes new labor categories, distinct from work on Wikipedia. Working on Wikidata, for example, is constrained by database and ontology logics. In contrast, Wikipedia work "is constrained by meeting Wikipedia's standards for formatting, quality, notability, neutrality, citations, etc., and the results are viewed primarily by humans" (Zhang et al., 2022, p. 2). On Wikidata, much of the work to populate fields is conducted by bots. However, on Wikidata, human editors "ensure that the information is accurate and consistent with the intended purpose of the field" (Zhang et al., 2022, p. 2). Most of the work to populate content on Wikipedia is conducted by human editors, even though many tasks are performed by bots (Geiger & Halfaker, 2017). Zhang et al. (2022) thus classify two types of Wikidata workers: Architects and Masons, who differ on their level of desired cognitive engagement, their preference for collaboration or solo work, and their degree of domain expertise. Whereas Architects develop the ontological infrastructure, "modeling and proposing properties," Masons "perform data entry work [. . .] e.g., adding or editing items, references, and properties, or linking to external databases" (Zhang et al., 2022, p. 10). Zhang et al. found that, unlike Wikipedia editors, Wikidata editors in their study did not know how their contributions were used:

> End-users often interact with Wikidata contributions through other software-based tools that use all or part of Wikidata as a database. These invisible machine intermediaries make it difficult for contributors to access information about how their contributions are used by others (p. 14).

McDowell and Vetter (forthcoming) convincingly argue that when Wikidata extracted data from Wikipedia and housed it in a central store under a less restrictive copyright license that enables upstream users to use Wikidata without acknowledging the source of that data, it shifted relationships between Wikimedia volunteers and the information they were producing. This process has alienated editors from their labor, breaking "the fundamental agreement to donate labor under the guise of the commons and sharing" when the "product of that labor [is] utilized, transformed, monetized, and sold back to the community" (McDowell & Vetter, forthcoming). McDowell and Vetter refer to this process as a "re-alienation of the commons" and further state that Wikidata represents what Fuchs calls a "pseudo-commons" (Fuchs, 2020, p. 123) because it obscures its participation in exploitive digital capitalism. The linkages between Wikipedia-based editorial work and the ways that such commons-focused labor becomes subsumed under exploitive capitalistic business practices have also been examined by Lund (2017).

Second, Wikidata labor is invisible, even to Wikipedians, for two reasons. First, Wikidata's interactions with Wikipedia articles have not always communicated well at the article level. Some Wikipedia editors have complained that Wikipedia articles are being edited remotely (because infobox items, e.g., are being pushed from Wikidata to all Wikipedia articles containing those items) and that edits were not shown on editors' watch lists automatically, which meant that Wikipedia editors could not check them for accuracy. In an example detailed in Ford (2022, pp. 121–122), the

Wikipedia editor, SarahSV, found that one of Wikidata's bots was opposing her careful attempts to curate an infobox about the book *Night* that detailed Elie Wiesel's time in concentration camps with his father during the Holocaust:

> SarahSV noticed that the infobox was suddenly listing the genre as "autobiographical novel" when she had previously left the field empty. She had done so because scholars couldn't agree on the genre, and Holocaust deniers call it a novel in their efforts to fictionalize it. SarahSV investigated and realized that a Wikidata bot had copied the genre data (that it was an "autobiographical novel") from Italian Wikipedia, where it was classified as a novel. Because the English Wikipedia infobox had no field for genre, the Wikidata algorithm, seeing this as missing information that needed to be filled in, unhelpfully added the translated field to it, thereby undoing SarahSV's careful efforts. The problem, wrote SarahSV, was that she couldn't work out how to remove it, and it "wasn't easy to see when the change had been made; when I looked through the article history, the Wikidata addition was showing up in old versions of the article."

For SarahSV, Wikidata's workings were so obscure that figuring out how to correct the errors in standardizing meaning across multiple-language versions was challenging. This error was made because she had meaningfully left the field for genre blank, and the bot had coded this as missing data and filled it in, thereby significantly changing the meaning of the item and undermining the consensus reached by editors about its representation.

Wikidata is also obscure to some editors because talk pages are only available in English, limiting the ability of editors from other languages to discuss, manage and govern Wikidata (Graham, 2012). The entry on climate change (Q125928), for example, will contain statements and interwiki links from and to all the (currently) 91 language versions of that article on Wikipedia. When discussions need to develop consensus around inappropriate content or disagreements, they occur in English, resulting in obscurity for non-first-language English speakers.

Wikidata is invisible to the public for yet other reasons. Facts presented in popular platforms like Google, Bing, and Alexa that may have originated in Wikidata are not required to have been attributed because the Creative Commons Public Domain Dedication governs its content. This dedication means that, even if facts in Google's knowledge panels were sourced from Wikidata, they would not need to carry a citation or a link back to Wikidata like Wikipedia requires with its Creative Commons Attribution Share-Alike license. Facts originating in Wikipedia may also lose their attribution via Wikidata since Wikidata extracts facts from Wikipedia, which requires attribution, but stores them in a system that does not require attribution.

Wikidata also exacerbates the problem of disappearing provenance data because it has also lost many citations to its statements extracted initially from Wikipedia (Ford, 2020; Kolbe, 2015). This process happened because the primary mechanism for transferring content from Wikipedia to Wikidata is via automated bots. Because citation data are not well standardized, citations on Wikidata are more often to a whole Wikipedia project (e.g., Italian Wikipedia) than to a significant source (e.g., Thomas Hobbes's *Leviathan*). Wikidata thus demonstrates an irony of data infrastructure in the context of machine learning in that its goal of providing open factual data to third-party providers under a public domain dedication may, in the long term, result in its downfall; third parties can and do ingest its data without attribution (McMahon et al., 2017). Such parties can legally store Wikidata's facts in their proprietary databases and thus lose their dependence on Wikidata and Wikipedia as a source over time, potentially negatively impacting data workers and the political economy of data labor.

Wikidata content may also become invisible because its data are used to power machine learning and AI applications like Google's knowledge panels in Google Search, answers to Apple's Siri virtual assistant, and significant language models like ChatGPT and Bard. As Wikidata's content is ingested by knowledge graphs that power these applications, they merge data from different sources, lose the traces of their originating statements, and start to learn independently, generating new content for themselves. Some argue that Wikidata fulfills its function of making knowledge freely accessible to all, yet others say that Wikidata threatens Wikimedia's most significant asset (Wikipedia) when it makes its data available to third-party users using a more liberal public domain dedication (Kolbe, 2015). This action cuts Wikipedia off from donations of time and money when potential editors and donors do not visit Wikipedia to get the information served to them directly by the third-party platform. Others (Ford, 2020) have pointed to the most significant risk in weakening the verifiability principle because of Wikipedia's datafication via Wikidata. When Wikidata and third-party users like Google do not acknowledge the source of factual statements that seem to be "handed down from God" (Dewey, 2016), this effectively removes the rights of users to question where those facts have come from and the means of changing how facts are represented.

In summary, attending to infrastructures raises essential questions about the outcomes of semantic infrastructures in the context of labor necessary for their maintenance and functioning. For Wikidata (and its sister projects, Wikipedia, and Wikimedia Commons, etc.), the problems that arise when more powerful commercial platforms ingest its content are contingent on two key issues: the liberal copyright rules associated with Wikidata and the practices of knowledge representation via machine learning that weaken principles of verifiability and the problems of latency that prevent third parties from querying Wikidata directly and thus making their usage of the resource more transparent. Although Wikidata imagines itself as a sustainable project developed for the long term, there are threats to its continued existence when it serves third-party providers who may not need it in the future. This service introduces a significant weakness of semantic infrastructure projects like Wikidata developed in the context of machine learning practices. Wikidata labor practices and knowledge representation thus remain open for ethnographic and qualitative research and quantitative network analysis. More research is needed to understand Wikidata's data labor: how Wikidata interfaces with other Wikimedia projects, how this seemingly technical project introduces new meanings through the automation that dominates Wikidata practice, and the extent to which it can sustain itself in the face of increasingly powerful upstream commercial providers that use its data and the data of its sister projects without credit.

Conclusion

An article about the history of Wikidata was recently published from the perspective of some of its founding architects, at the 2023 World Wide Web Conference. The authors noted that despite the "subjective perspective" of the article, it would "still play a major role in filling the gaps, offering explanations, and deriving objectives for the future" (Vrandečić et al., 2023, p. 616). Yet, there are still gaps in understanding Wikidata as a significant force of meaning-making on the internet today. The authors gloss over the impact of Wikidata on Wikipedia ("There were fears and fantasies of complete automation of Wikipedia article writing, a forced uniformity and alignment across the different Wikipedia language editions, and the loss of nuance and cultural context in structured data," p. 619) and present Google as having "a lasting positive impact on the interest into Wikidata" (Vrandečić et al., 2023, p. 619), but these claims are yet to be investigated.

Understanding Wikidata as a semantic infrastructure can help to frame questions about the impact of Wikipedia and other public information sources' datafication. This work is essential to recognizing the implications of how facts are created, stored, manipulated, and distributed to online platforms like search engines and virtual assistants and, ultimately, how the world's knowledge is mediated. As Bowker et al. (2010, p. 99) note, understanding "the nature of infrastructural work involves unfolding the political, ethical, and social choices that have been made throughout its development." Wikidata matters for mediation because it has become the shared infrastructure for the travel of knowledge, and how it enhances facts has implications for their meanings and for the vocabulary available to people for describing the world to one another.

Recognizing Wikidata as a semantic infrastructure is also helpful in understanding the ultimate outcomes of such infrastructure. We asked what the contingencies for knowledge representation and data labor are in achieving infrastructural goals for Wikidata. Regarding knowledge representation, Wikidata's representations depend on how data are organized in ontologies, the design of participation mechanisms, and the data collection, curation, and maintenance practices. Regarding data labor, Wikidata relies on largely invisible work, invisibility exacerbated by its copyright rules, its role as a source for upstream, commercial platforms, and the current practices of knowledge representation using machine learning on the internet. But how third-party providers use Wikidata's labor is also obscure to editors, introducing an increased distance between the volunteers and the outcomes of their labor. In each case, Wikidata's goals, rules, and practices are crucial to how power is exercised and experienced. The meanings that it circulates (and the existence of a free pipeline for commercial platforms to train their data) depend on such socio-technical features.

Conceptualizing Wikidata as a semantic infrastructure is a valuable frame for the challenges it continues to face. Wikidata is fragile because it is developed as a public good and yet faces challenges from downstream platforms that, using AI, ingest its facts and structures without credit. We do not know what Wikidata's future looks like. Still, we know that its outcomes are contingent upon the material features of its copyright rules and its goal to serve as the top-layer ontology providing meaning across the internet. We have only sketched out two key areas in which Wikidata has come to matter to the practices of knowledge development. There is, for example, also a need to understand how meanings established on Wikipedia's multiple-language versions are interrupted or mistranslated when they move to Wikidata and upstream providers (such as the increasingly critical question-answering systems on board virtual assistants and smart speakers), how Wikidata's design features have evolved and continue to evolve, and how they impact on the availability of consensus and information diversity at local, regional, and global levels.

Future research on Wikidata is essential for two key reasons. First, it represents a critical case for studying the practices and implementations of knowledge automation and knowledge distribution in the context of an internet where a few commercial players heavily dominate. Second, there is a pressing need to examine alternatives to exploitive platform capitalism that produce valuable public knowledge goods. It is still an open question whether Wikidata represents this alternative. But it is clear that Wikidata is much more than simply a "technical" project and has significant implications for meaning-making in a future likely to be dominated by automated knowledge machines.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Temple University Grant-in-Aid.

ORCID iD

Andrew Iliadis https://orcid.org/0000-0002-8345-6251

References

Abizadeh, A. (2023). The grammar of social power: Power-to, power-with, power-despite and power-over. *Political Studies*, *71*(1), 3–19. https://doi.org/10.1177/0032321721996941

Ahmed, W., & Poulter, M. L. (2022). Representation of non-western cultural knowledge on Wikipedia: The case of the visual arts. *Digital Studies/Le Champ Numérique*, *12*(1), 1–27. https://doi.org/10.16995/dscn.8078

Alaimo, C., & Kallinikos, J. (2021). Managing by data: Algorithmic categories and organizing. *Organization Studies*, *42*(9), 1385–1407. https://doi.org/10.1177/0170840620934062

Allhutter, D. (2019). Of "working ontologists" and "high-quality human components": The politics of semantic infrastructures. In D. Ribes & J. Vertesi (Eds.), *DigitalSTS: A field guide for science & technology studies* (pp. 326–348). Princeton University Press.

Arrieta-Ibarra, I., Goff, L., Jiménez-Hernández, D., Lanier, J., & Glen Weyl, E. (2018). Should we treat data as labor? Moving beyond "free." *American Economic Association Papers & Proceedings*, *108*, 38–42. https://doi.org/10.1257/pandp.20181003

Barnard, J. (2020, April 1). How to get your brand in Google's knowledge graph without a Wikipedia page. *Search Engine Journal*. https://www.searchenginejournal.com/get-brand-in-google-knowledge-graph-without-wikipedia-page/356530/

Barysevich, A. (2021, November 2). How to maximize your reach using Google's knowledge graph. *Search Engine Journal*. https://www.searchenginejournal.com/maximize-reach-using-googles-knowledge-graph//144579

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2010). Toward information infrastructure studies: Ways of knowing in a networked environment. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of internet research*. Springer. https://doi.org/10.1007/978-1-4020-9789-8_5

Brasileiro, F., Almeida, J. P. A., Carvalho, V. A., & Guizzardi, G. (2016). Applying a multi-level modeling theory to assess taxonomic hierarchies in Wikidata. In *Proceedings of the 25th international conference companion on World Wide Web* (pp. 975–980). https://doi.org/10.1145/2872518.2891117

Cazier, C. (2016, February 10). Wikidata 101. *Search Engine Land*. https://searchengineland.com/wikidata-101-241844

Clark, J. A., Williams, H. K. R., & Rossmann, D. (2022). Wikidata and knowledge graphs in practice: Using semantic SEO to create discoverable, accessible, machine-readable definitions of the people, places, and services in libraries and archives. *Information Services & Use*, *42*(3–4), 377–390. https://doi.org/10.3233/ISU-220171

Cooke, R. (2021, January 14). Wikipedia is the last best place on the internet. *Wired*. https://www.wired.com/story/wikipedia-online-encyclopedia-best-place-internet/

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, *8*(2), 1–14. https://doi.org/10.1177/20539517211035955

Dewey, C. (2016, May 11). You probably haven't even noticed Google's sketchy quest to control the world's knowledge. *The Washington Post*. https://www.washingtonpost.com/news/the-intersect/wp/2016/05/11/you-probably-havent-even-noticed-googles-sketchy-quest-to-control-the-worlds-knowledge/

Edward, T. (2015, May 1). Leveraging Wikidata to gain a Google Knowledge Graph result. *Search Engine Land*. https://searchengineland.com/leveraging-wikidata-gain-google-knowledge-graph-result-219706

Eriksson, M., Fleischer, R., Johansson, A., Snickars, P., & Vonderau, P. (2019). *Spotify teardown: Inside the black box of streaming music*. MIT Press.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., & Vrandečić, D. (2014). Introducing Wikidata to the linked data web. In *The semantic web–ISWC 2014: 13th international semantic web conference* (pp. 50–65). Springer. https://doi.org/10.1007/978-3-319-11964-9_4

Flyverbom, M., & Murray, J. (2018). Datastructuring—Organizing and curating digital traces into action. *Big Data & Society*, *5*(2), 1–12. https://doi.org/10.1177/2053951718799114

Ford, H. (2020). Rise of the underdog. In J. Reagle & J. Koerner (Eds.), *Wikipedia @ 20: Stories of an incomplete revolution* (pp. 43–54). MIT Press.

Ford, H. (2022). *Writing the revolution: Wikipedia and the survival of facts in the digital age*. MIT Press.

Ford, H., & Graham, M. (2016). Provenance, power and place: Linked data and opaque digital geographies. *Environment and Planning D: Society and Space*, *34*(6), 957–970. https://doi.org/10.1177/0263775816668857

Fuchs, C. (2020). The ethics of the digital commons. *Journal of Media Ethics*, *35*(2), 112–126. https://doi.org/10.1080/23736992.2020.1736077

Geiger, R. S., & Halfaker, A. (2017). Operationalizing conflict and cooperation between automated software agents in Wikipedia: A replication and expansion of "even good bots fight." In *Proceedings of the ACM on human-computer interaction* (pp. 1–33). https://doi.org/10.1145/3134684

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In *FAT* '20: Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 325–336). https://doi.org/10.1145/3351095.3372862

Graham, M. (2012, April 6). The problem with Wikidata. *The Atlantic*. https://www.theatlantic.com/technology/archive/2012/04/the-problem-with-wikidata/255564/

Graham, M., & Sengupta, A. (2017, October 5). We're all connected now, so why is the internet so white and western? *The Guardian*. https://www.theguardian.com/commentisfree/2017/oct/05/internet-white-western-google-wikipedia-skewed

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Harper.

Helmond, A. (2015). The platformization of the web: Making web data platform ready. *Social Media + Society*, *1*(2), 1–11. https://doi.org/10.1177/2056305115603080

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., José Emilio Labra Gayo Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). *Knowledge graphs*. Morgan & Claypool.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2021). Towards accountability for machine learning datasets. In *FAccT '21: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 560–575). https://doi.org/10.1145/3442188.3445918

Iliadis, A. (2018). Algorithms, ontology, and social progress. *Global Media and Communication*, *14*(2), 219–230. https://doi.org/10.1177/1742766518776688

Iliadis, A. (2019). The Tower of Babel problem: Making data make sense with Basic Formal Ontology. *Online Information Review*, *43*(6), 1021–1045. https://doi.org/10.1108/OIR-07-2018-0210

Iliadis, A. (2022). *Semantic media: Mapping meaning on the internet*. Polity.

Iliadis, A., & Acker, A. (2022). The seer and the seen: Surveying Palantir's surveillance platform. *The Information Society*, *38*(5), 334–363. https://doi.org/10.1080/01972243.2022.2100851

Iliadis, A., Acker, A., Stevens, W., & Kavakli, B. (2023). One schema to rule them all: How Schema.org models the world of search. *Journal of the Association for Information Science and Technology*. Advance online publication. https://doi.org/10.1002/asi.24744

Irani, L. C., & Silberman, M. S. (2013). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *CHI '13: Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 611–620). https://doi.org/10.1145/2470654.2470742

Jones, I., Hecht, B., & Vincent, N. (2022). Misleading tweets and helpful notes: Investigating data labor by Twitter Birdwatch users. In *CSCW'22 companion: Companion publication of the 2022 conference on computer supported cooperative work and social computing* (pp. 68–71). https://doi.org/10.1145/3500868.3559461

Khan, T., Abimbola, S., Kyobutungi, C., & Pai, M. (2022). How we classify countries and people—and why it matters. *British Medical Journal—Global Health*, *7*(6), 1–6. https://doi.org/10.1136/bmjgh-2022-009704

Kolbe, A. (2015, December 2). Whither Wikidata? *The Signpost*. https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2015-12-02/Op-ed

Konieczny, P., & Klein, M. (2018). Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata human gender indicator. *New Media & Society*, *20*(12), 4608–4633. https://doi.org/10.1177/1461444818779080

Kopp, O. (2022a, April 22). How does Google process information from Wikipedia for the Knowledge Graph? *Online Marketing Consulting*. https://www.kopp-online-marketing.com/wikipedia-knowledge-graph

Kopp, O. (2022b, April 8). Natural language processing to build a semantic database. *Online Marketing Consulting*. https://www.kopp-online-marketing.com/natural-language-processing-to-build-a-semantic-database

Larkin, B. (2013). The politics and poetics of infrastructure. *Annual Review of Anthropology*, *42*, 327–343. https://doi.org/10.1146/annurev-anthro-092412-155522

Lund, A. (2017). *Wikipedia, work and capitalism*. Springer.

McDowell, Z. J., & Vetter, M. A. (2021). *Wikipedia and the representation of reality*. Routledge.

McDowell, Z. J., & Vetter, M. A. (2022). Fast truths and slow knowledge: Oracular answers and Wikipedia's epistemology. *Fast Capitalism*, *19*(1), 104–112. https://doi.org/10.32855/fcapital.202201.009

McDowell, Z. J., & Vetter, M. A. (forthcoming). The re-alienation of the commons: Wikidata and the ethics of "free" data. *International Journal of Communication*.

McMahon, C., Johnson, I., & Hecht, B. (2017). The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 142–151. https://doi.org/10.1609/icwsm.v11i1.14883

Miceli, M., Schuessler, M., & Yang, T. (2020). Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction*, *4*(115), 1–25. https://doi.org/10.1145/3415186

Monea, A. (2016). The graphing of difference: Numerical mediation and the case of Google's Knowledge Graph. *Cultural Studies ↔ Critical Methodologies*, *16* (5), 452–461. https://doi.org/10.1177/1532708616655763

Pecánek, M. (2020, September 24). Google's Knowledge Graph explained: How it influences SEO. *Ahrefs*. https://ahrefs.com/blog/google-knowledge-graph/

Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T., & Pintscher, L. (2016). From freebase to Wikidata: The great migration. In *WWW '16: Proceedings of the 25th international conference on World Wide Web* (pp. 1419–1428). https://doi.org/10.1145/2872427.2874809

Pérez, S. (2012, March 30). Wikipedia's next big thing: Wikidata, a machine-readable, user-editable database funded by Google, Paul Allen and others. *TechCrunch*. https://techcrunch.com/2012/03/30/wikipedias-next-big-thing-wikidata-a-machine-readable-user-editable-database-funded-by-google-paul-allen-and-others/

Plantin, J.-C. (2019). Data cleaners for pristine datasets: Visibility and invisibility of data processors in social science. *Science, Technology, & Human Values*, *44*(1), 52–73. https://doi.org/10.1177/0162243918781268

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310. https://doi.org/10.1177/1461444816661553

Plantin, J.-C., & Punathambekar, A. (2019). Digital media infrastructures: Pipes, platforms, and politics. *Media, Culture & Society*, *41*(2), 163–174. https://doi.org/10.1177/0163443718818376

Poddębniak, G. (2023, May 12). How to get in the Google Knowledge Graph. *Onely*. https://www.onely.com/blog/google-knowledge-graph/

Poirier, L. (2019). Classification as catachresis: Double binds of representing difference with semiotic infrastructure. *Canadian Journal of Communication*, *44*(3), 361–371. https://doi.org/10.22230/cjc.2019v44n3a3455

Roth, M. (2012, March 30). The Wikipedia data revolution. *Diff*. https://diff.wikimedia.org/2012/03/30/the-wikipedia-data-revolution/

Safavi, T., & Koutra, D. (2021). Relational world knowledge representation in contextual language models: A review. *arXiv*. https://arxiv.org/abs/2104.05837

Sambasivan, N., Kapania, S., & Highfill, H. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *CHI '21: Proceedings of the 2021 CHI* *conference on human factors in computing systems* (pp. 1–15). https://doi.org/10.1145/3411764.3445518

Simonite, T. (2019, February 18). Inside the Alexa-friendly world of Wikidata. *Wired*. https://www.wired.com/story/inside-the-alexa-friendly-world-of-wikidata/

Tharani, K. (2021). Much more than a mere technology: A systematic review of Wikidata in libraries. *The Journal of Academic Librarianship*, *47*(2), 1–8. https://doi.org/10.1016/j.acalib.2021.102326

Thornton, K., & Seals-Nutt, K. (2018). Science stories: Using IIIF and Wikidata to create a linked-data application. In *Proceedings of the ISWC 2018 posters & demonstrations, industry and blue sky ideas tracks* (pp. 1–4). https://ceur-ws.org/Vol-2180/paper-68.pdf

Thylstrup, N. B. (2019). *The politics of mass digitization*. MIT Press.

Thylstrup, N. B. (2022). The ethics and politics of data sets in the age of machine learning: Deleting traces and encountering remains. *Media, Culture & Society*, *44*(4), 655–671. https://doi.org/10.1177/01634437211060

Tripodi, F. B. (2022). *The propagandists' playbook: How conservative elites manipulate search and threaten democracy*. Yale University Press.

Tripodi, F. B. (2023). Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, *25*(7), 1687–1707. https://doi.org/10.1177/14614448211023772

Turki, H., Hadj Taieb, M. A., Shafee, T., Lubiana, T., Jemielniak, D., Aouicha, M. B., Labra Gayo, J. E., Youngstrom, E. A., Banat, M., Das, D., & Mietchen, D. (2022). Representing COVID-19 information in collaborative knowledge graphs: The case of Wikidata. *Semantic Web*, *13*(2), 233–264. https://doi.org/10.3233/SW-210444

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489

Vrandečić, D., Pintscher, L., & Krötzsch, M. (2023). Wikidata: The making of. In *Companion proceedings of the ACM Web conference 2023 (WWW '23 companion)* (pp. 615–624). https://doi.org/10.1145/3543873.3585579

Waller, V. (2016). Making knowledge machine-processable: Some implications of general semantic search. *Behaviour & Information Technology*, *35*(10), 784–795. https://doi.org/10.1080/0144929X.2016.1183710

Whose Knowledge? (2022). *Decolonizing the internet's structured data*. https://whoseknowledge.org/resource/dti-structured-data-report/

Wikidata. (2022a, November 18). *Help:Statements*. https://www.wikidata.org/wiki/Help:Statements

Wikidata. (2022b, June 25). *Wikidata:Contribute*. https://www.wikidata.org/wiki/Wikidata:Contribute

Wikidata. (2022c, July 10). *Wikidata:Introduction*. https://www.wikidata.org/wiki/Wikidata:Introduction

Wikidata. (2022d, March 30). *Wikidata:WikiProject ontology*. https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology

Wikimedia. (2023, June 12). *Wikimedia statistics*. https://stats.wikimedia.org/#/wikidata.org

Wikipedia. (2023, February 22). *Wikipedia:Verifiability*. https://en.wikipedia.org/wiki/Wikipedia:Verifiability

Zhang, C. C., Houtti, M., Smith, C. E., Kong, R., & Terveen, L. (2022). Working for the invisible machines or pumping information into an empty void? An exploration of Wikidata contributors' motivations. In *Proceedings of the ACM on human-computer interaction* (pp. 1–21). https://doi.org/10.1145/3512982

Author Biographies

Heather Ford is Associate Professor at the University of Technology Sydney (UTS) in the School of Communications, Coordinator of the UTS Data and AI Ethics Cluster, Affiliate of the UTS Data Science Institute, and Associate of the UTS Center for Media Transition. She was formerly Google Policy Fellow at the Electronic Frontier Foundation and Executive Director of iCommons. She is the author of *Writing the Revolution: Wikipedia and the Survival of Facts in the Digital Age* (MIT Press, 2022).

Andrew Iliadis is Assistant Professor at Temple University in the Department of Media Studies and Production within the Klein College of Media and Communication and serves on the faculties of the Media and Communication Doctoral Program, Cultural Analytics Graduate Certificate Program, and Science, Technology, and Society Network. He is the author of *Semantic Media: Mapping Meaning on the Internet* (Polity, 2022) and co-editor of *Embodied Computing: Wearables, Implantables, Embeddables, Ingestibles* (MIT Press, 2020).