Selected Papers of #AoIR2022:
The 23rd Annual Conference of the
Association of Internet Researchers
Dublin, Ireland / 2-5 Nov 2022

# SEMANTIC MEDIA: POLITICAL ECONOMY PERSPECTIVES ON PLATFORMIZED FACT PRODUCTION

Andrew Iliadis
Temple University

Heather Ford
University of Technology Sydney

Doris Allhutter
Institute of Technology Assessment, Austrian Academy of Sciences

Zachary McDowell
University of Illinois Chicago

Matthew Vetter
Indiana University of Pennsylvania

Large media platforms are now in the habit of providing facts in their products and representing knowledge to various publics. For example, Google's Knowledge Graph is a database of facts that Google uses to provide quick answers to publics who use their products, while Wikipedia has a product called Wikidata that similarly stores facts about the world in data formats through which various apps can retrieve the data. Microsoft, Amazon, and IBM use similar fact storing and retrieval techniques in their products. This panel introduces papers that take a political economy perspective on such platformaized versions of fact production and examines the underlying infrastructures, histories, and modeling techniques used in such knowledge representation systems.

Knowledge representation, long a central topic in archiving work in library and information sciences, is a key feature of platforms and practiced by internet companies more broadly. Much of this work has historically centered on metadata models that seek to organize and describe information in standardized ways. In the context of expanding

Iliadis, A., Ford, H., Allhutter, D., McDowell, & Vetter, M. (2022, November). *Semantic Media: Political Economy Perspectives on Platformized Fact Production*. Panel presented at AoIR 2022: The 23rd Annual Conference of the Association of Internet Researchers. Dublin, Ireland: AoIR. Retrieved from http://spir.aoir.org.

this data organizing and labeling work into the wider web, one of the main facilitators was the “Semantic Web” project proposed by Tim-Berners Lee and the World Wide Web Consortium (W3C). Today, many of the same principles, technologies, and standards that were proposed by those early projects in metadata modeling from groups like W3C are found at companies like Google and Facebook, organizations like Wikipedia, government portals, and beyond.

These platform metadata models are typically produced by industry professionals (e.g., taxonomists, ontologists, knowledge engineers, etc.) who help structure information for algorithmic processing on platforms and their recommender systems. Such structured information is supposed to add a layer of contextual expressivity to web data that would otherwise be more difficult to parse, though the issue of context control is not unproblematic in relation to statements of facts. In many of these automated systems, metadata models contribute to articulating ready-made facts that then travel through these systems and eventually reach the products that are engaged by everyday web users. This panel connects scholars working in information, media studies, and science and technology studies to discuss these semantic technologies.

The first paper presents data gathered from interviews with semantic web practitioners who build or have built metadata models at large internet and platform companies. It presents results from a qualitative study of these platform data management professionals (collectively referred to as “metadata modelers”) and draws from unstructured interviews (n=10) and archival research. The paper describes the image of a metadata ecology along with selected work-related contestations expressed by interview subjects regarding some of the difficulties and intractable problems in metadata modeling work. The paper includes a discussion of the political economy of platform semantics through an examination of critical semantic web literature and ends with some policy concerns.

The second paper translates the method of tracing “traveling facts” from science studies to the context of online knowledge about evolving, historic events. The goal is to understand the socio-political impact of the semantic web as it has been implemented by monopolistic digital platforms and how such practices intersect in the context of Wikipedia, where the majority of knowledge graph entities are sourced from. The paper describes how the adoption (and domination) by platform companies of linked data has catalyzed a re-shaping of web content to accord with the question and answer linked data formats, weakening the power of open content licenses to support local knowledge and consolidating the power of algorithmic knowledge systems that favor knowledge monopolies.

The third paper discusses building a semantic foundation for machine learning and examines how information infrastructures that convey meaning are intimately tied to colonial labor relations. It traces the practice of building a digital infrastructure that enables machines to learn from human language. The paper describes examples from an ethnographic study of semantic computing and its infrastructuring practices to show how such techniques are materially and discursively performative in their co-emergence with techno-epistemic discourses and politico-economic structures. It examines sociomaterial process in which classifications, standards, metadata, and methods co-

emerge with processes of signification that reconstitute and/or shift hegemonic ecologies of knowledge.

The fourth paper evaluates and examines the ethics of “free” data (CC-0) in Wikidata by evaluating the sources and usage of data from and within Wikidata. From knowledge graphs to AI training, Wikidata is the semantic web platform that is being used across the Internet to power new platforms. Through a consideration of the ways in which Wikidata scrapes Wikipedia’s “share alike” knowledge through scraping metadata and the significant donations and partnerships from large technology firms (Google in particular), this paper addresses ethical concerns within the largest semantic web platform, how these transformations of knowledge alienate donated volunteer labor, and offers some ways in which these issues might be mitigated.

## “We Became What You Might Call the Semantic Guardians”: How Metadata Modelers Talk about Platform Content

Andrew Iliadis
Temple University

This paper draws from a qualitative study of platform data management professionals (e.g. taxonomists, ontologists, knowledge engineers, etc., herein collectively referred to as “metadata modelers”) who build or have built metadata models at large internet and platform companies (n=10). Using data gathered from unstructured interviews and archival sources, the paper focuses on metadata modeler experiences in data formalization, and their technical application on web platforms, aiming to provide a sociotechnical approach to understanding how metadata modelers view and talk about their metadata modeling work. Interviews were conducted over a two-year period with metadata modeling professionals at large platform companies or internet organizations. The interviews were conducted over an internet-connected telephone and recorded using CallGraph. They were then transcribed using Rev, and the transcribed interviews were then coded using an inductive grounded theory process (Strauss & Corbin, 1997).

Metadata models are an important yet hidden feature of web platforms (Eriksson, 2016). The ability to algorithmically search, retrieve, and reason with web data meaningfully often requires these machine-readable, manually crafted metadata models. These models provide a layer of contextual granularity that is supposed to enhance data accessibility, particularly as large amounts of unstructured and heterogeneous data proliferate via digital services and web products. Platforms use models to act as mediated layers that signal/highlight content through vocabularies and axioms expressed in richly curated metadata markup (e.g. Google surfacing news article ‘snippets’ in search results, Airbnb providing information about travel ‘experiences’ related to trips to a certain city, Uber Eats’ ‘your favorites’ recommendations, etc.), rather than having users or administrators do the mundane work of navigating generic or unmarked search results and/or parse documents for text.

Algorithmically culled content presented on these platforms often depend on semantic parsing via metadata models that have been manually structured on the administrative

backend for automated presentation on user frontend interfaces. This infrastructural enhancement is supposed to increase web data's contextual expressivity, assisting admins/users in finding the items for which they might be searching on a host platform without having to navigate away to another source (e.g., via in-platform knowledge panels or voice-assisted search). In these practices, scholars have long noted the interpretive and hermeneutic aspects of such data models and schemas (Acker, 2015), which are implicitly imbued with an *aboutness* (Hjørland, 2001) concerning the data to which they are affixed. Notably, there have been several classic studies by science and technology scholars that sought to critically examine data models and associated knowledge representation infrastructures using qualitative, ethnographic, and critical methods (Forsythe, 1993; Adam, 1998).

Internet companies and businesses more broadly have described facing challenges related to data integration and harmonization because of idiosyncratic data labelling practices, varieties of data typing (formats), and legacy software infrastructures. While issues related to limitations in data storing and warehousing are somewhat technically less problematic in recent years owing to engineering solutions related to data compression and server space evolution, the ability to logically and meaningfully reason with heterogenous data has created persistent semantic challenges for querying and retrieving large datasets. To exploit these large troves of diverse data that have been siloed or stored in data lakes, companies and organizations have turned to solutions in the form of web semantics standards that increase the scalability of their data reasoning capabilities.

New efficiencies in metadata representation, indexing, caching, and querying have evolved to become standards at large internet companies who now specialize in so-called "knowledge graphs" that present information about the world though semantic annotation and interoperability. Yet, metadata models are fraught with easily imaginable errors and difficulties that can emerge when multiple parties must coalesce around semantics and a determinate set of language rules. Language is embedded in historical, social, and cultural contexts, and context in human-computer interaction has long been theorized as an essential component to technological challenges related to user experiences and interaction (Dourish, 2004). While a single overarching vocabulary and grammar could not possibly suit the myriad formalizations that undergird discursive constructs, particularly when expressed in varying contested social contexts, the weight of the potential impossibility of a universalizing semantics has seldom deterred platform metadata modelers, who often view themselves as engineers tasked with the responsibility of enhancing data accessibility and exchange on the web through their platforms and their products, regardless of the long-term feasibility of such an infrastructural semantic project.

The interviews conducted with metadata modelers for the present project paint a slightly different picture of metadata modeling work and show evidence of the slight disillusionment of these workers with the larger semantic project. While metadata modelers historically have tended to downplay intractable semantic problems when describing projects like the "Semantic Web" in popular articles (Berners-Lee et al., 2001), preferring instead to focus on the imagined future benefits of the technology,

examples taken from these interviews describe several apparent failures. Two such examples from the interview transcripts are presented below:

> *"We became what you might call the semantic guardians…who tried to make sure that the semantics…aligned properly. And I have to say, we only partially succeeded…I think it is one of the failures for us. We weren't able to get them cleanly aligned in a way that people found natural. And so the result has been rather...it's rather complicated and awkward for people who come to the Semantic Web, cold, and plaintively say, 'how can I use this?' And you have to start saying, 'sit down and I'll take you through a graduate-level course in the foundations of logic before you could even get started.' We tried, we should have succeeded in making the whole thing a lot simpler and we didn't."*

> *"So apart from all these little technical issues of error and worries about the complexity of the language and so on…this is I think a central difficulty for the Semantic Web, which is that people, even well-meaning people, well trained, smart, intelligent people, best will in the world, don't just spontaneously agree on metaphysics. So they will tend to write down what they know in the formalism you give them, in different ways, and I can't see any way at all of policing this. There's no way of conveying to the world, 'you guys should all do it this way,' which a lot of the upper-level ontologies…want to do. They want to say to the world, 'hey, this is the way you do it, guys, this is the way you think about how time changes things and so forth and I've fixed it in this ontology, now, run with this,' but people won't."*

Beyond the examples from the interview transcripts which showed evidence of the frustration of these workers, the interviews also painted a picture of a metadata ecology consisting of various logics, syntaxes, serializations, enterprise software, upper level versus domain ontologies, and standards that add a cumbersome complexity to the field of metadata modeling work. What emerges is the image of a kludgy data field that informs the sleek, fact producing products that millions of people use when searching for answers and information.

**References**


Acker, A. (2015). Toward a hermeneutics of data. *IEEE Annals of the History of Computing*, *37*(3), 70-75.

Adam, A. (1998). *Artificial Knowing: Gender and the Thinking Machine*. Routledge.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34-43.

Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, *8*(1), 19-30.

Eriksson, M. (2016). Close reading big data: The Echo Nest and the production of (rotten) music metadata. *First Monday*. https://journals.uic.edu/ojs/index.php/fm/article/view/6303

Forsythe, D. E. (1993). Engineering knowledge: The construction of knowledge in artificial intelligence. *Social studies of science*, 23(3), 445-477.

Hjørland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, field, content... and relevance. *Journal of the American Society for Information Science and Technology*, *52*(9), 774-778.

Strauss, A., & Corbin, J. M. (1997). *Grounded Theory in Practice*. Sage.

## TRACING "TRAVELING FACTS" ON THE SEMANTIC WEB

Heather Ford
University of Technology Sydney

In a 2012 blog post titled "Things, Not Strings," the senior vice president of engineering for Google, Amit Singhal, wrote that Google was now using Wikipedia and other "public" data sources to seed a knowledge graph that would provide "smarter search results" for users (Singhal, 2012). Other named sources included the CIA World Factbook and Freebase (a now retired online database, formerly owned by Google). In addition to returning a list of possible results, Google would present a "knowledge panel" on the right-hand side of the page that "summarize[s] relevant content around that topic, including key facts you're likely to need for that particular thing" (Singhal, 2012).

Google's Knowledge Graph proved a unique implementation of the semantic web, a vision of a "web of data" to replace the "web of documents" that had characterised the web's original design (Berners-Lee, Hendler and Lassila, 2001). The semantic web is based on the ideal that computers need to be able to process the semantics (meaning) that humans attach to their words. Using semantic web logic, the computer would "know whether we meant Paris, the perfume, Paris, the place or Paris, the celebrity" by structuring information in a way that computers could understand, and then interconnecting those structured databases so that computers could understand what was meant by users' queries (Berners-Lee et al., 2001). To the web's founder, Tim Berner's Lee, promoting computer "understanding" required the development of what he called "linked data". In a 2009 TED talk, Berners-Lee said that the web was about sharing documents but that there was "still huge unlocked potential" that could be realized by sharing data rather than documents. Data was about relationships, explained Berners-Lee, and "the really important thing about data is the more things you have to connect together, the more powerful it is" (Berners-Lee, 2009). The idea of linked data was about "people doing their bit to produce a little bit, and it all connecting."

Two years later and Google appeared to provide the semantic web with a major boost when it launched the knowledge graph. But rather than the original web being built by enthusiasts all "doing their bit" as Berners-Lee had presented it (Berners-Lee, 2009), Google was using machine learning to extract billions of entities from around the web and doing this to reinforce its power. Within seven months of its launch, Google's knowledge graph had tripled in size to cover 570 entities and 18 billion facts imported from sources like Wikipedia.

Google used AI and machine learning in particular to recognize statements from a variety of sources as belonging to the same person, place, event, or thing. That data was then used to power applications that answered queries by human users speaking to the machine in what linguists and computer scientists call “natural language”—in order to distinguish it from other language types, such as the computing programming languages that provide instructions to machines. The knowledge graph enabled Google to answer questions like “Who is Desmond Tutu?” “What happened in Ukraine today?” and “What is the capital of Australia?” directly. In addition to powering knowledge panels in Google search, the knowledge graph was also used to answer spoken voice queries in Google Assistant on android phones and Google’s smart speakers.

Now knowledge graphs power the delivery of answers from Q&A systems on major virtual assistants and smart speakers including Apple’s Siri and Amazon’s Alexa. It is predicted that there will soon be more smart speakers than people on the planet (Shulevitz, 2018). Smart speakers have become increasingly popular and their Q&A applications are one of their most popular features. Knowledge graphs reached the peak of Gartner’s 2020 Hype Cycle for Artificial Intelligence ().

In this paper, I argue that the rise of the semantic web as it has been developed and driven by platform companies via automatically generated knowledge graphs and simplistic answers in Q&A systems has resulted in a series of significant consequences for the web that remain largely unexamined. Data scientists Connor McMahon, Isaac Johnson, and Brent Hecht (2017) found that facts in Google’s knowledge panels were predominantly sourced from Wikipedia even though they were “almost never cited” beyond the opening description of the phenomenon. According to the authors, this has had a significant impact on Wikipedia’s sustainability as visitors do not need to go to Wikipedia to obtain answers, and thus miss the opportunity to donate their time or money to the non-profit organisation hosting it.

More than that, I argue that the adoption (and domination) by platform companies of linked data has catalysed a re-shaping of web content to accord with the question and answer and linked data formats, weakening the power of open content licences to support local knowledge and consolidating the power of algorithmic knowledge systems that favour knowledge monopolies. This has been supported by the ideologies of open data, big data and AI and the socio-technical influences of automation and machine learning that dominate the web ecosystem today.

I do this using methods and theories from Science and Technology Studies relating to the development of scientific knowledge or “science-in-the-making”, but applied in this case to the construction and travel of facts relating to historic events as they occur and are represented in key sites across the web. In particular, I apply Morgan’s (2010) conceptual framework relating to “how facts travel” in order to understand how power and agency are distributed in the construction and distribution of facts that become so privileged in current knowledge systems. I start by locating semantic data represented as facts and categorisations of historic events using the case of the 2011 Egyptian revolution and then the 2022 invasion of Ukraine in Google Search (including the knowledge panels, “people also asked” and featured snippets) and Google Assistant

(answers to questions). I analyse the discourse and materiality of these semantic representations using grounded theory analyses of screenshots and documents including Google's blog, Google research and other corporate materials. I then trace these facts back to their common origins in Wikipedia, Wikidata and Quora and analyse the discourse and practice surrounding their construction, debate and wrangling. This work is supported by a long term ethnographic study of Wikipedia's data projects.

The contributions of this paper are twofold. First, I demonstrate the method of tracing "facts-in-the-making" as a useful mechanism for understanding how truth in the age of the semantic web is constructed. Second, I provide a political-economic analysis of the changes wrought by the semantic web and identify the features of linked data development that prevent the development of shared knowledge the semantic web was founded to solve.

**References**

Morgan, M., (2010) "Travelling facts," in *How well do facts travel?: The dissemination of reliable knowledge*, Howlett, P., & Morgan, M. S. (Eds.).Cambridge University Press.

Singhal, A., "Introducing the Knowledge Graph: Things, Not Strings," The Keyword (blog), May 16, 2012, https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.

McMahon, C, Johnson, I., and Hecht, B., (2017), "The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship between Peer Production Communities and Information Technologies," in *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017* (AAAI Press), 142–151.

Shulevitz, J. (2018). Alexa, Should We Trust You? *The Atlantic*.

## BUILDING A SEMANTIC FOUNDATION FOR MACHINE LEARNING - OR HOW MEANING AND COLONIAL LABOR RELATIONS ARE MADE INFRASTRUCTURAL

Doris Allhutter
Institute of Technology Assessment, Austrian Academy of Sciences

My talk traces practices of building a digital infrastructure that enables machines to learn from human language and everyday discourse. This emerging 'semantic infrastructure' is a 'meaning-centered' foundation relating concepts or entities, such as data from text, speech, and images. It already is a result of machine learning techniques and again feeds resources into further (semi-)automated processes. There is a multilayered apparatus in place for a semantic infrastructure to emerge: this apparatus consists of existing web architectures, the tools, methods and practices used, the data, standards and protocols. Nonetheless, this apparatus includes narratives of an economically beneficial and smart future growing from epistemic traditions in computer science and AI research, as well as economic practices that embed and coemerge with

these technologies. All this is held together by affective investments into a spirit of collectivity and peer production. These intertwined structures, materialities, practices, and affects configure a sociomaterial apparatus that is historically contingent. Using examples from an ethnographic study (Allhutter 2019), I show how the apparatus of semantic computing and its infrastructuring practices are materially and discursively performative in their co-emergence with techno-epistemic discourses and politico-economic structures.

**Algorithmic Performances of the 'Everyday'**

Learning systems from training data that represent commonsense knowledge aims at endowing machines with human-like understanding. Everyday language and commonsense knowledge are particularly valuable resources for training intelligent systems, and, I suggest, they are particularly interesting for two reasons: Firstly, from a technical perspective, commonsense knowledge is a resource that can be used to mutually train disparate systems. For example, a system extracting everyday knowledge from natural language can learn that a 'horse' is an 'animal' and that a horse is commonly found in a 'stable on a farm'. Image recognition can then, e.g. access an image of a horse and learn to recognize what the building in the background is. In networked infrastructures, systems thus communicate with each other in an effort to understand and create 'meaning'. Secondly, using everyday language and common sense is fascinating due to their inherent power relevant dimensions: commonsense knowledge may not be subject to or backed by scientific findings or expert knowledge, but it represents widely uncontested knowledge that has become hegemonic in a particular geohistorical context. It implicitly informs our everyday practices and ad hoc decisions. In her political theory of the everyday, Brigitte Bargetz (2016) describes everyday practices as a crucial site of political contestation: they are how power structures are enacted (2016, 208). The everyday is "a mode of exercising power" (2016, 35) that, at the same time, carries a potential for agency and political resistance. Representing common sense in terms of axiomatic relationships between the objects and concepts in a domain and designing (semi)automated methods to reason about these relations reflects a particular way of conceptualizing the world. It is part of a sociomaterial process in which classifications, standards, (meta)data, and methods coemerge with processes of signification that reconstitute and/or shift hegemonic ecologies of knowledge.

The political nature of practices of semantic infrastructuring becomes visible by analyzing how infrastructures or particular enactments of an infrastructure coemerge with societal structures, individuals, bodies, and their knowledge practices. In Karen Barad's (2003) view, some things, views, practices matter—they are made possible—and others are excluded and do not materialize. The performativity of infrastructure and its agentive capacities show in the way in which it accommodates some practices, people, and viewpoints more than others (Star and Ruhleder 1996). This becomes particularly pertinent in light of research that shows how machine learning amplifies structural discrimination and reinforces injustice, sexism, racism, classism, and ableism (Hu Kohler-Hausmann 2020, Benthall & Haynes 2019). I suggest that acknowledging that semantic infrastructuring (or making meaning infrastructural) affects and is affected

by power relations allows us to analytically grasp the hegemonic ways in which difference, and in particular ideologies of human difference, become performative.

## The Coloniality of 'Human Computation'

With its roots in AI and machine learning, the epistemic narratives of automation and machine intelligence have provided a foundation for the field of semantic computing. However, 'human computation' and the global digital economy have become an integral part of its methodologies. Semantic elements of the meaning-centered infrastructure (e.g. commonsense ontologies) are fabricated in hybrid, semi-automated processes relying on algorithmic agencies and human labor. A crucial question is how high the cost actually is for encoding all the relevant knowledge so that it can be exploited by machines. Thus, outsourcing and the crowd workers of the Global South (Scholz 2016) play a substantial role in building the semantic foundations that are supposed to generate economic prosperity for the Global North. Due to its embeddedness in digital marketplaces and tools (such as ontology editors), low pay is inherent to microwork. With its vulnerable working conditions, unregulated work times, and a high dependency on opaque technical evaluation systems, crowdwork epitomizes the persistence of the "coloniality of labor" and engenders "a place of 'exteriority' or 'colonial difference'" (Gutiérrez-Rodríguez 2010, 44). The creation of the semantic infrastructure is contingent on a transnational division of labor. It emerges in relation to colonial pasts and presents and enacts, continues, and transforms global economic processes. The digitization of knowledge work and the implementation of crowdwork into the apparatus of semantic computing are part of the post-Fordist transformation of a transnational division of labor.

## Infrastructural Power

Thinking together how meaning and colonial labor relations are made infrastructural in current practices of computing, shows some of the socio-material choices made when establishing a semantic foundation for machine learning. It is crucial to see how they create difference and inequality through a set of epistemic and economic practices. To take account of the entanglement of historically grown, structural power relations and their emergent materializations in sociotechnical systems, I use the notion of 'infrastructural power'. It attempts a theorization of 'intra-acting' modes of 'power' that relates structural and transitional elements in an attempt to capture the enfolded micro, meso and macro-levels of socio-material practices in their historical contingency.

## References

Allhutter, D. (2019). Of 'Working Ontologists' and 'High-quality Human Components'. The Politics of Semantic Infrastructures. In *DigitalSTS: A Field Guide for Science & Technology Studies*, Princeton and Oxford: Princeton University Press, 326-348.

Barad, K. (2003). *Posthumanist Performativity. Toward an Understanding of How Matter Comes to Matter.* Signs 28 (3): 801–31.

Bargetz, B. (2016). *Ambivalenzen des Alltags. Neuorientierungen für eine Theorie des Politischen*, Bielefeld: transcript.

Benthall, S., & Haynes, B. D. (2019, January). *Racial categories in machine learning*. In Proceedings of the conference on fairness, accountability, and transparency (pp. 289-298).

Gutiérrez-Rodríguez, E. (2010). *Migration, Domestic Work and Affect. A Deconolonial Approach on Value and the Feminization of Labor*. New York: Routledge.

Hu, L., & Kohler-Hausmann, I. (2020). *What's sex got to do with fair machine learning?.* arXiv preprint arXiv:2006.01770.

Scholz, T. (ed.) (2016). *Platform Cooperativism. Challenging the Corporate Sharing Economy*. New York: Rosa Luxemburg Foundation.

Star, S.L., & Ruhleder, K. (1996). *Steps toward an Ecology of Infrastructure: Design and Access for Large Information Spaces.* Information Systems Research 7 (1): 111–34.

## Extraction and Alienation of the Knowledge Commons: Wikipedia, Wikidata, and the Ethics of “Free” Data

Zachary McDowell
University of Illinois Chicago

Matthew Vetter
Indiana University of Pennsylvania

Yochai Benkler recently noted that “Wikipedia and commons-based peer production more generally continue to offer an existence proof that there can be another way” – an alternative to the more prevailing forces of market exchange and surveillance capitalism that characterize the current web (p. 43). Benkler’s take encapsulates the ethos of the commons that inspires the labor that creates Wikipedia (as well as spaces like FLOSS/FOSS ecosystems). However, recent developments in the Wikimedia ecosystem complicate (and even threaten) the ethos, and even the sustainability, of the digital commons. Wikidata, a sister project of Wikipedia, serves as a radically open, structured knowledge database that appears at first glance to continue the openness of the commons, but radically shifts the way in which the data is used. An enormous database with nearly 100 million data items that can (and are) used freely in machine learning networks to create new products, Wikidata is the largest semantic web platform humanity has ever created. If Wikipedia was the promise of Web 2.0, Wikidata represents Web 3.0, for better or worse.

Launched in 2012, Wikidata is a collaboratively edited knowledge base that is the structured data repository for all Wikimedia projects, particularly Wikipedia. Wikidata employs user-generated metadata standards which reflect consensus in the scientific community as well as already-established standards. As the backbone for Wikipedia, it

is the repository of information that is utilized by a variety of projects that represent information on a variety of platforms (e.g.: Facebook Factchecking, Google search results). The data model for entities in Wikidata utilize specific subfields' data structures so that data can be passed back and forth between databases. Thus, Wikidata uses APIs to ``talk to" entities in other databases developed and maintained by NIH, NCBI, Ensembl, Homologene, among others, helping to reinforce core concepts around scientific structured data.

What Wikidata shifts is the way in which the commons are both preserved and utilized, creating concerns for an "extraction" of the commons, rather than Benkler's "another way." At first glance Wikidata sounds like a huge step forward in providing truly free data. However, the ways in which large tech companies (in particular those training machine learning systems) utilize this data, and where the data originated highlight ethical concerns over data utilization and production. In this paper, we examine the structures of Wikidata, calling for more critical scrutiny of the project in terms of 1) its usage of CC0 "No Rights Reserved" license, and 2) how knowledge from Wikidata is extracted, re-appropriated, and commodified beyond the intent of its original creators.

That which is common has always been extracted – this is not new. Whether through folk art or other artistic appropriation, or through mining, land ownership, oil extraction, pollution, and water extraction, people and companies have extracted and profited from what is common since the dawn of time. Benkler's statement that "there can be another way" centers around the type of openness that creates more openness. A CC-BY-SA license is open in a way that is also "closed" – it shuts off information for particular types of usages in order to preserve the openness. This "other way" ensures that "what is common" creates more "common" and creates a new type of "oikos" (home) in the "oikonomos" (economy), one that focuses on sharing and giving away for the common good, and walling off that which is common from those who wish to extract it.

Instead of Wikipedia's CC-BY-SA ("share alike") license which requires that derivatives and uses of the information retain the same license, Wikidata utilizes a license that has no requirements. This might sound ideal, but in reality Wikidata appropriates that particular FOSS imaginary of sharing, but instead delicenses data by assigning it a CC0 license allowing companies to extract, commodify, and otherwise use this data in ways to create systems without requirements to utilize the license or reference the works which were utilized. Recognized early on in Wikidata's inception, Andreas Kolbe writes for the Wikipedia Signpost in 2015: "The no-attribution CC0 license means that third parties can use the data on their sites without indicating their provenance, obscuring the fact that the data came from a crowdsourced project subject to the customary disclaimers" (n.p.). Perhaps even more significant, the CC0 license allows for a commodification and re-appropriation of content originally licensed under a CC-BY-SA license and created by a volunteer community in Wikipedia. In particular, Wikidata scrapes Wikipedia metadata to populate its system with information, skirting the copyright issue. Although metadata isn't copyrightable currently (and there are ample reasons why it shouldn't be), the original information which has been utilized was created by countless volunteer hours under the guise of this "share alike" license.

In light of Wikimedia foundation's ties to tech giant donors such as Google, and how Google actively utilizes Wikimedia project data to train AIs (see "Announcing WIT"), this paper raises concerns over the ethics of utilizing the commons to create new "products" that are then sold back to those that toiled tirelessly to create what now feeds tech giant's machines. Marx's concept of alienation here is appropriate, as the outcome of this socially donated labor are transformed ("objectified," via metadata and into Wikidata) but also "dispossessed" (literally depriving the information from its ownership under CC-BY-SA) and then transformed into capital (by companies using the data to make new products) (Marx 1993, p. 832). As Sartre notes, alienation "begins with exploitation" (Sartre 2004, p. 227), and the methods of extraction and usage of labor donated under the guise of sharing seems extremely exploitative. This creates what we refer to as "re-alienation of the commons," in which the fundamental agreement to donate labor under the guise of the commons and sharing, to what Benkler refers to as "another way," has become broken, utilized, transformed, monetized, and sold back to the society that worked so hard to create something shared. This "other way" attempted to subvert this massive alienating system by creating a community of sharing, but instead it became a new space for extraction.

**References**


Announcing WIT: A Wikipedia-Based Image-Text Dataset. (n.d.). Google AI Blog. Retrieved February 18, 2022, from http://ai.googleblog.com/2021/09/announcing-wit-wikipedia-based-image.html

Benkler, Y. From utopia to practice and back. (2020). In J. Reagle and J. Koerner (Eds.), *Wikipedia @ 20: Stories of an incomplete revolution* (pp. 43-54). MIT Press. https://doi.org/10.7551/mitpress/12366.001.0001

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

Marx, K. (1993 [1939]). *Grundrisse: Foundations of the critique of political economy* (rough draft). Penguin Classics.

Sartre, J.-P. (1984). *Critique of Dialectical Reason, Vol. 1: Theory of Practical Ensembles* (First edition). Verso.

Øversveen, E. (2021). Capitalism and alienation: Towards a Marxist theory of alienation for the 21st century. *European Journal of Social Theory*, 13684310211021580. https://doi.org/10.1177/13684310211021579