Selected Papers of #AoIR2020:
The 21st Annual Conference of the
Association of Internet Researchers
Virtual Event / 27-31 October 2020

PLATFORMIZING KNOWLEDGE: MESS AND MEANING IN WEB 3.0 INFRASTRUCTURES

Andrew Iliadis
Temple University

Wesley Stevens
Temple University

Jean-Christophe Plantin
London School of Economics

Amelia Acker
University of Texas at Austin

Huw Davies
University of Edinburgh

Rebecca Eynon
University of Oxford

Panel Introduction

This panel focuses on the way that platforms have become key players in the representation of knowledge. Where the retronym “Web 1.0” is generally used to refer to the web before the advent of social media, “Web 2.0” is typically used to represent a shift toward digital convergence and user interaction—though scholars have shown that the division is ultimately contingent on one’s historical perspective (Ankerson, 2015; Blank & Reisdorf, 2012; boyd, 2015; Cammaerts, 2008; Fisher, 2018; Gehl, 2011; Song, 2010; van Dijck, 2013; Zimmer, 2009).

Media and communication researchers have further theorized the Web 1.0-2.0 transition by emphasizing the modularity of social media platforms like YouTube, Facebook, and Twitter (Langlois et al, 2009; Gillespie, 2010; Langlois & Elmer, 2013; McKelvey, 2011). Following research into the cultural significance of software (Chun, 2011; Fuller, 2008; Galloway, 2004; Kitchin & Dodge, 2011; Manovich, 1999, 2001), Web 2.0 researcher’s

Iliadis, A., Stevens, W., Plantin, J-C., Acker, A., Davies, H., Eynon. R. (2020, October). *Platformizing knowledge: mess and meaning in web 3.0 infrastructures*. Paper presented at AoIR 2020: The 21th Annual Conference of the Association of Internet Researchers. Virtual Event: AoIR. Retrieved from http://spir.aoir.org.

emphasis on platform programmability (Bogost & Montfort, 2009) and the materiality of informational media (Dourish, 2017; Dourish & Mazmanian, 2013) as a definitional aspect of the web highlighted a historical shift from unidirectional and largely static web content to a more dynamic and multidirectional one with increased affordances (Bucher & Helmond, 2017).

Specifically, inquiries into the social significance of digital tools like application programming interfaces (APIs) have shown that such technologies facilitate platformization (Berry et al, 2015; Bucher, 2013), which “entails the extension of social media platforms into the rest of the web and their drive to make external web data ‘platform ready’” (Helmond, 2015). Building on work in the tradition of ethnography of infrastructure (Star, 1999), research into digital information infrastructures has developed adjacent to these platform-oriented approaches (Bowker et al, 2010; Edwards et al, 2009), with a specific subset focusing on online knowledge infrastructures (Edwards et al, 2013; Karasti et al, 2016) and their interoperability.

Recently, there have been calls to combine infrastructure and platform-based frameworks to understand the nature of information exchange on the web through digital tools for knowledge sharing (Plantin, 2018; Plantin et al, 2018a; Plantin et al, 2018b). The present panel builds and extends work on platform and infrastructure studies in what has been referred to as “knowledge as programmable object” (Plantin, et al., 2018b), specifically focusing on how metadata and semantic information are shaped and exchanged in specific web contexts (Bates et al, 2016; Edwards et al, 2011; Hui, 2016; Leonelli, 2016).

As Bucher (2012; 2013) and Helmond (2015) show, data portability in the context of web platforms requires a certain level of semantic annotation. Semantic interoperability is the defining feature of so-called “Web 3.0”—traditionally referred to by computer scientists as the semantic web (Antoniou et al, 2012; Szeredi et al, 2014). Since its inception, the semantic web has privileged the status of metadata for providing the fine-grained levels of contextual expressivity needed for machine-readable web data, and can be found in products as diverse as Google's Knowledge Graph, virtual assistants like Siri and Alexa that rely on Wikidata, online research repositories like Figshare, and other sources that engage in platformizing knowledge.

The first paper in this panel examines the international Schema.org collaboration. The second paper investigates the epistemological implications when platforms organize data sharing. The third paper discusses private platforms’ extraction and collection of user metadata and the enclosure of data access. The fourth paper argues for the use of patents to inform research methodologies for understanding knowledge graphs.

References

Ankerson, M. S. (2015). Social Media and the “Read-Only” Web: Reconfiguring Social Logics and Historical Boundaries. *Social Media and Society*, 1(2). https://doi.org/10.1177/2056305115621935

Antoniou, G., Groth, P., Harmelen, F. van, & Hoekstra, R. (2012). *A Semantic Web Primer*. Cambridge: MIT Press.

Bates, J., Lin, Y.-W., & Goodale, P. (2016). Data journeys: Capturing the Socio-material constitution of data objects and flows. *Big Data & Society*, 3(2), 205395171665450. https://doi.org/10.1177/2053951716654502

Berry, D. M., Borra, E., Helmond, A., Plantin, J.-C., Walker Rettberg, J., & Walker, J. (2015). The Data Sprint Approach: Exploring the field of Digital Humanities through Amazon's Application Programming Interface. *Digital Humanities Quarterly*, 9(4). Retrieved from http://eprints.lse.ac.uk/65438/

Blank, G., & Reisdorf, B. C. (2012). The Participatory Web: A user perspective on Web 2.0. *I Information, Communication & Society*, 15(4), 537–554. https://doi.org/10.1080/1369118X.2012.665935

Bogost, I., & Montfort, N. (2009). *Racing the Beam: The Atari Video Computer System*. Cambridge: MIT Press.

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* (pp. 97–117).

boyd, d. (2015). Social Media: A Phenomenon to be Analyzed. *Social Media and Society*, 1(1). https://doi.org/10.1177/2056305115580148

Bucher, T. (2012). Want to be on the Top? Algorithmic Power and the Threat of Invisibility on Facebook. *New Media and Society*, 14(7), 1164–1180. https://doi.org/10.1177/1461444812440159

Bucher, T. (2013). Objects of Intense Feeling: The Case of the Twitter API. *Computational Culture* (a Journal of Software Studies), 1–17. Retrieved from http://computationalculture.net/article/objects-of-intense-feeling-the-case-of-the-twitter-api

Bucher, T., & Helmond, A. (2017). The Affordances of Social Media Platforms. *SAGE Handbook of Social Media*, (June 2016), 1–41.

Cammaerts, B. (2008). Critiques on the Participatory Potentials of Web 2.0. *Communication, Culture & Critique*, 1(4), 358–377. https://doi.org/10.1111/j.1753-9137.2008.00028.x

Chun, W. H. K. (2011). *Programmed Visions: Software and Memory*. Cambridge: MIT Press.

Dourish, P. (2017). *The Stuff of Bits: An Essay on the Materialities of Information*. Cambridge: MIT Press.

Dourish, P., & Mazmanian, M. (2013). Media as Material: Information Representations as Material Foundations for Organizational Practice. In *How Matter Matters: Objects, Artifacts, and Materiality in Organization Studies* (p. 92-118).

Edwards, P. N., Bowker, G. C., Jackson, S. J., & Williams, R. (2009). Introduction: An Agenda for Infrastructure Studies. *Journal of the Association for Information Systems*, 10(5), 364–374. DOI:10.17705/1jais.00200

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., … Calvert, S. (2013). *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*, (May), 25–28. https://doi.org/http://hdl.handle.net/2027.42/97552

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science*, 41(5), 667–690.

Fisher, E. (2018). When Information wanted to be Free: Discursive Bifurcation of Information and the Origins of Web 2.0. *Information Society*, 34(1), 40–48. https://doi.org/10.1080/01972243.2017.1391910

Fuller, M. (2008). Software Studies: A Lexicon. *The MIT Press*, 334. https://doi.org/10.7551/mitpress/9780262062749.001.0001

Galloway, A. R. (2004). *Protocol: How Control Exists After Decentralization*. Cambridge: MIT Press.

Gehl, R. W. (2011). The Archive and the Processor: The Internal Logic of Web 2.0. *New Media and Society*, 13(8), 1228–1244. https://doi.org/10.1177/1461444811401735

Gillespie, T. (2010). The Politics of “Platforms.” *New Media and Society*, 12(3), 347–364. doi:10.1177/1461444809342738

Helmond, A. (2015). The Platformization of the Web: Making Web Data Platform Ready. *Social Media and Society*, 1(2). doi:10.1177/2056305115603080

Hui, Y. (2016). *On the Existence of Digital Objects*. Minnesota: University of Minnesota Press.

Karasti, H., Millerand, F., Hine, C. M., & Bowker, G. (2016). Knowledge Infrastructures: Part IV. *Science and Technology Studies*, 29(4), 1–9.

Kitchin, R., & Dodge, M. (2011). *Code/Space: Software and Everyday Life*. Cambridge: MIT Press. https://doi.org/10.1080/00343404.2012.696477

Langlois, G., & Elmer, G. (2013). The Research Politics of Social Media Platforms. *Culture Machine*, 14, 1–17.

Leonelli, S. (2016). *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.

Manovich, L. (1999). Database as Symbolic Form. *Convergence*, 5(2), 80–99. https://doi.org/10.1177/135485659900500206

Manovich, L. (2001). *The Language of New Media*. Cambridge: MIT Press. https://doi.org/10.1386/nl.5.1.25/1

McKelvey, F. (2011). FCJ-128 A Programmable Platform? Drupal, Modularity, and the Future of the Web. *Fiberculture*, (18), 232–254.

Plantin, J.-C. (2018). Google maps as cartographic infrastructure: From participatory mapmaking to database maintenance. *International Journal of Communication*, 12, 489–506.

Plantin, J.-C., Lagoze, C., & Edwards, P. N. (2018a). Re-integrating Scholarly Infrastructure: The Ambiguous Role of Data Sharing Platforms. *Big Data & Society*, 5(1), 205395171875668. https://doi.org/10.1177/2053951718756683

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

Song, F. W. (2010). Theorizing Web 2.0: A Cultural Perspective. *Information, Communication & Society*, 13(2), 249–275. https://doi.org/10.1080/13691180902914610

Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, 43(3), 377–391. https://doi.org/10.1177/00027649921955326

Szeredi, P., Lukácsy, G., Benkő, T. (2014). *The Semantic Web Explained: The Technology and Mathematics behind Web 3.0*. Cambridge: Cambridge University Press.

van Dijck, J. (2013). Facebook and the Engineering of Connectivity: A Multi-layered approach to Social Media Platforms. *Convergence*, 19(2), 141–155. https://doi.org/10.1177/1354856512457548

Zimmer, M. (2009). Renvois of the Past, Present and Future: Hyperlinks and the Structuring of Knowledge from the Encyclopédie to Web 2.0. *New Media and Society*, 11(1–2), 95–113. https://doi.org/10.1177/1461444808099573

ONE SCHEMA TO RULE THEM ALL: ANALYZING GOOGLE, MICROSOFT, YAHOO, AND YANDEX'S SCHEMA.ORG STRUCTURED METADATA PROJECT

Andrew Iliadis
Temple University

Wesley Stevens
Temple University

Increasing data interoperability on platforms through structured metadata modeling is a key feature of web 3.0. Historically, several domain-specific metadata initiatives have facilitated data interoperability for individual industries. These structured metadata schemas typically allow developers to “wrap” web data to provide semantic and contextual expressivity of content, thus improving information search and retrieval in their respective fields. Some drawbacks of this practice are that a plurality of metadata schemas produces layered code that can overburden administrators, while lack of a single metadata vocabulary results in unequal distribution of schemas due to intermittent application. In this paper, we use network analysis and archival research to provide a historical timeline and an analysis of a global project to construct a universal standard schema on the web (Schema.org). Throughout, we discuss the political economy of semantics and raise some potential ethical concerns.

The Semantic Web

The Semantic Web is a loosely connected group of technologies, standards, methods, organizations, and people responsible for improving data interoperability across the web. Today, many products that people use benefit from semantic web outputs, including knowledge panels, virtual assistants, databases, and dashboards.

Since its inception, the semantic web has privileged the status of metadata for providing the fine-grained levels of contextual expressivity needed for machine-readable web data. Sir Tim Berners-Lee, the inventor of the web, expressed that “the first form of semantic data on the Web was metadata” (Berners-Lee & Fischetti, 1999). After inventing the web in 1989, Berners-Lee founded the World Wide Web Consortium (W3C) in 1994, one of the web’s main international standards organizations. W3C describe the semantic web as a ‘web of data’ and have led several semantic web metadata initiatives, publishing standards like Resource Description Framework (RDF) data model and the languages that use RDF to describe semantic information on the web, including the widely used Web Ontology Language (OWL) and the more recent Shapes Constraint Language (SHACL). In 2006, Berners-Lee issued four rules for a proper semantic web that could be checked against a five-star linked open data system. These rules included using Uniform Resource Identifiers (URIs) as names for things, using Hypertext Transfer Protocol (HTTP) URIs so individuals can look up names, providing metadata using RDF, and including links to enable further discovery. A five-star, linked open data rating from the W3C would require that the web data be open licensed, available as machine-readable structured data, exist in a non-proprietary format, use the technical standards of the W3C for identifying data, and, most importantly, link the data to other data on the web, thus creating context.

Since these early beginnings, W3C’s standards have effectively bootstrapped the semantic web into what it is today, undergirding web data in several domains. Though the project has yet to fully live up to the early vision of ubiquitous adoption promoted by Berners-Lee, there is clear evidence that standards created by W3C have impacted areas where semantic web data must be exchanged by machines. Yet, as Mccarthy (2017) shows, the project might more accurately be described as a series of “entangled” semantic webs, due in part to the diverse, specialized domains where idiosyncratic vocabularies are created (economics, publishing, etc.), along with tensions between proprietary and public metadata schemas. This situation is starting to change with the introduction of globally coordinated projects to align semantics across fields.

Schema.org and Key Structured Data Events

A modern outgrowth of the semantic web, Schema.org was launched on June 2, 2011 as a joint project between Google, Microsoft, Yahoo, and Yandex to standardize structured data on the web. Since its beginnings, Schema.org has acted as a self-described “collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond,” according to their website. A list of all the core changes to the vocabulary in each of the Schema.org releases is available (visit schema.org/docs/releases.html), from versions 0.91 (released on 21/4/12) to 6.0 (released on 21/1/20). While the Schema.org project is global in nature and the result of individual contributions from people at different organizations (some commercial competitors), Google is most closely associated with Schema.org’s daily operations, provide the most resources and personnel, and are one of the largest users of its outputs in their products such as Google’s Knowledge Graph and virtual assistants. Further, the Schema.org metadata model in many ways collects and build on the work of smaller, domain specific initiatives such as Dublin Core, Friends of a Friend, and GoodRelations.

To get a better sense of Schema.org, we spent time over a five-year period talking to semantic web experts, developers, and users to better understand the history of structured data on the web. Several of the individuals and groups we spoke to are part of search engine optimization (SEO) communities (like SEO Skeptic and Graph Lounge), or were involved (then or in the past) in building the underlying tools and infrastructure for the semantic web at places like the W3C. Before analyzing the structure of Schema.org itself, we wanted to get a better sense of the historical development of large platform companies’ engagement with structured data on the web, and to do this we pulled CSV data from SEO Skeptic, who had been tracking platform companies’ releases and announcements, to create our own data visualization of a timeline of key structured data events over a ten year period, from 2009-2019 (Figure 1). We labelled and color coded the data to identify which platform companies and organizations were involved in each of the events. These included Google, Microsoft, Yahoo, Yandex, Schema.org, Freebase, Pinterest, Apple, Wikipedia, Facebook, Twitter.

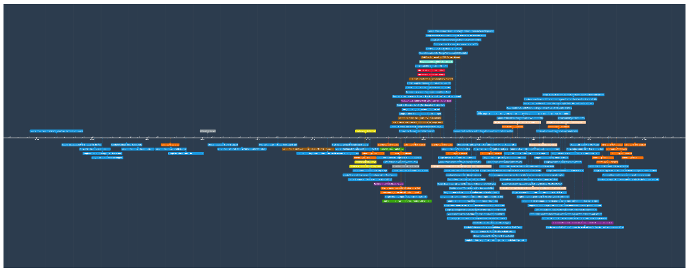
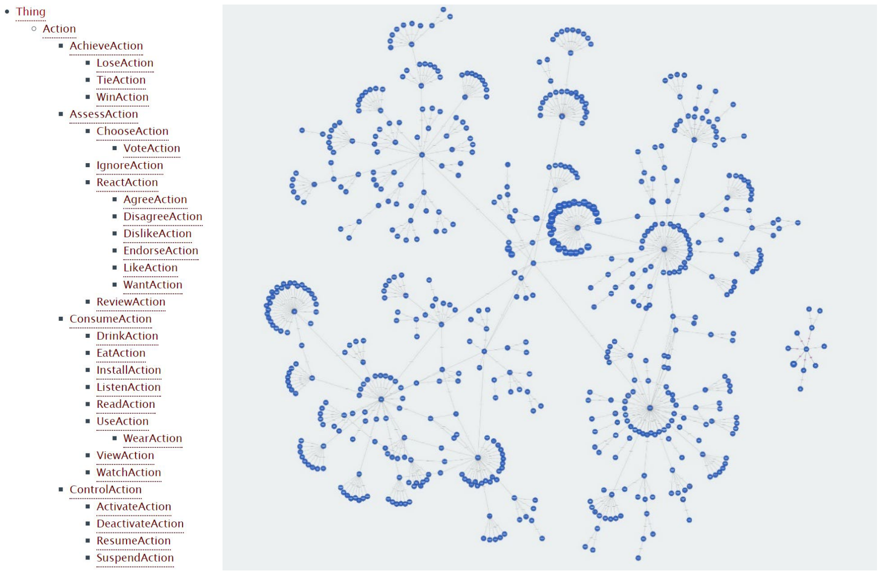


Figure 1 Timeline of Key Structured Data Events (2009-2019)

The timeline clearly showed that Google was the dominant force involved in over 90 key events involving announcement and effective dates of structured data products and services. Some of these included the introduction of “Rich Snippets” (12/5/2009), the launch of the “Knowledge Graph” (16/5/2012), and the beta version release of the “Fact Check Markup Tool” (2/10/2018). While the visualization lacks certain small product updates and releases from organizations like Wikipedia (it does contain most of the major announcements), the timeline overwhelmingly shows that Google has been involved in more significant product and feature announcements (apart from updates and releases) than other companies and organizations in the structured data field. We coded Schema.org as a separate entity since it is technically an international collaboration, and identified 20 key structured data events, including its version release date history, allowing us to see its progress alongside Google and the others. If one were to label Schema.org as being, for all intents and purposes, a Google initiative, then Google’s dominance of the structured data field increases by a third.

Schema.org and Semantic Networks

Where the timeline allowed us to see a bird’s eye view of Schema.org activity in the structured data field while also providing us an idea of the overall investment in semantics by some of the main platform companies, we also wanted to bring a magnifying glass to Schema.org’s embedded semantic structure. Our second task was to examine in detail the semantics of Schema.org’s structured data and organization. To do this, we downloaded the raw Schema.org data model hierarchy in machine readable form as an RDF file (available at schema.org/docs/developers.html). We then processed the RDF data model in WebVOWL, a tool for web-based visualization of ontologies. This allowed us to create a semantic network visualization of Schema.org’s internal organization, overall vocabulary and centrality of terms, which we compared to the hierarchy and available documentation (Figure 2).



- Thing
  - Action
    - AchieveAction
      - LoseAction
      - TieAction
      - WinAction
    - AssessAction
      - ChooseAction
        - VoteAction
      - IgnoreAction
      - ReactAction
        - AgreeAction
        - DisagreeAction
        - DislikeAction
        - EndorseAction
        - LikeAction
        - WantAction
      - ReviewAction
    - ConsumeAction
      - DrinkAction
      - EatAction
      - InstallAction
      - ListenAction
      - ReadAction
      - UseAction
        - WearAction
      - ViewAction
      - WatchAction
    - ControlAction
      - ActivateAction
      - DeactivateAction
      - ResumeAction
      - SuspendAction

Figure 2 Partial Schema.org Hierarchy (left) and Schema.org Semantic Network (right)

We then took turns iteratively and individually reviewing the hierarchy and semantic network while taking notes and then met to discuss our observations. Each of us noted several interesting features of Schema.org's ontology. First, we both noted that as a form of knowledge representation Schema.org is created, disputed, and revised as evidenced by the extensive available documentation of its release history. Second, we broadly noted that the project privileges ontological and epistemological commitments and understandings of things relating to work and everyday life and that these assumptions are subsequently baked into platforms by algorithms that pull structured metadata in search results (Schema.org is used by over 10 million websites, apps, and products, according to their website). By laying bare these conceptual maps, it became clear that Schema.org privileges certain modalities of things like labor, leisure, and knowledge over others (though, as we mentioned, the hierarchy is always revised).

The semantic network illustrates the datafication of life within several domains, from tech industries to creative work. For instance, the "CreativeWork" category in the network is the third largest node and includes entities related to media texts, production, and consumption. Here, creative works are understood as units or types of labor, illustrating tech companies' role in shaping normative standards of cultural production and consumption in recent decades (Craig & Cunningham, 2019; Napoli & Caplan, 2017). Similarly, the "Place" categorization privileges locations related to culture, tourism, and commerce (i.e. where people are likely to spend money and leisure time). These largely public locations are those often involved in vacation and leisure contexts (e.g. "TouristAttraction," "LandmarksOrHistoricalBuildings," "CivicStructure"). Likewise, the "Event" item is largely focused on cultural events and texts, specifically the arts (e.g. "ComedyEvent," "Festival," "ScreeningEvent," "ExhibitionEvent").

The "Intangible" category seems to hierarchize purchasing behaviors and commodities, a common theme among all the second-level entities that appear (beyond the top-level entity "Thing"). Moreover, some items seem far less tangible than others (e.g. the difference between making a reservation, purchasing a ticket, or enumeration). This category also includes sub-level items such as "Brand," "MerchantReturnPolicy," and "MediaSubscription." What is notable here is not necessarily how well these items fit within the assigned category, but the extent to which each category is defined by purchasing behaviors, cultural commodities, and creative labor. The privileging of creative work and focus on cultural labor made possible through the services provided by these tech companies illustrates the complex webs of meaning-making within these spheres as well as the role these corporations play in legitimizing and making those meanings accessible to the people who use them so frequently.

Several interesting entities appear, including "ClaimReview," which enables developers to markup news stories with fact checks so they appear as such in search results. The display of this feature was announced by Google on 13/10/2016. The use of this tool grew in connection to the "fake news" debates surrounding the 2016 US presidential election (Bing introduced support for fact checks using ClaimReview on 14/9/17). ClaimReview positions Schema.org and Google as a key arbiter of what can be considered credible, leveraging their already widespread legitimacy as both a harbinger and source of information. Moreover, Google's inclusion of knowledge panels as a way of keeping online traffic on their own search engine for quickly accessing information is illustrative of the company’s attempt to maintain their foothold and monopolize information as a kind of semantic “middle” layer between users and primary sources. For example, the Google knowledge panels that are partially populated by Schema.org markup and that contain information about things like biographical details and facts, are a major part of internet infrastructure for users today. “62 percent of mobile searches in June 2019 were no-click” and “people ages 13 to 21” are “twice as likely as respondents over 50 to consider their search complete” once they’ve seen a “knowledge panel” (Kelley, 2019). Semantics provided by sources like Schema.org will continue to play a key role as users ask Google and virtual assistants for things “on the fly.”

Conclusion

By engineering software that analyzes structured metadata and using it to provide ontological descriptions of the world, companies like Google have designed the technological infrastructure that privileges certain conceptualizations of things like work, labor, news, life, and technology, while limiting or ignoring alternatives (Iliadis, 2018, 2019). Through its ability to surface information in retrieval, Schema.org is in a way a regulator and gatekeeper, potentially limiting access to information that is not defined or included in the markup. Our job now is to locate more examples of this limitation and we plan to do so in future work. For example, we are interested in surfacing semantic errors in information retrieval on products like Google’s knowledge panels and Alexa’s answers which rely on Wikidata markup. These semantic infrastructures assert what is possible, normal, and appropriate while consolidating the datafication of information in ways that benefit companies and corporations. The Schema.org partnership represents a critical juncture in terms of how tech companies have grown to play an integral role regarding the way labor, information, cultural commodities, and media products are understood and accessed. In doing so, platform companies have not only created a hierarchy of what people want and how often they search for it, but underscore these processes with their own marketing and cultural logics (Jenkins, 2004; Klinger & Svensson, 2014; van Dijk & Poell, 2013).

References

Berners-Lee, T., and Fischetti, M. (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. New York: Harper.

Craig, D. and Cunningham, S. (2019). *Social Media Entertainment: The New Intersection of Hollywood and Silicon Valley*. NYU Press.

Iliadis, A. (2018). Algorithms, Ontology, and Social Progress. *Global Media and Communication*. 14(2), 219-230. 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. doi.org/10.1108/OIR-07-2018-0210

Jenkins, H. (2004). The Cultural Logic of Media Convergence. *International Journal of Cultural Studies*, 7(1), 33-43.

Kelley, L. (2019). The Google Feature Magnifying Disinformation. *The Atlantic*, September 23. Available at https://www.theatlantic.com/technology/archive/2019/09/googles-knowledge-panels-are-magnifying-disinformation/598474/

Klinger, U., & Svensson, J. (2015). The Emergence of Network Media Logic in Political Communication: A Theoretical Approach. *New Media & Society*, 17(8), 1241-1257.

Mccarthy, M. T. (2017). The Semantic Web and Its Entanglements. *Science, Technology and Society*, 22(1), 21–37. https://doi.org/10.1177/0971721816682796

Napoli, P., & Caplan, R. (2017). Why Media Companies Insist They’re Not Media Companies, Why they're Wrong, and Why it Matters. *First Monday*, 22(5).

van Dijck, J. and Poell, T., 2013. Understanding Social Media Logic. *Media and Communication*,1(1), 2-14.

THE EPISTEMOLOGICAL POWER OF DATA SHARING PLATFORMS

Jean-Christophe Plantin
London School of Economics

Critical internet researchers have shown the larger politics involved in using social media data for research, including the risks of neglecting the corporate logic that shapes the “social” (Langlois & Elmer, 2013; Marres, 2017) or resulting in a divide between “data rich” and “data poor” institutions (boyd & Crawford, 2012). Existing research has also detailed how conditions of access and use of such data conflict in many ways with the requirements for reliable scholarship (Acker & Kreisberg, 2019; Borra & Rieder, 2014; Driscoll & Walker, 2014; Gaffney & Puschmann, 2013; Gerlitz & Rieder, 2013; Hogan, 2018). This paper complements this research, but goes beyond analyzing the politics or the methodological issues emerging when using social media platforms as data source. It focuses instead on platforms dedicated to data sharing—defined as the deposition, preservation, and access for use and reuse of research data (Tenopir et al., 2011)—and investigates the epistemological implications when such platforms organize data dissemination and reuse.

The central case study is the platform Figshare, created in September 2011 by Dr Mark Hahnel, a PhD graduate from Imperial College London in stem cell biology. On one hand, it exists as Figshare.com, a website that invites individual researchers to self-archive their outputs (such as datasets, graphics, presentation slides, and almost anything else) on a personal profile. On the other hand, it exists as a middleware technology sold as Figshare for institutions and Figshare for publishers. With these offers, institutions such as universities, research institutions, libraries, data archives, repositories, and scientific publishers can contract with the company to get a series of services on top of their existing data infrastructures—such as customized web portals, management software for researchers’ data, or services that mint persistent identifiers (e.g., DOIs). Acting as a platform that mediates between a plurality of scholarly actors, but also that is “plugged-in” on top of other research institutions, Figshare illustrates how platforms are an increasingly important configuration for data sharing (Plantin et al, 2018).

Based on this case study, this paper investigates the epistemological implications of the reliance on such platforms in scholarship, by asking: When the logic of platforms is transposed from the web economy to the scholarly world, how does this configuration shape research? Do platforms facilitate all types of research practices, or are they more compatible with some scientific paradigms than with others? Is there an epistemological program embedded in the architecture of platforms?

Extending literature in internet studies that described how digital platforms act as gatekeepers (Bucher, 2013; Galloway, 2006; Gillespie, 2010; McKelvey, 2011; van Dijck & Poell, 2013), the following research shows that, beyond a discourse of neutrality, platforms shape the mediation between scholarly parties based on specific normative incentives. The demonstration is organized in three stages. First, data sharing platforms emphasize the large collection of heterogeneous data sets—facilitated by minimal selection, low deposit requirements for researchers, and minimal internal processing—and their publication through APIs. Second, the architecture of data sharing platforms, by providing automatic access to deposited research data through API, operationalizes “big data” research practices, based on pattern recognition over large quantity of data. Third, Complementing the analysis of the platform architecture with discourses from Figshare founder and manager, the paper reveals the epistemological inclinations of the platform logic towards “big data” research. It concludes that as Figshare positions their APIs at a central configuration for scholarship (with the larger aim to making science “programmable”), it is crucial to uncover the gatekeeping function that platforms play within the large infrastructure for knowledge, despite being typically hidden behind rhetoric of openness, accessibility, and neutrality.

The methodology to analyze this platform combines document analysis with interviews. The documents comprise Figshare blog posts, op-eds by its founder Mark Hahnel or interviews of him in specialized press, the data deposit interface, the Figshare API documentation, and an interview I conducted with Mark Hahnel in London in 2016. Combined altogether, this mixed-method approach reveals how the technical architecture and the discourses of the two entities shape the research practices and types of knowledge that can emerge from their mediation.

References

Acker, A., & Kreisberg A. (2019). Social Media Data Archives in an API-Driven World. *Archival Science*. https://doi.org/10.1007/s10502-019-09325-9

Borra, E., & Rieder, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, 66(3), 262–278.

boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, 15(5), 662–679.

Bucher, T. (2013). Objects of Intense Feeling: The Case of the Twitter API: Computational Culture. *Computational Culture*, 3. Retrieved from http://computationalculture.net/article/objects-of-intense-feeling-the-case-of-the-twitter-api

Driscoll, K., & Walker, S. (2014). Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *International Journal of Communication*, 8(0), 20.

Gaffney, D., & Puschmann, C. (2013). Data Collection on Twitter. In K. Weller, A. Bruns, J. Burgess, & M. Mahrt, *Twitter and Society* (pp. 55–68). New York, NY: Peter Lang International Academic Publishers.

Galloway, A. R. (2006). *Protocol: How Control Exists After Decentralization* (New Ed edition). Cambridge, MA: The MIT Press.

Gerlitz, C., & Rieder, B. (2013). Mining One Percent of Twitter: Collections, Baselines, Sampling. M/C Journal, 16(2). Retrieved from http://journal.media-culture.org.au/index.php/mcjournal/article/view/620

Gillespie, T. (2010). The politics of ‘platforms.’ *New Media & Society*, 12(3), 347–364. doi:10.1177/1461444809342738

Hogan, B. (2018). Social Media Giveth, Social Media Taketh Away: Facebook, Friendships, and APIs. *International Journal of Communication*, 12, 592–611.

Langlois, G., & Elmer, G. (2013). The Research Politics of Social Media Platforms. *Culture Machine*, 14(0). Retrieved from http://www.culturemachine.net/index.php/cm/article/view/505

Marres, N. (2017). *Digital Sociology: The Reinvention of Social Research*. Malden, MA: Polity Press.

McKelvey, F. (2011). A Programmable Platform? Drupal, Modularity, and the Future of the Web. *Fibreculture Journal*, 18. Retrieved from http://eighteen.fibreculturejournal.org/2011/10/09/fcj-128-programmable-platform-drupal-modularity-and-the-future-of-the-web/

Plantin, J.-C., Lagoze, C., & Edwards, P. N. (2018). Re-integrating Scholarly Infrastructure: The Ambiguous Role of Data Sharing Platforms. *Big Data & Society*, 5(1), 205395171875668. https://doi.org/10.1177/2053951718756683

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE*, 6(6), e21101.

van Dijck, J., & Poell, T. (2013). Understanding Social Media Logic. *Media and Communication*, 1(1), 2–14.

PRIVATE PLATFORMS, METADATA, AND THE ENCLOSURE OF DATA ACCESS

Amelia Acker
University of Texas at Austin

Two of the most urgent issues for internet researchers concern the rise of private platforms and their models for data access that rely on profit-driven standards and ontologies. Capitalist in their design, platforms aggregate user data en masse so that it can be transformed, bought, sold, and accessed (sometimes for free) as an integral part of a profit model of data expropriation away from users to data brokers (Gehl, 2014). The standards and ontologies used to classify and exchange data in private platforms are different from web standardization efforts and international standards engineering of the early Internet and open web (Russell, 2014). Platforms are frequently described as 'walled gardens' for user experiences. These walls are extended to inhibit and enclose methods for data access as well, because platforms rely upon the 'accumulation by dispossession' of data from individuals (Thatcher, O'Sullivan, & Mahmoudi, 2016).

The generation and reuse of metadata through the asymmetric creation and collection of user data in platforms is now an epistemic mark of our present data culture. Increasingly we see that the semantic relationships between users' data and data derived from their environments (such as work, gym, church, or commute) are classified with corporate taxonomies and then become central to the user experience and product design of personalized platform products from Amazon, Yelp, or YouTube. For example, Spotify recently announced its personas tool, which builds on years of user-centered design techniques to cluster and examine listening behaviors of user groups in the U.S. (Torres de Souza, Hörding, & Karol, 2019). Personas are now part of the platform's internal vocabulary to support the identification of users, categorize their listening habits, and drive algorithmic recommendations. Yet the classified personas label of Spotify users is not available to users who generated this tool. While metadata standards like the personas product taxonomy that describe people and their behaviors have always been essential to social platforms, the intensification and datafication of ICTs we now see proves that some user categories can and have been used in support of profiling, social sorting, and redlining minority and vulnerable populations with platforms (Eubanks, 2018; Noble, 2018).

With profit-driven platforms premised on targeted advertising based on user metadata, the power of legacy standards and the categorization of users has far-reaching consequences when deployed in big data applications, where large-scale data-mining and data analytics are dependent upon initial taxonomies and classification systems to sort users. More and more, we see engagement metadata being gamed for misinformation campaigns, media manipulation and disinformation efforts, even large-scale discrimination tactics across platforms (Acker & Donovan, 2019). For example, ethnic affinity categories are among the options that data brokers and advertisers can use to promote content and direct targeted content to users on platforms (Angwin, Mattu, & Parris Jr., 2016). Platforms typically classify users by a range of affinity categories that may be unknown to users themselves. These metadata structures can be used to control the information that is accessed by some users and not others, as well as categorizing content for specific users based on personalized predictions. These metadata are then used to display new content, push news alerts, and personalize search results to users through affinity categories on platforms such as Google, YouTube, Instagram, even navigation apps.

Grouping users into audiences for targeted advertising is not new. Data markets such as newspaper, television media, food service, and insurance markets existed for decades before mobile networks and social platforms. But with knowledge infrastructures that enable near-constant data creation and collection, platforms can leverage large-scale social networks, environmental sensors, and rapid data-processing to create new gateways of control and access to collections of data. So, in our data culture of constant creation and collection of data in platforms, metadata standards and ontologies that underwrite these networked infrastructures remain hidden and are often not accessible by most of the creators who produce them or researchers who wish to investigate them (Acker, 2018).

Another characteristic of this moment is that data creators are ceding control and access to their data to profit-driven platforms, whilst platform intermediaries' profit from consuming and providing access to these collections to data brokers who curate them for long-term value. Despite the fact that knowledge infrastructures can now extract data at scale, intermediaries and brokers assert control over collections of user data by enclosing access and inhibiting oversight, criticism, and research (Acker & Kreisberg, 2019; Bruns, 2019). This distance between creators who produce data and collection contexts where data and metadata are accumulated and stored in private platforms is where KI research can have swift and lasting impact. These uncertain archives are made of data and metadata accrued from big data apparatuses such as transportation and mapping apps, social media, mobile devices, learning management systems, and internet infrastructures assembled by data intermediaries to resell and repurpose user data to data brokers. Few of these data archives are accessible to the creators who produced these digital traces and are impacted by these categories the most. Indeed, platform users cannot opt out of affinities, audience profiles, schemas, corporate taxonomies or personas once they have been classified as such, because these metadata standards and ontologies belong to the intermediaries and brokers who control them. While these metadata are not always understood or preserved for the people who created them, they are actively leveraged by platform intermediaries, data brokers, and third-party data consumers to create personalization profiles, predictive analytics, algorithmic recommendations, and user data collections.

Internet infrastructures reinforce and redistribute authority, power, and control with gateways like metadata standards and ontologies. Today, creating data in platforms is a form of belonging, but there are few avenues in place for users to control or access metadata generated about them, or to withdraw from categories once they have been sorted and enrolled into them. If the space that separates users from collections of their data is an epistemic mark of our current moment, then the access and enactment of metadata in private platforms will continue to be how this space grows and more walls are built. It is imperative for internet researchers to confront not just the commodification of data in private platforms, but the impact of enclosure of data access regimes we are witnessing with their rise.

References

Acker, A. (2018). A Death in the Timeline: Memory and Metadata in Social Platforms. *Journal of Critical Library and Information Studies*, 2(1), 27.

Acker, A., & Donovan, J. (2019). Data craft: A theory/methods package for critical internet studies. *Information, Communication & Society*, 0(0), 1–20. https://doi.org/10.1080/1369118X.2019.1645194

Acker, A., & Kreisberg, A. (2019). Social Media Data Archives in an API-driven world. *Archival Science*. https://doi.org/10.1007/s10502-019-09325-9

Angwin, J., Mattu, S., & Parris Jr., T. (2016, December 27). Facebook Doesn't Tell Users Everything It Really Knows… [Text/html]. Retrieved August 31, 2018, from *ProPublica* website: https://www.propublica.org/article/facebook-doesnt-tell-users-everything-it-really-knows-about-them

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 0(0), 1–23. https://doi.org/10.1080/1369118X.2019.1637447

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press.

Gehl, R. W. (2014). *Reverse Engineering Social Media: Software, Culture, and Political Economy in New Media Capitalism*. Philadelphia, Pennsylvania: Temple University Press.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism* (1 edition). New York: NYU Press.

Pomerantz, J. (2015). *Metadata*. Cambridge, Massachusetts; London, England: MIT Press.

Russell, A. L. (2014). *Open Standards and the Digital Age: History, Ideology, and Networks*. New York, NY: Cambridge University Press.

Thatcher, J., O'Sullivan, D., & Mahmoudi, D. (2016). Data Colonialism through Accumulation by Dispossession: New Metaphors for Daily Data. *Environment and Planning D: Society and Space*, 34(6), 990–1006. https://doi.org/10.1177/0263775816633195

Torres de Souza, M., Hörding, O., & Karol, S. (2019, March 26). The Story of Spotify Personas. Retrieved May 2, 2019, from *Spotify Design* website: https://spotify.design/articles/2019-03-26/the-story-of-spotify-personas/

TOWARDS A SOCIOLOGICAL UNDERSTANDING OF KNOLWDGE GRAPHS

Huw Davies
University of Edinburgh

Rebecca Eynon
University of Oxford

Why Should We Care about Knowledge Graphs?

Many of the world's richest and influential technology companies such as Google, Facebook, Microsoft, eBay, and IBM draw value from their knowledge graphs. Knowledge graphs can be used to unify and link different sources and types of data, build sophisticated models for machines to understand our social world, deanonymize data in ways that challenge our privacy, and help AI pretend to be human. Facebook, for example, is using its knowledge graph to create a "structured understanding of music and lyrics" (Noy, Gao, Jain, Narayanan, Patterson, & Taylor, 2019) that can be used to "detect when people are referencing them" during conversations on its platform so that the company can monetise "serendipitous moments between individuals" (ibid). Yet, despite this technology's importance, the inner workings of knowledge graphs are unfamiliar to sociologists (Halford, Pope, & Weal (2013).

What are Knowledge Graphs?

Knowledge graphs synthesize more recognisable ways of managing data. For example, they are databases that can be interrogated via structured queries and their properties can be analysed like social networks. However, unlike databases and social networks, knowledge graphs are also knowledge bases within which data is labelled and characterized with formal semantics. These three affordances combined, allow owners of knowledge graphs to search across and within large and previously incompatible datasets and analyse and interpret data in ways that enable machines to infer new knowledge about our world.

Knowledge graphs can operationalise five forms of semantic modelling instrument: entities, classes, relationships, categories, and ontologies. Classes enable entities to be put into logical hierarchical categories. For example, a person is an entity that belongs to the class 'human' who has the relationship 'works for' an entity 'company' that belongs to the class 'corporation': her job within the company maybe within the category of 'managerial'. Entity descriptions often include a classification of the entity with respect such a class hierarchy. For example, a manager has a 'higher salary' than an administrator. The relationships between entities are usually tagged with types, which provide further information about the nature of the relationship. For example, the administrator is 'subordinate to' the manager. It is possible to add 'human-friendly text' to further clarify design intentions for the entity and improve searches. Ontologies further formalize these ways of organizing data to make them more robust, standardized, and legible to machines. Ontologies serve as prescribed taxonomies that integrate logic so that the model without knowing any specific salary figures would logically assume two mangers in the same role would have the same salary. However, such systems are only as good as their logics. For example, the sociology of gender pay gaps would need to be integrated into the model for the machine to understand structural sexism in the workplace and why male and female managers may not be paid the same.

Machines are not good at processing such anomalies. Therefore, knowledge graphs often operationalise socially constructed and sometimes problematic ways of reflecting human knowledge. Human subjectivity becomes hard coded into systems which are then treated as embodiments of objective processes. Graham and Ford (2016), for example showed, Jerusalem's highly politically sensitive status was hard coded into Google's knowledge graph. Yet sociologists with expertise in knowledge production and all its socially constructed ambiguities know relatively little about knowledge graphs and adjacent semantic web technologies (Halford, Pope, and Weal (2013). How can this be changed; how can knowledge graphs be rendered more legible to sociologists?

A Way into Knowledge Graphs

Given their commercial value to companies, proprietary knowledge graphs are a typical 'black box' technology that they don't want to reveal to competitors. However, there are ways of studying knowledge graphs that offer sociologists opportunities to critically engage with them and there are opportunities to develop knowledge graphs as a sociological method.

To develop their knowledge graphs, many of the technology companies apply for patents. In 2018, for example, Facebook, applied for a patent its model of social class to be integrated into its knowledge graph United States Patent Office (2020). See Figure 1.

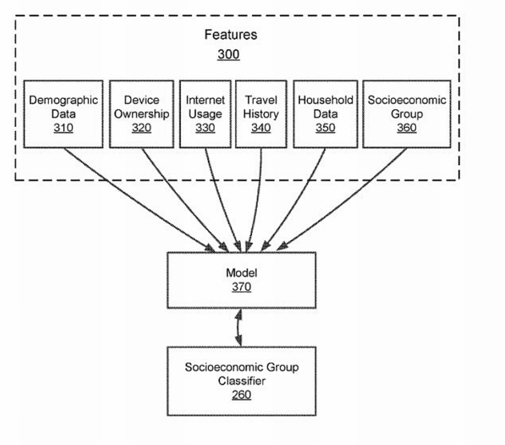


Figure 1

Drawing on such examples can facilitate a better sociological understanding of knowledge graphs and highlight the salience and relevance of knowledge graphs to a broader research community. We can query Facebook's ontological categories such as 'Socio-economic Group' and the data such as 'Travel History' that Facebook uses to justify these categories.

Knowledge Graphs a Sociological Method

Knowledge graphs are also available to sociologists as a method. This is enables us to expose their inner workings to a non-specialist audience and show opportunities for us to engage with technologists in the construction of computerised knowledge. In our example, we model the political economy field educational technology (Davies and Enyon, 2020). To produce our knowledge graph, we use a graph database engine called Neo4j and combine data from various sources including interviews, financial records, and automated web searches. We show how we classified our entities, characterized relationships between entities with semantic tabs, and added contextual data such as hyperlinks to justify our decisions and render them transparent. Using graph measuring techniques such as PageRank, we are able to gain new insights into this political economy such as its dominant capital investors. See Figure 2 for a screen shot from this graph. PageRank identified the central node in this graph, Learn Capital as a dominant investor in the political economy of edtech. The pink and blue nodes are edtech companies that Learn Capital is investing in. The purple and green nodes are other investors.

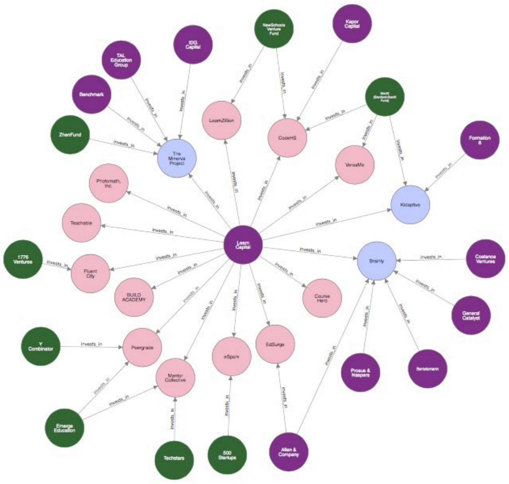


Figure 2

These two strategies, analyzing patents and developing our own knowledge graphs, promise sociologists a better, more meaningful understanding of this important technology.

References

Davies, H., & Eynon R. (2020). The Mobilisation of AI in Education: A Bourdieusean Field Analysis. *Sociology*. (in 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

Halford, S., Pope, C., & Weal, M. (2013). Digital Futures? Sociological Challenges and Opportunities in the Emergent Semantic Web. *Sociology*, 47(1), 173–189. https://doi.org/10.1177/0038038512453798

Noy N., Gao Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2018). Industry-scale Knowledge Graphs Lessons and Challenges. *Semantic Web Conference* in Asilomar, California, in October 2018 Retrieved from http://iswc2018.semanticweb.org/panel-enterprise-scale-knowledge-graphs/

US Patent and Trademark Office (2020). Systems, Computer-Readable Media, and Methods for Activation-Based Marketing. Retrieved from http://appft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=%22socioeconomic+group%22&OS=%22socioeconomic+group%22&RS=%22socioeconomic+group%22