

Semantic Web Practices: Infrastructural Politics and the Future of the Web

Science, Technology, & Human Values

1-25

© The Author(s) 2024



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/01622439241249573

journals.sagepub.com/home/sth



Susan Halford¹ , Mark Weal² ,
Faranak Hardcastle³, Nicholas Gibbins²,
Samantha Pearman-Kanza² and Catherine Pope³

Abstract

In the past thirty years, the Web has developed from its inception as a layer of protocols on top of the Internet to use by more than 5 billion people and organizations. This has driven the creation of vast quantities of data and led to deep concerns about the politics of digital data and computational methods. To date, critical investigation of these concerns has focused on large commercial platforms built on top of the Web, and their use of machine learning methods. Meanwhile, less attention has been paid to the underlying design and protocols of Web itself, and how these might be implicated in the very same process and concerns. We explore ongoing endeavors to transform the Web from a library of documents intended for humans to a “semantic Web” using symbolic artificial intelligence to enable

¹University of Bristol, United Kingdom

²University of Southampton, United Kingdom

³University of Oxford, United Kingdom

Corresponding Author:

Susan Halford, University of Bristol, 13 Berkeley Square, Bristol BS8 1QU, United Kingdom.

Email: susan.halford@bristol.ac.uk

machine reasoning across multiple heterogeneous data sources. In principle, this would transform the production and circulation of knowledge at Web scale. We present the findings from an experimental, interdisciplinary study exploring the epistemological politics and sociomaterial practices involved in situated accomplishment of the semantic Web. Our findings have consequences for the future of the Web and the future of Web-based platforms.

Keywords

semantic Web, digital infrastructure, platforms, symbolic AI, interdisciplinarity, autoethnography

Introduction

The past thirty years have seen the remarkable growth of the World Wide Web. From its inception as a new layer of protocols on top of the Internet, the Web is now used by 5 billion people¹ and is embedded in economic, social, and political life around the world. It is, by any account, an infrastructure for the twenty-first century, “as crucial to modern life as electricity, telephones and sewers” (Plantin et al. 2018, 10). In turn, the rapid growth of the Web has generated digital data at an extraordinary rate and scale, fueling a new round of artificial intelligence (AI) and raising deep concerns about the politics of digital data and computational methods. To date, the focus of these concerns has been the new “platforms” built on top of the Web, particularly large commercial organizations such as Facebook, Amazon, and Google. Indeed, a sharp distinction is made with the nonproprietary, distributed and open qualities of the “original” Web, sometimes represented as a “public commons” for users. In contrast, the new platforms draw users into proprietary walled gardens where their activities leave digital traces, which can be commercialized through machine learning methods. Indeed, it is sometimes suggested that as these platforms continue to grow, they will rework the “open Web” putting its very future at risk (Helmond 2015; Plantin et al. 2018).

Our starting point is that this distinction between the open Web and platforms obscures critical questions about the Web as infrastructure, and its implication in related concerns about the politics of data and computational methods. In this account, the Web sits in the background as a neutral enabler for platforms, which are the primary focus of attention. From an

infrastructural perspective, the fact that it is taken for granted is significant. As Jackson et al. (2007) note, the most effective infrastructures come to “appear as timeless, un-thought, even natural features of contemporary life” (Jackson et al. 2007, n.p.). However, this invisibility belies the activity that goes into producing such infrastructures, which are never fixed or finished “things” but made and remade in ongoing sociomaterial practice. As such, infrastructures embody values and politics, the outcome of choices is shaped by debate and controversy (Bowker et al. 2010), carrying these choices forward, often with profound consequences for what they do, and do not, enable (Winner 1986; Star 1999).

This paper takes the Web as the focus of infrastructural inquiry. This has particular importance at present, as long-standing ambitions gain ground in transforming the open Web from its original form into a semantic Web. While the Web was originally conceived as a network of links between documents and intended for human readers, a semantic Web adds structured machine-readable meaning to the data inside documents. Enormous quantities of data are published on the open Web, from a vast array of sources, on all subjects: far more data, and more diverse data than in any of the new platforms. A semantically enabled Web would extract these data from their documentary siloes and—taken to its logical conclusion—would enable machine reasoning across all data published on the open Web. In contrast to the machine learning AI that has been the focus of critical attention in studies of platforms, this application of symbolic AI would—in principle—transform the Web into a single-linked database for computational analysis. Proponents’ claims are underpinned by a positivist epistemology where data might be readily modeled and linked for the greater good. However, this positivist epistemology belies the infrastructural work required to achieve a semantic Web—particularly the epistemological and ontological work required to create semantic linked data—and its consequences. To critically examine the infrastructural work involved in building the semantic Web, this paper presents the findings from an interdisciplinary empirical experiment designed to apply semantic Web techniques to social research questions, specifically related to social class, aging, and health.

How this work is done matters, with significant implications for the global production, control and circulation of knowledge. What is at stake is how knowledge is represented and circulated in the largest information construct in human history. However, shifting analysis from a high-level recognition of what is at stake to empirical evidence of how these processes play out is a challenge. This is technically complex work, not usually open or readily amenable to critical sociological investigation. Directly addressing

this challenge, we report on an interdisciplinary and experimental research project, designed to explore the “situated accomplishment” of the semantic Web as “a grandiose theoretical concept” (Woolgar and Lezaun 2013, 322) by tracing “the fluid and unstable processes” (Van Heur, Leydesdorff, and Wyatt 2012, 342) through which semantic linked data are produced. Overall, our paper unfreezes the sociomaterialities of semantic Web as an emergent infrastructure. This has significant consequences for the future of the open Web, because online data are described and modeled by tools which require conventions and formats to be put in place, unlike the original permissionless Web (Halpin and Monnin 2016). It is also significant for the future of the platforms that have dominated debates on data and AI to date not least because the open Web provides the underlying infrastructure for these platforms (Mukherjee 2019). Further, the major corporate platforms themselves are also investing in semantic Web technologies (Halpin and Monnin 2016). Relatively little is known about these highly sensitive commercial operations, but our paper provides some insights into the potential consequences of semantic Web technologies for data and the knowledge that is derived from them. We show the importance of understanding the sociomaterial practices involved in creating the semantic Web.

We begin by discussing the Web as infrastructure, the recent shift in attention from the original Web to digital platforms, and how this shift raises critical questions about the ontological and epistemological politics of digital data and artifacts. We then introduce the semantic Web and its knowledge practices. Following an overview of our methodology, the paper focuses on the empirical material, identifying five key junctures in the project, which we use to investigate wider issues involved in creating infrastructure, including its consequences from both computational and social research perspectives. We conclude by reflecting on the significance of these findings and widen out to consider how these same processes might shape the semantic Web at scale, including implications for the future of platforms.

From the Open Web to Platforms

The history of the Web is often framed in public discourse as an “invention,” originating in the proposal² for a new “global information sharing system” made by Tim Berners-Lee in 1989. However, like all infrastructures, the Web did not appear *de novo* but was a response to the limitations of the Internet and pieced together from existing technical artifacts. The Internet had been steadily expanding the physical network of

linkages between computers since the 1960s but file transfer protocols remained complex, cumbersome, time-consuming, and a specialist/expert preoccupation. By contrast, Berners-Lee's proposal combined a simple, standardized document format, building on an existing markup language (SGML), with an hypertext and hypertext transfer protocols (HTTP) built on top of the Internet. Using a "uniform resource locator" referencing the Internet domain name system, documents could be published and accessed remotely regardless of bespoke operating systems and software.

The early 1990s saw the gradual enrolment of individuals and organizations running their own Web servers, using the new protocols and growing the nascent Web infrastructure, for example, through the launch of effective search engines, standardization of protocols through the World Wide Web Consortium (W3C), and co-alignment with Internet governance (Musiani 2015; Berners-Lee 2000; Brügger et al. 2019). Despite some early attempts to patent and commercialize the Web, over time, it has been enacted and grown through an extensive network of actors beyond the control of any individual corporation or government (Plantin and Punathambekar 2019).

In short, the Web is not a "thing" but a set of loosely orchestrated and dynamic sociomaterial practices. In 2005, the advent of social media sites led to the branding of Web 2.0 (O' Reilly 2005), initially distinguished from the original Web through the radical extension of user-generated content. Where Web 1.0 had enabled those with the skills and resources to publish a website online, Web 2.0 enabled anyone with access to the Web to create online content through simple interfaces. The growth in use that followed was extraordinary. Facebook acquired 315 million accounts in its first five years (2004-2009), reaching 1 billion by 2012 and just under 3 billion by 2023. Increasingly, these were not just websites but a distinct form of commercial enterprise built on top of the Web, now widely known as platforms (Gillespie 2010). Critical analysis of these platforms is extensive, and centers largely on sociopolitical concerns about data and computational methods. Detailed information on individuals and organizations becomes the property of commercial companies, who aggregate and interrogate the data using machine learning to construct new social forms of social analytics, marketing, and targeting. This has produced dramatic effects, for example, the use of Facebook data by Cambridge Analytica as part of the Brexit campaign in the 2016 referendum on UK membership of the European Union (Risso 2018). More insidiously, online behavioral tracking has become a standard commercial activity (even with governance such as the

General Data Protection Regulations in the European Union), and sharing of individual data in the practice of everyday life is routinized.

At a more detailed level, critical platform studies pay attention to their sociomateriality. For example, how the technical architecture of platforms shapes online socialities as well as the data that are created and the forms of knowledge derived from these. The establishment of “friends” and “followers” or “likes,” for example, can have profound effects on social interactions and identities (Van Dijck and Poell 2013) and has created artifacts for social media analytics, however questionable their value may be (Halford et al. 2018). Following Bogost and Montford’s (2009) injunction to take platforms seriously as computational infrastructures, Helmond (2015) describes the computational work that drives platforms forward, showing how platforms are—increasingly—extending their reach into the open Web, instating themselves on external websites (e.g., through widgets) and using a variety of methods to draw Web data into their own databases.

This is important analysis. Where we differ is in the characterization and positioning of the “open Web” in the analysis of platforms. In these accounts, it appears almost as if the Web “stopped” in the mid-2000s and is now a relic from an earlier era, benignly enabling the rise and growth of the new platforms, but now a passive prey to the inexorable rise of platform logic (Holmes 2013), which taken to its logical conclusion will mark the demise of the Web (Plantin et al. 2018). This stands in contrast to historical studies which emphasize the ongoing sociomaterial production of the Web (cf., Barnet 2019; Musiani and Schafer 2019). Extending this point, the Web must be seen as a live, enacted infrastructure, rather than an ossified thing. Indeed, this construction of the Web as static or superseded obscures its implication in some of the same questions about the politics of data, AI, and infrastructure as those raised in studies of platforms.

The Rise of the Semantic Web

In this paper, we focus on a different and longer-term challenge to the original Web, specifically ongoing challenges to its grounding as a document-based system, rather than a data-based system. While vast volumes of data were published inside documents, the established protocols offered no way of link across these data. A “semantic Web” would add machine-readable meaning to data published online and use computational tools to support machine reasoning across the Web. Machines would be able to take data from one source to complement or extend data from other

sources, combining information at Web scale “to make inferences, choose courses of action and answer questions.”³ To achieve this, a standardized knowledge representation system was required to model data entities (people, places, and things), their properties, and their relationships. While the original Web, demanded only limited standardization to make simple links between documents, the semantic Web demanded a “full blown language for knowledge representation” (Halpin and Monnin 2016, 6) to enable computational reasoning across Web data.

Like any infrastructure, the semantic Web has been subject to controversy and dispute. In the semantic Web community, this has been characterized by two different approaches reflecting longer-term divisions on the philosophy of AI. The initial approach to building a semantic Web was shaped by the declarativist tradition “built on a foundation of logical axioms that precisely described the permitted inferences of any statement made” and “first order logic that could express any and all knowledge to be published on the Web” (Halpin and Monin 2016, 129). From this perspective, a semantically enabled Web would depend on the development of consistent and durable global semantics. This approach raises two key challenges. First, the scale of knowledge engineering work required given that adding structured and consistent meaning to data is a major undertaking. Doing so in a shared and consistent way demands coordination and collaboration across multiple sites of data publishing in order to realize interoperability. Second, concerns were raised about the rationalization of a permission-less Web (Halpin and Monnin 2016) into a structured information system. What forms of knowledge could and would be expressed by semantic Web technologies and what may that mean for the future of the Web? (Shirkey 2006; Halford et al. 2013; Ford and Graham 2016; Mccarthy 2017). The Web infrastructure would no longer be indifferent to content but, rather, would depend on prescriptive ways of describing entities and their relationships. These would be reproduced at scale and—at the same time—obscure to those not familiar with their philosophical and computational underpinnings. In this context, there were concerns about what would happen to partial, inconclusive or context-sensitive information (Ford and Graham 2016) in the push for global semantics. In contrast to a declarativist approach, a growing “linked data” community was making use of semantic Web technologies for individual, localized projects marking a shift toward a “proceduralist” philosophy. Rather than depending on generalizable statements, proceduralists understand intelligence as know-how, where the tools for reasoning are inseparable from domain knowledge, focusing on data infrastructures rather than logical modeling (Winograd 1975). This is a

more specific and partial approach driven by local priorities which values pragmatism and simplicity over formalism and consistency (Poirier 2017).

These two approaches to the semantic Web resonate with wider and long-standing distinctions in Computer Science between “neat” and “scruffy” (Poirier 2018). This brings philosophical distinctions down to earth, focusing on differences between theoretical and applied approaches in computational practice. In short, neat is principled, consistent, and interoperable; scruffy gets something done albeit in a particular context. Poirier (2018, 359) describes the history of the semantic Web in the 2010s as a “turn for the scruffy,” linking this both to a proceduralist position on knowledge representation and in pragmatic recognition of the political–economic considerations that shaped semantic Web activity, such as end goals, deliverables, funding, and market demand. In short, existing accounts suggest that the semantic Web is emerging in distinct, diverse forms. What we lack is empirical evidence of the consequences of these different approaches for the forms of knowledge produced. In the remainder of this paper, we present findings from an interdisciplinary experimental project designed to explore how the semantic Web is “done.”

Methodology

Our project *Social Sciences, Social Data and the Semantic Web* was designed specifically to explore the epistemological and ontological stakes of the semantic Web. In designing our research, we faced the usual challenges of studying infrastructure, which by definition “typically sits in the background . . . is invisible, and . . . frequently taken for granted In such a marginalised state its consequences become difficult to trace and politics are easily buried in technical encodings” (Bowker et al. 2010, 98). Our methodology was designed to unfreeze (Star 1999) the semantic Web by surfacing the invisible work involved in re/producing semantic linked data. Specifically, our project was interdisciplinary, experimental, and autoethnographic.

The project was designed by a team of two computer scientists and two sociologists and employed two researchers with recent interdisciplinary PhDs in social and computational sciences. It had small funding⁴ from a UK Economic and Research Council program expressly intended to support blue skies research. This gave us “the courage to try something new [in] a permissive environment” that supported interdisciplinary research (Bijsterveld and Aagie 2023, 2). The key research questions were framed by sociology, but it did not provide the disciplinary expertise to untangle the

technical complexities of semantic Web technologies. The computer scientists were experienced with semantic Web technologies, but their training had not prepared them for the epistemological or substantive questions raised by this project. For all of us, this was without a doubt the most challenging project we had ever worked on, and the riskiest by far as it fell way outside the usual norms and rewards of our home disciplines.

Working together, we sought to explore how semantic Web technologies “materially organize and instantiate relations between people, things, perspectives, and technologies” (Gray, Gerlitz, and Bounegru 2018, 1). All members of the team had worked together previously and shared a commitment to interdisciplinary collaboration for critical examination of Web technologies. Our collaboration had developed over time, through shared teaching and research activities that had led to “shifts in how we read, value concepts, critically combine methods, cope with knowledge hierarchies and adopt writing styles” (Bijsterveld and Aagie 2023, 3-4). In Schubert and Kolb’s (2019) terms, this was a symmetrical engagement in which computer science and sociology brought distinct and in-depth expertise to the project and neither was in the service of the other.

Our approach was collaborative experimentation (Balmer et al. 2015) based on converting two conventional social data sets into semantic linked data to make the implicit infrastructuring processes visible. Including social scientists as participants in the process, rather than as external observers of a computer science project allowed us to open-up challenges and resolutions that would otherwise remain unarticulated (Balmer et al. 2015). Our approach was experimental in two senses. First, we were agnostic regarding different approaches within the semantic Web. We did not have a pre-defined approach for how the semantic Web should be done, our aim was simply to trace how it happened. Second, the process itself was the object of inquiry rather than the means to an end. The experiment was to see how this process evolved and it is this that we concentrate on here.

The experiment was grounded in a specific case: creating semantic linked data for research on social class, aging, and health. We chose this topic for two reasons. First, it is characterized by conceptual disputes, complex research questions and highly dynamic data. This was a deliberately tough test case for semantic Web technologies, intended to render taken-for-granted infrastructural practices more visible than they may otherwise have been. Second, there were two publicly available data sets available for experimentation: the Great British Class Survey (Savage 2018) and the English Longitudinal Survey of Ageing (ELSA Wave 6 2012). The GBCS was a pioneering study driven by Bourdieusian approaches

to social class, which moved beyond occupational categorizations to include economic, social, and cultural capitals as a basis for defining a new class structure for the UK (Savage 2018). However, it had little information on aging and health. The English Longitudinal Survey of Ageing was rich with information on age and health and included economic, social, and cultural data but relied on occupational definitions of social class (NS-SEC). These are two separate surveys, and we cannot know if any of the same individuals appear in both data sets. We did not set out to link information on individuals. Rather, our aim was to critically examine the ontological and epistemological consequences of using semantic Web technologies to infer GBCS social classes to explore health inequalities in the ELSA population.⁵

Our approach was autoethnographic. We documented and reflected on our activities throughout the project.⁶ Data collection took place from January 2019 to January 2020. Three main types of data were collected: (1) audio recordings and photographs from team meetings and the project Advisory Board; (2) field notes and photographs from technical meetings; and (3) a full record of all technical development in GitHub. Initial thematic analysis of the data was undertaken by the lead ethnographer and discussed in depth at full-team data workshops. In this process, we identified five critical junctures across the lifetime of the project where important epistemological and ontological questions about our use of semantic Web technologies surfaced. In what follows, we describe each of these junctures and open-up what was at stake for our project. Because these insights are presented for a social science audience here, the conventions of writing lean more in this direction. However, because our argument focuses on the materiality of the semantic Web, we focus on technical detail to a degree that means what follows may be less familiar for a social science audience. However, it is precisely this attention to detail that allows us to unpack the situated accomplishment of the semantic Web and to explore what is at stake in these knowledge practices.

Creating Semantic Linked Data

In this section, we explore five junctures during the project. All entailed debate about the best way forward and involved both the sociologists and the computer scientists. Although the focus was always substantive—how best to represent the social research data—the choices centered on differences between declarativist and proceduralist approaches, shaped within a wider landscape characterized by local pressures and contingencies. In what follows we describe these choices chronologically.

We began by thinking about how to represent social class and aging in a machine-readable way. In the semantic Web, “ontologies”⁷ provide a formal description of the classes of object, their properties, and relationships in a field. The *first key juncture* concerned whether to create a conceptual ontology of social class, aging and health, or an applied ontology derived from the data sets we planned to use. The former aligns with a declarativist perspective, using logic to model definitive statements about what can exist. The latter aligns with a proceduralist approach, using a specific artifact (survey instruments in this case) to create a model of the field.

In computer science:

What we mean when we talk about an ontology is an engineered artifact, it is a shared conceptualization of some domain or area of study or area of interest. (Computer scientist 2, Advisory Board Meeting notes, March 15, 2019)

This can often involve importing established schema from outside the semantic Web:

On other projects I’ve worked on it was “here’s your list of seed types.” It’s not something you can argue about . . . it just exists, it’s schemas with pre-existing hierarchies and you just have to convert it into semantic form, whereas we spent a lot of time discussing . . . how best to represent the objects in the ontology. (Computer scientist 3, Team Meeting, March 22, 2019)

And there were already identified data sets for the project:

As soon as I saw the data sets and realized that they needed to be linked together in some way, my assumption went down the data route . . . maybe that’s different from the original view of the semantic Web (Computer scientist 3, Team Meeting, March 22, 2019)

From a sociological perspective, the means to construct an ontology were less clear:

We could have analyzed the data in different sorts of ways, there was a fight in the team . . . [and] of course there are other people who may have done it differently. (Advisory group member 1, Advisory Board Meeting, March 15, 2019)

Collectively, we were all acutely aware of outputs committed to the funder and the tight deadlines. An applied ontology was easier and quicker

than a conceptual ontology. This was a scruffy decision, driven by pragmatic needs rather than taking a principled approach to create a comprehensive ontology which would have been “neater” and (likely to be) of greater use beyond our project. The critical consequence of this decision was to narrow the representation of the domain to a particular artifactual source, deferring key decisions about what was in the ontology to the GBCS survey (as we modeled this first). This meant the original theoretical and methodological approaches in GBCS became hardwired as “the field,” rather than part of a contested field. We also inherited some questionable classifications from GBCS (e.g., a binary representation of gender) and a specific set of activities that were important from a Bourdieusian perspective (e.g., social engagement and cultural consumption). Once these decisions are represented in the ontology, they implicitly restrict what is knowable whenever the ontology is used, whether for its original purpose or as it is repurposed by others.

Once we made the decision to base our ontology on GBCS, *a second key juncture* concerned debate over *how* to model the ontology. Initially, we began by organizing this around cultural, social, and economic capitals based on the structure of the GBCS survey, intending to pick out the implicit semantics for the ontology (informed by reading the underpinning sociological research). This proved difficult. Semantic Web technologies lacked the expressivity to capture theoretical nuances:

I think those things [social class] are contested among sociologists and the danger is that if we model that as a top-level thing, somebody else is going to come along and say “well actually I think *that* should be cultural capital and *that* shouldn’t be” . . . I’m worried that some of the terms we are using here . . . are a bit more contested. (Computer scientist 1, Team Meeting notes, April 3, 2019)

By sticking to the GBCS questionnaire, we risked reifying decisions already made by social researchers in the 2010s. Not only definitions of social, economic, and cultural capitals but also tying survey responses (facts, preferences, and activities) to these categories. In turn, this raised issues of interoperability:

In looking for GBCS’s measures of cultural capital in ELSA, I have encountered various dilemmas . . . about if and how variables could be considered relevant to GBCS measures of cultural capital. (Social scientist 2, email, April 30, 2019)

If the ontology was to be useful for different users and contexts, a more generalized way of representing knowledge that didn't concretize theoretical positions derived from the GBCS survey was needed. We agreed on a radical simplification, removing social, economic, and cultural capitals as the organizing features of the ontology, separating out the underlying theory and the survey instrument as design principles. We chose to use the more generic classes of "person" and "activities" as the common denominator. Effectively, the ontology became distanced from explicitly Bourdieusian class analysis to provide the (apparently) least contested description of the data. This made it easier to link across our two data sets—since we knew that ELSA did not have the same terms and measures as GBCS. It also meant that this ontology could be used more widely by others.

So, you take a data set, in this case GBCS, and represent it in such a way that we could use that representation for the job of data integration between different data sets so we could do reasoning with it. (Computer Scientist 2, Team Meeting notes, May 21, 2019)

For better or worse, this represented a shift back from the procedural approach—modeling the data—to a more declarativist position, making definitive claims about what entities exist (person and activity) rather than modeling a survey instrument. A more declarative approach opens the potential for reuse. The ontology no longer had any explicit Bourdieusian theoretical framing nor contextual detail about its origins. It now defined facts about people (income and other sources of wealth, for example) as well as things that people do (activities) but not from an explicit theoretical position. By now, this is not an ontology of class but an ontology of facts about individuals and the activities they engage in—albeit still based on GBCS—that may be used for a variety of purposes. By not imposing rigid categorical or conceptual decisions, we leave it open to others to overlay their views or categorizations on top of the base ontology. A consequence of this is that information about the Bourdieusian framing of the data collection becomes lost as categorical aspects of the instruments are removed from the modeling. However, reuse may require the extension of the base ontology and should there be requirements for a broader set of activities to support different ways of interrogating the data.

Once we had a basic ontology design, drafted the *third key juncture* was about how to model time. This was in part prompted by the longitudinal nature of the ELSA data set, but more generally by the semantic Web imperative to create resources that can be reused and repurposed over time.

When a fact was recorded (e.g., house price or income) is important. With the declarative emphasis on standardized and consistent models, we sought to use an existing ontology. However, those created for describing temporal concepts (topological ordering, instants or intervals) were too specific, and Web Ontology Language-Time—the declarative standard for the semantic Web—was more extensive, demanding a time stamp for each piece of data: “I mean OWL-Time is a bit [heavy] . . . it is not quite clear what all the things in OWL-Time do” (Computer scientist 2, Team Meeting, May 21, 2019).

This made it very costly computationally at the expense of scale and speed. Instead, we turned to the smaller yet declarative “fluents” ontology, created to manage inference across key aspects of time in a computationally efficient way (Welty and Fikes 2006). This ontology is declarative, neat, and formal (based on situational calculus). Adopting it meant reconfiguring the ontology, adding temporal parts for every individual and allowing a common time stamp for data collected in a single survey, rather than an individual time stamp for each piece of data. A range of other temporalities may also appear, including frequency (of particular activities, for example) or duration (periods of employment or residency, for example). Adopting this modeling approach established critical path dependencies for later in the design process as explained below. Decisions on modeling time had consequences for the tractability of types of temporal queries or how that data might be consumed by larger systems with different overarching temporal models.

The *fourth key design juncture* occurred at the point of reintroducing social class into our toolkit. By taking Bourdieusian capitals out of the data-based ontologies, we had simplified and generalized them to support reuse by other researchers who could now query the data in a highly flexible way. However, we *were* interested in Bourdieusian capitals and the GBCS social classes (and we suspected that many other social researchers would be too). Specifically, we wanted to create the computational means to derive GBCS social classes from ELSA (as a test case), which would then allow us to integrate GBCS classes with analysis of health and aging. We were also interested in overlaying different understandings of class on top of existing data sets, in our case comparing GBCS with NS-SEC to examine health inequalities among older adults. Given our previous decision to model the data and not the domain, and then to focus on “activities” as the core of our ontology, we now had to create a *separate* ontology for social class. Both the GBCS and NS-SEC class models proved easy to model declaratively, drawing on the underpinning research to define the different class hierarchies conceptually. But to produce actual class GBCS dispositions from ELSA, we needed to engage with empirical data.

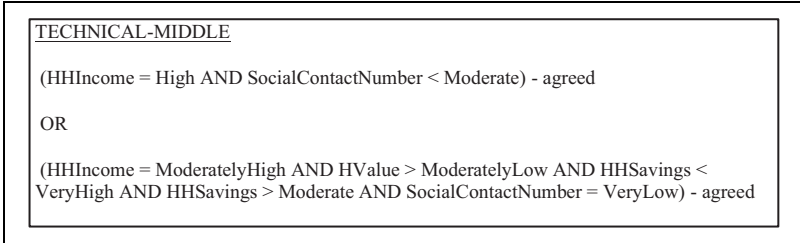


Figure 1. Fragment of description of ontology inference rules derived from Great British Class Survey literature.

The semantic Web tool of choice for this is inference, which gives machines a set of abstract rules to work out knowledge they don't have from knowledge that they do have. That meant providing rules to derive GBCS social classes from capital measures (social, cultural, and economic) in a data set (ELSA) that did not use this model of class analysis. These rules are written inside the ontology and published along with it to make them discoverable and usable by other machines to interrogate and examine other data sets. We modeled the inference rules from the sociological research (Figure 1).

So far, so neat. However, while inference works with logic, some forms of inference need statistical information about the data set (e.g. categorical thresholds). The symbolic approach to AI embedded in semantic Web technologies is not well suited to complex mathematical calculations, such as inferring class membership based on values measured against thresholds. These processes involved knowledge that did not reside within the data sets:

In the descriptions for economic capital for seven classes, words like “very high” and “high” are used, but do you have any qualifiers for this? Currently, my approach was to assume the numbers in Table 6 were an average of the scores and then fit the categories around that, but obviously that will only work if those numbers are the average. (CS1, questions to GBCS team, June 7, 2019)

Our pragmatic solution was to supplement our inference rules in the ontology with an algorithm informed by empirical evidence from the findings of the GBCS to calculate the capital measures for individuals that the inference rules would then use to reach a class disposition. This supplements the declarative and neat inference rules that would allow machines to

make deductions, with proceduralist and scruffy algorithms providing a specific set of instructions that tell machines exactly what to do with a specific data set when directed to do so. The thresholds used in these algorithms were derived for the specific data set through analysis, and this knowledge resides external to the semantic Web data and would require generating for any new data set, placing further limitations on reuse.

So far, we have described the preparatory work involved in creating a package of tools to support the sociological analysis of GBCS social classes. We have created three ontologies, some inference rules, and an algorithm. Following the descriptions in the two ontologies, we used R2RML⁸ to convert the GBCS and ELSA data sets from their spreadsheet forms to semantic-linked data (transforming the terms and structures in the data sets into their ontology equivalents). The final challenge was to run queries across the data sets. Our proof of concept was to derive GBCS class dispositions for ELSA using inference rules. We tested this using a small sample from GBCS, to see how our methods performed compared to the original class recorded there, and it performed well when comparing our ascribed class to the most likely class generated by the GBCS K-means clustering (Savage et al. 2013). However, when we ran queries across larger volumes of data in ELSA, the process became intractable. Our interpretation was that the fluents approach to time caused this problem:

The Fluents ontology has had a great impact on the project, I'm still convinced that's why the reasoner doesn't work... It created a lot of extra triples⁹ and I'm still convinced that's the reason that the Reasoner kept breaking. (Computer scientist 3, Team meeting transcript, October 15, 2020)

We've got a very semantic Web approach which means we could do lots of reasoning, if only the reasoning scaled! Could we have come up with a simpler approach? Yes, we probably could. Would it allow us to do the same things [given compute capacity]? It wouldn't. (Computer scientist 2, Team meeting transcript, October 15, 2020)

Although it had been designed to be an efficient declarative statement of time, there is no record of the Fluents ontology being used to query across even moderate-sized data sets. While this had seemed to be a declarative and neat solution earlier on, now it was challenging the whole outcome of the experiment. The semantic Web is better suited for some things than others. Complex numerical calculations or modeling temporal data are challenging. As a result, they are often avoided, or dealt with outside of the semantic Web, reducing data interoperability.

Consequently, the *fifth juncture* concerned whether to persevere with inference¹⁰—demonstrating the declarative power of the semantic Web—or move to use algorithms outside of the ontology to provide a specific solution for our project. Given the pressures of time, staffing, and limits of the reasoning technologies, we chose to devise an additional algorithm—deriving class dispositions from capital measures—to achieve our proof of concept. This was more proceduralist and scruffy approach. It did not demonstrate the power of machine reasoning. Although our algorithms are reusable, they are not discoverable in the semantic Web because they require specific execution to generate values and are not encoded in semantic Web schema. They would also have to be rewritten for different data sets and contexts. From a social research perspective, this meant we now had the means to interrogate ELSA with knowledge from GBCS. However, our (unusual) social scientific use of semantic Web technologies is compromised to the extent that it is marginal to the (growing) semantic Web, not completely discoverable by others or not part of the knowledge base. This means social researchers are less able to benefit from the full computational opportunities of the semantic Web.

Discussion and Conclusions

The Web is now a core infrastructure for the modern world. Like any infrastructure, how it is done matters. This paper takes the Web as the focus of critical infrastructural inquiry to explore the politics of data and AI methods in the evolution of the open Web. These questions here are distinct from those about the politics of data and AI in big corporate platforms such as Facebook, Amazon, and Twitter. These platforms are centralized and proprietary, and critical concerns focusing on the use of machine learning methods to extract information and meaning from users' content and activities within the platforms. The emergent semantic Web uses symbolic AI to generate a new layer of standardized tools and protocols for data linkage and inference across multiple, diverse, distributed, and open Web data sources, with a focus on scientific and practical knowledge (Berners-Lee et al. 2001) rather than deriving information on users' activities online. Nonetheless, the potential implications are profound. No less than a new approach to how knowledge is represented, created, and circulated across the open Web.

Our project interrogated the situated accomplishments of the semantic Web (Woolgar and Lezaun 2013, 322) of the semantic Web, in one small instance, exploring the sociomaterial practices of the SW and their

consequences for the knowledge created. This is the first project to explore this empirical detail as far as we are aware. Overall, our project demonstrated two central points. First, while high-level debates about declarative and proceduralist approaches have emphasized two very different approaches to the semantic Web (Halpin and Monin 2016; Poirier 2017), such binary framings were neither accurate nor helpful in explaining how the semantic Web was done in practice. In our experience, creating semantic linked data was not a choice between declarativist and proceduralist approaches or between neat and scruffy. Rather, our experiment shows that these different approaches are contextual and relational. There was a continual pull between the local demands of the project—intellectually and practically—and the wider goal of producing interoperable and reusable artifacts and tools. In the end, specific choices were made to embed local and specific ontological and epistemological decisions in what would appear as a generic set of tools. Second, that the epistemic constraints of the Web demand workarounds for certain kinds of data and querying. Symbolic AI is not well-equipped to deal with quantitative data, so this work was outsourced to algorithms. Also, the computational costs of modeling time made these queries intractable. The significance of these points is that certain kinds of data or queries may be marginalized by the semantic Web or hidden outside the transparent and open structures of the semantic Web. More widely, the data and tools that we created carry these specific ontological and epistemological choices and consequences forward, if and when they are reused and become part of the wider semantic Web infrastructure.

The promotion and publication of semantically linked data on the open Web is growing fast (www.lod-cloud.net, <https://opendatacharter.net>, and www.wikidata.org). By publishing data and underlying ontologies as linked and open, the explicit intention is that others will reuse and repurpose them. Most effectively (for SW proponents) particular ontologies will become the single standard for describing particular domains, both empirical fields and structuring processes such as time or space. Our small study of the knowledge practices involved in creating semantically linked data, and their consequences, highlights the importance of raising how these same questions and processes might shape the semantic Web at scale. The data sets and ontologies published on the open Web have also been produced through concrete knowledge practices, shaped by the priorities and contingencies of the organizations and individuals involved. This means they inherit and bear particular decisions about how to represent contested entities, relations, and fields. Linked data artifacts carry decisions about the meanings

that are “wired into” data sets and will—in principle—extend beyond their origins into any other context of use. These decisions can be very difficult to excavate once they have been materialized into a particular ontology (as we saw with the disappearance of any explicit reference to Bourdieu in our ontology, albeit that the legacy remained implicit and unarticulated in some categories that remained in the ontology). Extending the implications of our findings to other contexts where semantic Web technologies are used, similar knowledge processes will shape those artifacts and shape activities in any other contexts where they are re-purposed. This could be explored further by interrogating the RDF data and underlying ontologies that are published in the linked data cloud as part of the wider commitment to data sharing and interoperability.

Finally, the emergence of a semantic Web also has consequence for the big commercial platforms and our understanding of these. There are two different points to be made here. First, semantically linked data provide a new resource for emerging “infrastructuralized platforms” (Plantin et al. 2018) that have been created on and by the open Web. The linked data cloud offers open data that can be ingested by the platforms which will, in turn, inherit the epistemological and ontological consequences of symbolic AI and the particular assemblages produced. This will extend and reproduce the knowledge practices of particular SW projects/activities inside the platforms, even if these platforms present as walled gardens (Holmes 2013), for example, in the hypothetical case that a major social media platform used Wikidata to construct biographies, it would inherit the inherent classifications of the original ontologies. Researchers have pointed out how platform engineering reaches out into the Web, for example, through the use of widgets (Helmond 2015), which confirms that SW engineering may shape platform activities. This raises further interesting and important questions about the co-constitution of platforms and the Web (Mukherjee 2019).

Second, there is evidence that platforms themselves are using SW methods for their own operations, alongside more fully established machine learning methods (Poirier 2017). However, we know little about how these methods are deployed inside platforms, for example, which data are used or how, what conceptualizations or models are in play. What we can reasonably suggest, based on the findings or our own experiment and the wider platform studies literature, is that these investments in SW technologies will be shaped by questions of local contingency and expected to produce actionable knowledge that can be fed into platforms’ corporate assemblages, shaping how search engines work, how recommendations are made (e.g., on Facebook feeds). In turn, this may instate epistemological and

ontological effects of SW methods into wider questions about fake news, filter bubbles, and discriminatory decision-making tools. Further, the epistemic limits of the semantic Web technologies we have explored in this study (and likely other limits not yet apparent) will shape their use inside large corporate platforms and subsequent consequences. For example, the intractability of temporal reasoning described above may narrow the forms of analysis inside platforms, and consequently what kinds of information appears through these platforms, and therefore what can and can't be known through them. For example, if Google is used to search for historical information and it draws on semantic Web technologies to generate this data, there would be no way of knowing how these data were constructed or modeled, nor the epistemic limits that may be embedded in the methods used (ontologies or forms of inference, for instance).

Our key point is to focus on the sociomaterialities of the open Web as a dynamic infrastructure and pay attention to how these sociomaterialities are embedded in the ongoing politics of data and AI. Attention to the emerging semantic Web raises critical questions about how knowledge is produced online, by whom, and with what consequences that both sit alongside and contribute to more widely debated concerns about platforms, data, and AI. Unpicking the sociomaterialities of these knowledge practices is painstaking and challenging work that demands the integration of expertise in social and computational sciences. As we have shown, the devil is in the detail—understanding knowledge practices demands sociological engagement with technical processes and computational openness. To understand knowledge practices is to question taken-for-granted ways of doing. Even then, we can never expect to unpick every last detail. The key is to generate sufficient knowledge of the kinds of processes and questions that should be asked to ensure that the infrastructures of the open Web are not taken for granted or sink into the background and instead remain the object of critical investigation.


Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Economic and Social Research Council (ES/M0003809/1).

ORCID iDs

Susan Halford  <https://orcid.org/0000-0001-5102-1790>

Mark Weal  <https://orcid.org/0000-0001-6251-8786>

Notes

1. Accessed July 1, 2022. <https://www.statista.com/statistics/617136/digital-population-worldwide/January2022>.
2. The original proposal for the WWW, HTMLized (w3.org), available at <https://www.w3.org/History/1989/proposal.html>.
3. Recording of a talk by Tim Berners-Lee at the First International Conference on the World-Wide Web hosted by the European Organization for Nuclear Research (CERN) in 1994. Accessed August 1, 2022. <https://videos.cern.ch/record/2671957>.
4. We are grateful to the Economic and Social Research Council for supporting this research, grant number ES/M0003809/1.
5. Unlike Great British Class Survey, this assigns individuals to social class as if it's a fact, rather than a likelihood—epistemologically, this is a different way of generating information than latent class analysis, from inductive to deductive.
6. This research received ethical approval from the University of Bristol and the University of Southampton.
7. The term ontology is used in a related but different way in social science, and this distinction had long been used in the team as a marker of the complexities of interdisciplinary research. Here, we describe what “building an ontology” means in the semantic Web. Later, we consider some of the ontological implications of this from a sociological perspective.
8. A mapping language to facilitate conversion from structured data (e.g., csv or database tables) into a linked data format.
9. Triples are the basic building blocks of Resource Description Framework (RDF), the standard SW modeling language.
10. The general inferential capabilities afforded by logic-based ontology languages (as opposed to algorithms) for deriving specific entailments.

References

- Balmer, Andrew S., Jane Calvert, Claire Marris, Susan Molyneux-Hodgson, Emma Frow, Matthew Kearnes, Kate Bulpin, Pablo Schyfter, Adrian Mackenzie, and Paul Martin. 2015. “Taking Roles in Interdisciplinary Collaborations: Reflections on Working in Post-ELSI Spaces in the UK Synthetic Biology Community,” *Science & Technology Studies* 28 (3): 3-25.

- Barnet, Belinda. 2019. "Hypertext before the Web—or, What the Web Could Have Been." In *The SAGE Handbook of Web History*, edited by Brügger Niels and Milligan Ian, 215-26. London, UK: Sage.
- Berners-Lee, Tim. (2000) *Weaving the Web: The Past, Present and Future of the World Wide Web*. London, UK: Texere.
- Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Web Semantic." *Scientific American* 284 (5): 34-43.
- Bijsterveld, Karin, and Swinnen Aagje. 2023. *Interdisciplinarity in the Scholarly Life Cycle*. Palgrave Macmillan Cham.
- Bogost, Ian, and Montfort Nick. 2009. *Platform Studies: Frequently Questioned Answers*. Digital Arts and Culture 2009. Irvine, CA: UC Irvine. Accessed April 25, 2024. <https://escholarship.org/uc/item/01r0k9br>
- Bowker, Geoffrey C., Karen Baker, Florence Millerand, and David Ribes. 2010. "Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment." In *International Handbook of Internet Research*, edited by Hunsinger Jeremy, Lisbeth Klastrup, and Matthew Allen, 97-117. New York: Springer.
- Brügger, Niels, Ian Milligan, Anat Ben-David, Sophie Gebeil, Federico Nanni, Richard Rogers, William J. Turkel, Matthew S. Weber, and Peter Webster. 2019. "Internet Histories and Computational Methods: A "Round-doc" Discussion." *Internet Histories* 3 (3-4): 202-22.
- Ford, Heather, and Mark Graham. 2016. "Semantic Cities: Coded Geopolitics and the Rise of the Semantic Web." In *Code and the City*, edited by Kitchin Rob and Sung-Yueh Perng, 200-14. London, UK: Routledge.
- Gillespie, Tarleton. 2010. "The Politics of "Platforms."" *New Media & Society*, 12 (3): 347-64.
- Gray, Jonathan, Carolin Gerlitz, and Liliana Bounegru. 2018. "Data Infrastructure Literacy." *Big Data and Society*, 5 (2): 2053951718786316.
- Halford, Susan, Catherine Pope, and Mark Weal. 2013. "Digital Futures? Sociological Challenges and Opportunities in the Emergent Semantic Web." *Sociology* 47 (1): 173-89.
- Halford, Susan, Mark Weal, Ramine Tinati, Les Carr, and Catherine Pope. 2018. "Understanding the Production and Circulation of Social Media Data: Towards Methodological Principles and Praxis." *New Media and Society* 20 (9): 3341-58.
- Halpin, Harry, and Alexandre Monnin. 2016. "The Decentralization of Knowledge: How Carnap and Heidegger Influenced the Web." *First Monday* 21 (12).
- Helmond, Anne. 2015. "The Platformization of the Web: Making Web Data Platform Ready." *Social Media + Society* 1 (2): 2056305115603080.
- Holmes, Ryan. 2013. "From Inside Walled Gardens, Social Networks Are Suffocating the Internet as We Know It." *Fast Company*, 9 August. Accessed April 25,

2024. <http://www.fastcompany.com/3015418/frominside-walled-gardens-social-networks-are-suffocating-the-internet-as-we-know-it>
- Jackson, Steven J, Paul N. Edwards, Geoffrey C. Bowker, and Cory P. Knobel. 2007. "Understanding Infrastructure: History, Heuristics and Cyberinfrastructure Policy." *First Monday* 12 (6).
- Mccarthy, Matthew T. 2017. "The Semantic Web and Its Entanglements." *Science, Technology, & Society* 22 (1): 21-37.
- Mukherjee, Rahul. 2019. "Jio Sparks Disruption 2.0: Infrastructural Imaginaries and Platform Ecosystems in 'Digital India.'" *Media, Culture & Society* 41 (2): 175-95.
- Musiani, Francesca. 2015. "Practice, Plurality, Performativity, and Plumbing: Internet Governance Research Meets Science and Technology Studies." *Science, Technology, & Human Values* 40 (2): 272-86.
- Musiani, Francesca, and Valérie Schafer. 2019. "Science and Technology Studies Approaches to Web History." In *The SAGE Handbook of Web History*, edited by Niels Brügger and Ian Milligan, 73-85. London, UK: Sage.
- O'Reilly, Tim. 2005, September 30. "What is Web 2.0." <http://www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-Web-2.0.html>
- Poirier, Lindsay. 2017. "A Turn for the Scruffy: An Ethnographic Study of Semantic Web Architecture." In *Proceedings of the 2017 ACM on Web Science Conference*, Troy, NY, June 25-28, 2017, 359-67. New York: ACM.
- Poirier, Lindsay. 2018. "Making the Web Meaningful: A History of Web Semantics." In *The Sage Handbook of Web History*, edited by Niels Brügger and Ian Milligan, 256-69. London, UK: Sage.
- Plantin, Jean-Christophe, Carl Lagoze, Paul N. Edwards, and Christian Sandvig. 2018. "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook." *New Media & Society* 20 (1): 293-10.
- Plantin, Jean-Christophe, and Aswin Punathambekar. 2019. "Digital Media Infrastructures: Pipes, Platforms, and Politics." *Media, Culture & Society* 41 (2) 163-74.
- Risso, Linda. 2018. *Harvesting Your Soul? Cambridge Analytica and Brexit*. In *Proceedings of the Brexit Means Brexit?*, edited by Christa Jansohn, 75-90. Accessed April 25, 2024. Mainz. http://www.adwmainz.de/fileadmin/user_upload/Brexit-Symposium_Online-Version.pdf#page=75
- Savage, Mike. 2018. *Social Class in the 21st Century*. London, UK: Penguin.
- Savage, Mike, Fiona Devine, Niall Cunningham, Mark Taylor, Yaojun Li, Johs Hjellbrekke, Brigitte Le Roux, Sam Friedman, and Andrew Miles. 2013. "A New Model of Social Class? Findings from the BBC's Great British Class Survey Experiment." *Sociology* 47 (2): 219-50.

- Schubert, Cornelius, and Andreas Kolb. 2019. "Designing Technology, Developing Theory: Toward a Symmetrical Approach" *Science, Technology, & Human Values* 46 (3): 528-54.
- Shirkey, Clay. 2006. "The Semantic Web, Syllogism and Worldview." Clay Shirkey's Writings about the Internet. Accessed April 25, 2024. http://eolo.cps.unizar.es/docencia/doctorado/Articulos/WebSemantica/Semantic%20Web%20-%20Shirky_%20The%20Semantic%20Web,%20Syllogism,%20and%20Worldview.pdf
- Star, Susan Leigh. 1999. "The Ethnography of Infrastructure." *American Behavioral Scientist* 43 (3): 377-91.
- Van Dijck, José, and Thomas Poell. 2013. "Understanding Social Media Logic." *Media and Communication* 1 (1) 2-14.
- Van Heur, Bas, Loet Leydesdorff, and Sally Wyatt. 2013. "Turning to Ontology in STS? Turning to STS through 'Ontology.'" *Social Studies of Science* 43 (3) 341-62.
- Welty, Chris, and Richard Fikes. 2006. "A Reusable Ontology for Fluents in OWL." In *Formal Ontology in Information Systems*, edited by Brandon Bennett and Christiane Fellbaum, 226-36. Amsterdam, the Netherlands: IOS Press.
- Winner, Langdon. 1986. "Do Artefacts Have Politics?" *Daedalus* 109 (1): 121-36.
- Winograd, Terry. 1975. "Frame Representations and the Declarative/Procedural Controversy." In *Representation and Understanding*, edited by Daniel G. Bobrow and Allan Collins, 185-10. New York: Morgan Kaufmann.
- Woolgar, Steve, and Javier Lezaun. 2013. "The Wrong Bin Bag: A Turn to Ontology in Science and Technology Studies?" *Social Studies of Science*, 43 (3): 321-40.

Author Biographies

Susan Halford is a Professor of Sociology. She was a founding Director of the Web Science Institute at the University of Southampton and is currently Co-director of the ESRC Centre for Sociodigital Futures at the University of Bristol. Her research focuses on the politics and practices of digital data, artifacts, and infrastructures, with specific attention to the processes of creating and intervening in emergent sociodigital futures.

Mark Weal is a Professor of Web Science in the Digital Health and Biomedical Engineering Group in Electronics and Computer Science at the University of Southampton. He is an Associate Director of the Web Science Institute and Co-director of the LifeGuide project to develop digital public health interventions. His research interests include the application of semantic Web technologies and information systems, including hypermedia systems and eLearning technologies.

Faranak Hardcastle is a Research Fellow in the CELS-Oxford research group at the Wellcome Centre for Human Genetics. Her research uses insights from Science and Technology Studies and Critical Data and Algorithm Studies to critically engage with the development and application of new healthcare and Web technologies.

Nicholas Gibbins is an Associate Professor in the Web and Internet Science Group in the University of Southampton's School of Electronics and Computer Science. His key areas of research are the infrastructure of the semantic Web and the design and development of knowledge-intensive applications. His teaching at Southampton closely follows his research: semantic Web, hypertext, databases, and AI.

Samantha Pearman-Kanza is a Senior Enterprise Fellow at the University of Southampton. She coordinates the AI 4 Scientific Discovery Network (AI4SD) and the Future Blood Testing Network. Her research applies computer science techniques to the scientific domain through the use of semantic Web technologies and AI. She has worked on interdisciplinary semantic Web projects in different domains, including agriculture, chemistry, and the social sciences.

Catherine Pope is a Professor of Medical Sociology and Senior Research Fellow at Green Templeton College, University of Oxford. She coled the Doctoral Training Centre for Web Science at the University of Southampton. An expert in qualitative and mixed methods for applied health research, she is a key contributor to developing methods for evidence synthesis. She has published empirical, theoretical, and methodological work for clinical, sociological, policy, and practitioner audiences.