ENACTING THE SEMANTIC WEB: ONTOLOGIAL ORDERINGS, NEGOTIATED STANDARDS, AND HUMAN-MACHINE TRANSLATIONS

by

Matthew T. McCarthy

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Sociology

at

The University of Wisconsin-Milwaukee

August 2017

ABSTRACT

ENACTING THE SEMANTIC WEB: ONTOLOGIAL ORDERINGS, NEGOTIATED STANDARDS, AND HUMAN-MACHINE TRANSLATIONS

by

Matthew T. McCarthy

The University of Wisconsin-Milwaukee. 2017
Under the Supervision of Professor Aneesh Aneesh

Artificial intelligence (AI) that is based upon semantic search has become one of the dominant means for accessing information in recent years. This is particularly the case in mobile contexts, as search based AI are embedded in each of the major mobile operating systems. The implications are such that information is becoming less a matter of choosing between different sets of results, and more of a presentation of a single answer, limiting both the availability of and exposure to, alternate sources of information. Thus, it is essential to understand how that information comes to be structured and how deterministic systems like search based AI come to understand the indeterminate worlds they are tasked with interrogating. The semantic web, one of the technologies underpinning these systems, creates machine-readable data from the existing web of text and formalizes those machine-readable understandings in ontologies. This study investigates the ways that those semantic assemblages structure, and thus define, the world. In accordance with assemblage theory, it is necessary to study the interactions between the components that make up such data assemblages. As yet, the social sciences have been slow to systematically investigate data assemblages, the semantic web, and the components of these important socio-technical systems. This study investigates one major ontology,

Schema.org. It uses netnographic methods to study the construction and use of Schema.org to determine how ontological states are declared and how human-machine translations occur in those development and use processes. This study has two main findings that bear on the relevant literature. First, I find that development and use of the ontology is a product of negotiations with technical standards such that ontologists and users must work around, with, and through the affordances and constraints of standards. Second, these groups adopt a pragmatic and generalizable approach to data modeling and semantic markup that determines ontological context in local and global ways. This first finding is significant in that past work has largely focused on how people work around standards' limitations, whereas this shows that practitioners also strategically engage with standards to achieve their aims. Second, the particular approach that these groups use in translating human knowledge to machines, differs from the formalized and positivistic approaches described in past work. At a larger level, this study fills a lacuna in the collective understanding of how data assemblages are constructed and operate.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AFAIK	As Far As I Know
API	Application Program Interface
ASA	American Sociological Association
ASCII	American Standard Code for Information Interchange
BCP 47	Best Current Practice for Language Tags
CEO	Chief Executive Officer
DSM	Diagnostic and Statistical Manual of Mental Disorders
ELI	European Legislation Identifier Ontology
fMRI	Functional Magnetic Resonance Imaging
GEON	The Geosciences Network
GPS	Global Positioning System
HTML	Hypertext Markup Language
IRB	Institutional Review Board
IETF	Internet Engineering Taskforce
IMO	In My Opinion
IRI	Internationalized Resource Identifier
JSON	Javascript Object Notation
JSON-LD	Javascript Object Notation for Linked Data
MDA	Multimodal Discourse Analysis

OWL	Web Ontology Language
RDF	Resource Description Framework
RDFa	Resource Description Framework in Attributes
SDTT	Structured Data Testing Tool
SSDA	Social Science Data Archive
STS	Science and Technology Studies
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WTO	World Trade Organization
WWW	World Wide Web
XML	Extensible Markup Language

ACKNOWLEDGEMENTS

First, I want to thank my advisor, Aneesh, for his unwavering support, confidence, and sound advice. Aneesh provided encouragement and wisdom during my failures and challenged me to do more to avoid becoming complacent with my successes. His impact on my research cannot be overstated. Not only was his wide range of knowledge instrumental in developing my own research agenda, his flexibility allowed me to explore topics that extended beyond the sometimes narrow confines of the discipline. Second, I would like to thank my committee members, Michael Zimmer, Timothy O'Brien, and Gordon Gauchat. Each provided me with invaluable assistance in creating and completing this project, as well as aid in other aspects of my studies. All three have provided different forms of advice and support since I have known them. I would also like to thank other faculty members who have helped me in significant ways throughout my graduate career. Noelle Chesley was a constant source of advice and mentorship in my professionalization and in my pedagogy. Kent Redding and Jennifer Jordan also deserve thanks for their advice and service on my committees over the years.

Perhaps the most important component of my success in graduate school is the support I received outside of the academy. I owe my parents, Christina and Kevin McCarthy an incalculable debt for their financial and emotional support in life, love, and scholarship. I would also like to thank D.J. Wolover, Ken Jackson, Joshua Cerretti, and Theresa Warburton for their friendship and commiseration. Having peers to navigate the ups and downs of graduate life with was tremendously helpful. Lastly, and by no means least, I would like to thank my partner, Megan Thomas, for her unwavering support, confidence, and affection during this time. There is no way that I would have successfully finished without it.

"The computer takes our own superlative power over worlds as the condition of possibility for the creation of worlds. Our intense investment in worlds — our acute fact finding, our scanning and data mining, our spidering and extracting — is the precondition for how worlds are revealed. The promise is not one of revealing something as it is, but in simulating a thing so effectively that "what it is" becomes less and less necessary to speak about, not because it is gone for good, but because we have perfected a language *for* it."

- Alexander R. Galloway: *The Interface Effect*

1. Introduction

More than once over the past year, the webmaster for Schema.org was approached about using the project to help automate political and journalistic fact checking and reducing the impact of fake news. The main concern being that fake news, and/or misleading claims, might be damaging to one's ability to inform oneself and to carry on as an engaged citizen, as incorrect, improper, or incomplete information can lead to incorrect attributions and false conclusions in the populous. Working though this conceptual morass is complicated. Recognizing something as false, requires interpretation on the part of the reader, however, this interpretation is seemingly not happening, or is not happening enough for this issue to gain such prominence. Popular expressions of anxiety have contributed to a collective blame shifting, placing the onus for filtering out fake news onto the owners of the platforms that deliver it. In the United States, companies like Google and Facebook have begun applying a host of technological fixes to the problem, allowing machines to algorithmically filter results. One solution, employed by Google, implicates Schema.org and relies on an infrastructural shift in the way that Web information is structured, represented, and recalled. This shift involves converting news articles and their claims into machine-readable data to indicate to an artificial intelligence (AI) or other computational agent, that a piece of content is analysis, background, opinion, reportage, or review. Furthermore, this conversion can take up individual, or sets of claims made to determine their truthfulness and to reference fact-checking entities whose job it is to review such claims.

How does a machine know what fake is, though? How does an AI know the difference between a true claim and one that has been debunked? How does a computer know if content is a sponsored advertisement, or content that is newsworthy, but sponsored? How can it understand how a fact checked claim relates to the contexts in which the claim was originally used or was reproduced? In sum, on what basis is a computer's understanding of the world built? Heather Ford and Mark Graham (2016) note a different, but similar, problem. They ask what effects the mediation of Web information has on the representation of different places. Specifically, they inquire about this just noted infrastructural shift in information recall and presentation. They investigate the ways that the city of Jerusalem is represented by Wikipedia, Wikidata, and Google, as well as the consequences of the way translations between those sites occur. In this process of translating information between Wikipedia, Wikidata, and Google, much definitional nuance is lost, as both Wikidata and Google obscure Jerusalem's complex historically grounded claims of ownership and importance and over-determine its political status (Ford and Graham 2016).

In this work, they also show that the provenance of important statistical data describing Jerusalem is obscured in different ways across these three domains. They argue that the population statistics, particularly of Jerusalem, reflects a certain political position that obscures the precise nature of the people who are doing the counting, and those who are counted. They show that these important details of provenance are lost in the translations from Wikipedia. Lastly, they also note that a user's agency to make changes and their ability to interrogate alternate information is limited as these translations occur. Where users have the ability to

actively engage with information on Wikipedia, no such agency exists in Wikidata and Google[1]. What causes this loss of nuance, obscured provenance, and reduction in a user's ability to change and interrogate information? How are these two problems of automated fake news filtration and the problems of information retrieval and representation linked?

These stories orient us to one of three major changes in the way that information currently exists on the Web. The change at the center them is one of infrastructure and its relation to data. This first change reflects a new implementation of the means for representing, interpreting, connecting, and retrieving data on the Web. It is an implementation where human interpretation is one more additional level removed from the underlying data, with an additional insertion point for computational agents to mediate information, one where information is actually premediated. This change is one that affects the thing actually being parsed in any search. Before I examine the new structural change introduced above, I will point to the second major change, the way that information on the Web is accessed and retrieved.

While search is still the dominant means through which information is accessed on the Internet, the nature of search has changed, and promises to change further. There is no novelty in claiming that search based information retrieval is changing, as work has extensively covered the effects of Web search, particularly as it pertains to Google and its marked shift from past search forms (Halavais 2009; Introna and Nissenbaum 2000; Viadhyanathan 2012). However,

[1] On this last point, I disagree. This study examines exactly the ways in which information seekers engage with the technologies underpinning Ford and Graham's work. Such development communities are open and participation driven. However, their analysis should give cause to consider an important related issue about both the expertise required to participate in a meaningful way, as well as the practical transparency the processes and technologies involved in semantic search.

there is value is stating the nature of those changes, insofar as those changes coincide with a change in interface. Just as the shift from the page to the screen prompted a movement to rethink information access, the shift from the desktop to the mobile device should prompt another (Galloway 2012).

Consider the following, the interface through which search takes place is now primarily mobile. Between 52% - 71% of all Web traffic is mobile depending on source and country. Over 50% of Google search is now mobile, though Google offers no geo-demographic detail beyond that (Chaffey 2016; Meller and Cohen 2015). One estimate by Hitwise reports that roughly 58% of all Web search is mobile, though they too do not offer more detailed distinctions (Fetto 2016). Moreover, these rates have increased year over year. Perhaps most telling in these regards, is that Google is switching its primary search index from desktop to mobile sites. This last fact has major consequences for what information appears and what does not, as content now faces increased competition for a newly constrained display space. All of this is to say that mobile search is more prominent than desktop search and its interface is newly limiting.

The third major change is in the method of access. It is not just that search and Web access shifted to a mobile interface, but mobile interfaces themselves are seeing a concurrent shift. While search sites still draw traffic, search is now integrated as a fundamental feature of mobile operating systems. Since 2011, these mobiles operating systems – Android, iOS, and Windows Mobile – have had search based artificial intelligence systems embedded into their platforms as an increasingly dominant means through which users interact with their devices. This has the practical consequence of allowing mobile search to occur outside of a Web browser, and instead promotes search through the mediating effects of predictive AI. While

researches do not know the precise nature of the many algorithms that each major search engine deploys, we do know that those search companies now rely on semantic data, rather than “raw” text, for their search results. Search is now mobile, AI driven, and semantic.

As a preliminary concern, why does this matter? Indexical search already filters information, and has done so algorithmically for decades. Adding to the issues covered in past work, I point to at least three main reasons (Halavais 2009; Introna and Nissenbaum 2000; Viadhyanathan 2012). First, as display space decreases, decisions about what to display and what to omit take on a new heightened significance. Second, the presentation of results matters. Rank order of results enrolls users and search companies in a trust relationship where users trust that the search engine is displaying the correct and most relevant information to their query. This has the effect of drawing users’ attention to the first few results, making highly unlikely that they view lower ranked results, with results on later pages even less likely to generate traffic (Jansen and Spink 2003; Jansen et al. 2000; Spink and Jansen 2004). This is particularly salient when search results are viewed on a mobile page, as the first few options are the only options immediately visible. Lastly, we do not know precisely how search algorithms sort and present results or how recent AI systems augment search. Some general knowledge exists about the processes, but exact details elude us, as they rest on a hidden combination of technologies and algorithms whose implementation are often trade secrets and differ depending on the company at hand. That is, they are notoriously black boxed. However, one such technology has not yet shifted completely into the background: the semantic web.

Recently, search engines have been drawing on linked data technologies, creating and enabling what is more conventionally referred to as the semantic web. The semantic web is an

additional layer to the World Wide Web (WWW) where content is given machine-readable context and meaning. Just as Hypertext Markup Language (HTML) provides standards for a describing a logical section of a Web document, the semantic web provides a series of standards for creating meaningful data and content from that Web content (Berners-Lee et al. 2001; Giri 2011; Halford et al. 2012; Legg 2013). This web of linked data converts an unstructured web of textual information to a structured web of semantic data by providing the definition of terms, properties, and the formal statements of relationships, all of which become marked and interlinked by Web developers through recourse to semantic ontologies. For example, a computational agent, such as a search based AI, would understand that *Dune* is a creative work authored by Frank Herbert in 1965 and has the fictional characters, Paul Atreides, Vladimir Harkonen, and Feyd-Rautha, among others. If the Web data was encoded correctly, an agent could start at any point in that chain of relationships and divine the entire set of them, all by dereferencing the parsed content through the ontology.

As search is now semantic and AI driven, it becomes increasingly important to understand how machines arrive at their understandings of the world and the consequence of those processes, even with the benign example above. Despite this importance, we as a public and as scholars are not aware of the ways these technologies are developed, much less how they declare the world. Ford and Graham (2016) provide an excellent example to show some of the practical and theoretical consequences of the move to the semantic web, but only point to a need for further research to understand how those processes occur and how their information subjects are brought into being. So, it is an understanding of those declarations that this study attempts to expose. Much like the laboratory studies from years past, this study

opens Pandora's black box before its contents become cemented and obscured (Knorr-Cetina 1999; Latour 1988; Latour and Woolgar 1986; Mol 2002). Thus, I am led to the following empirical question, how are these semantic data assemblages enacted? Such a question is not easy to answer, as it first involves unearthing the answers to a host of subsidiary questions. First among them, investigates how ontological states are declared in the processes of developing and using the semantic web. This question is primarily theoretical in that it attends to the particular delimiting of meaning space. The second question asks, how are the complexities of a flexible and non-deterministic world represented to a deterministic machine? This question is primarily empirical, as it attends to the various interactions between actual components of a given data assemblage.

These two overarching questions are both questions of ontologies. At first glance, they appear as distinct interpretations of the term, but under closer analysis they converge to describe the same underlying thing. This discussion is not driven by the classical questions of what it is to be, but rather how one becomes. A Deleuzian ontology concerns itself not with essences or the nature of a thing, but with capacities and how a thing is actualized as it enters into webs of relationships (DeLanda 2006, 2011; Deleuze 1964[1995]; Deleuze and Guattari 1987; May 2005). A thing, or multiplicity for Deleuze, has no essential character but that which is active in a given moment, as a part of a given assemblage. Thus, the question of what something is, takes on less significance than how something becomes, as it requires an understanding of the processes of connection. His approach to ontology rests on the assumption that the ontological is unstable and cannot ever be known, only modeled through its integration in an assemblage under the particular expressive logic of that assemblage. As

such, we can equate the contexts of activation and integration with the ontological state of the thing. So, the previous question on becoming can be reframed as a question of how the assemblage is enacted, or called into being (Mol 2002; Woolgar and Lezaun 2013). Accordingly, ontology is not fixed and it demands an understanding of its settlement, particularly as ontologies in this paradigm are multiple (Deleuze 1964[1995]; Hoffman 2015; May 2005; Mol 2002; Woolgar and Lezaun 2013). In simpler and more directly relevant terms, how is ontology declared by ontologies? The trick, as Deleuze would argue, is uncovering those enrollments and activations in assemblages. This is the task that I will turn to, shortly.

The answer to my questions requires an understanding of the data assemblage surrounding the semantic web. To understand a data assemblage, one needs to take stock of its various components and examine the ways in which those components interact to continually produce the assemblage. Broadly speaking, the semantic web is too large and dispersed a thing for a focused empirical study, as it, like all assemblages, is composed of many other assemblages. In this case, that means many other ontologies, users, consumers, technologies, etc. In an effort to set empirical boundaries and to focus this study, I engage with one such assemblage, Schema.org. As only one component of the larger semantic web, Schema.org is an ontology development project spearheaded by Google, Microsoft, Yahoo, and Yandex that attempts to create semantic data for the World Wide Web. Currently, it is among the largest and most used semantic web ontologies, covering approximately 10% of existing Web content, with major users including the Google, Microsoft, Yahoo, and Yandex search based ecosystems, in addition to a substantial number of other large Web domains (Guha 2014; Nevile 2014).

This study relies on the in-depth case study of the users, developers, vocabulary, markup standards, and digital artifacts directly associated with the Schema.org project. As both the objects of investigation and their impact on my research questions are diverse, multiple methodological approaches drawn from recent work on digital technologies are employed. I deploy netnographic, or digital ethnographic, methods across a number of digital sites that are central to Schema.org's development. Netnographic practice allows for a situated analysis of community engagement available in online forums, providing a set of observational data similar to, but more reliable than field notes and artifacts from traditional ethnographic methods (Kozinets 2010; Springer 2015). Additionally, I use an object based analysis, similar to multimodal discourse analysis, to interrogate the markup standards used by the project. This approach allowed me to identify and interrogate the ways that discursive patterns help to shape affordances and prohibitions, the ways of being and the constraints that might be placed on being, the ways of doing and the constraints placed on practice, and the ways of representing and the constraints on possible claims making through markup standards (Introna 2005, 2014; Introna and Hayes 2011; Kress and van Leewuen 2006; Lupton 2014; O'Halloran 2011).

The remainder of this work unfolds in the following way: first, in chapter two I provide a detailed explanation of how the semantic web works in both theory and practice. In addition to explaining its operation, I discuss the limited amount of work directly pertaining to the semantic web. This work, like Ford and Graham's (2016) study, offers us a set of critiques and potential consequences in implementing a semantic web. However, it is not focused on the processes of its development or use, and makes no reference to the various standards involved.

The most notable of its consequences include the interrelated processes of ontological over-determination and ontic occlusion (Ginsberg 2008; Knobel 2010). This section argues that ontologies, as taxonomic orderings of the world, are fully determinate systems, and what can be known and/or represented in them is bound completely by the terms set out in those ontologies. That is, they permit a predetermined range of claims that can be made against them. This is inadequate for representing the indeterminacy of natural language and the environment of meaning in which it exists.

Chapter three establishes my theoretical assumptions and engages with the extant literature relevant to data assemblages and their components. I expand on recent attempts to study data assemblages by situating my research within an assemblage theoretical perspective (Kitchin 2014a, 2014b; Kitchin and Lauriault 2014; Kitchin et al. 2015a, 2015b). An assemblage based approach to my main questions requires resolving a number of smaller questions relevant to literatures describing and explaining the various components of data assemblages. The small, but growing, work on data assemblages encourages researchers to take these various components as interrelated features of the larger assemblage, studying not what the components do in isolation, but how they interact to produce the assemblage of which they are a part. Among the ones drawn on in this chapter are, literatures on standards, infrastructures, classification, and communities of practice. The existing work on standards argues that standards play a central role in affording and constraining action and interpretation vis-à-vis technologies, and in so doing, they help to bridge diverse communities of practice and infrastructures as mechanisms for interoperability. However, the ways in which these communities negotiate these limitations is under-studied. Relatedly, classification systems play

a similar role, enabling the storage and retrieval of information. Problematically, they can render certain beings invisible when they are neglected by the system. Thus, what is included and excluded, as well as how those decisions are made, take on a central importance. Sociological research in these areas, particularly that which engages with science and technology studies, situates these classification systems, infrastructures, and standards in communities of practice that bring to bear a diverse set of knowledges, assumptions, and techniques in their interactions with these components. So, this literature prompts consideration of questions about how these communities of practice engage with, and negotiate, the affordances and constraints of not only the standards involved, but also the limitations of classification, and techniques of representation presented by those other components.

Chapter four provides a detailed description of how one goes about studying hidden infrastructure-like systems such as the semantic web. In this case and methodology chapter, I provide further detail about Schema.org as a specific implementation of the semantic web. In this chapter, I describe the case that guides my research. This part includes a description of Schema.org, the major institutions attached to the project, the communities of practice that are most closely involved in its development and its use, the vocabulary that constitutes the project, and the primary markup standards implicated in its creation and use. The next segment in this chapter describes the primary sites where my research occurred. This included Schema.org's GitHub repository, the World Wide Web Consortium (W3C) message boards, and sources of documentation for the three markup standards supported by Schema.org. Additionally, this chapter provides a detailed explanation of both the data collection process

and the preparation for analysis. In this section, I also provide a detailed outline of the two main methodological techniques used in this study as alluded to above. Lastly, I review the limitations of my methodology and include a statement of researcher's standpoint and bias that such a study necessarily brings about. I also include a description of the ways that my research changed as the investigation unfolded.

Answering my larger questions necessitates that I first attend to smaller questions. Specifically, I need to uncover the ways in which the various components of the Schema.org assemblage interact. Two central components are the communities of practice and the markup standards that enable Schema.org to function as a semantic web ontology. Chapter five explains the three markup standards supported by Schema.org and the way that they condition its development and use. This chapter extends recent work that both locates standards at the level of professional practice, and that investigates how blockages are negotiated and how the restrictions created by standardization are relaxed while still permitting the interoperability they are designed to enable. I show how the markup standards used with Schema.org are deployed across the two primary fields of practice implicated in the project. In line with existing work on standards and classification, I show the ways in which they set certain affordances and constraints on the development and use of the Schema.org ontology, and thus the affordances and constraints placed upon the enactment of the data assemblage. However, I also show the ways that these standards work against the thing they are supposed to enable, and how the communities of practice negotiate those restrictions, adding to recent work on this front (Halpin 2016). Furthermore, I detail some of the unintended consequences of those

negotiations and how they can serve to over-determine and occlude certain understandings of the world.

As previously noted, the particular form that any given assemblage takes is dependent on the interactions between the assemblage's components. Thus, following from chapter five, the study of semantic web ontologies must necessarily engage with the questions relating to processes of creation and the interactions between other components. In chapter six, I am concerned with the specific apparatuses that include, systems of thought, practices, and communities (Kitchin 2014b:25). Within these apparatuses are the conceptual models, rationalities, techniques, conventions, and the communities involved in applying those systems of thinking and forms of practice. As such, this chapter examines the interactions between those various elements of the Schema.org data assemblage. In this section I concentrate on the two main communities of practice relevant to Schema.org, the ontologists charged with the vocabulary's creation and maintenance, and the Web developers who deploy the vocabulary to convert their content into semantic data. Through this, I reveal the ways that Schema.org's ontologists adapt the Resource Description Framework (RDF) data model to their specific needs. Additionally, I examine their underlying philosophical approach to modeling semantic data, the problems they encounter in doing so, and the consequences of the ways in which they negotiate those problems. I argue here that these approaches have the consequences of creating indeterminacy surrounding a trade-off of semantic precision and coverage that cannot be resolved. Furthermore, their approach creates a certain path dependency that helps contribute to the over-determination and occlusion mentioned above.

Finally, in chapter seven, I draw together the various threads developed throughout this study, both to answer my motivating questions, as well as to outline my contribution to the emerging literature on data studies and the semantic web. Here I argue that the combination of negotiating the affordances and constraints of markup standards and the particular approach to modeling semantic data, leads to a premediated Web environment (Aneesh 2017). The representation of the world, and thus the version of it that becomes available, is being read behind the scenes and being presented as already digested information. Returning to my initial vignette and to Ford and Graham (2016), I expose and examine the processes and technologies by which a computer can know the nature of a news source and claim, as well as how a city like Jerusalem can be over-determined as simply the capital city of Israel, as opposed to a city of a rich, but contested linage. Not only is this study a problematization of the means for representing, but it is also a new provocation to the sociological study of information access. This is not pass normative judgment on the semantic web or the way that its various practitioners implement it, as they themselves recognize the imperfect nature of their work. Instead, it is recognition that the information environment is being encoded, and that encoding is rapidly being settled outside of visible Internet space.

2. What is the Semantic Web?

This chapter serves as an orientation to the semantic web and the data model underpinning it. As such, it begins to unravel the puzzle of its deployment and ossification. It charts a description of the semantic web and outlines its essential differences from the World Wide Web. Through a series of empirical examples, it provides an initial understanding of how a machine can come to know something, how the system that permits that knowing ontologically orders the world, and offers glimpses of potential consequences of such orderings.

Initially a proposal by Sir Tim Berners-Lee, the inventor of the World Wide Web, the semantic web is an additional layer sitting on top of the existing Web where content is given machine-readable context and meaning (Berners-Lee et al. 2001; Giri 2011; Halford et al. 2012; Legg 2013). On the World Wide Web, information is essentially "raw" text, i.e., what we experience is a web of textual documents that have been encoded for display and use through a set of website development standards such as HTML and Extensible Markup Language (XML). The meaning derived from this text is the sole result of processes of human interpretation. While those sets of standards applied to the web of text are often machine-readable, they supply no semantic content. Search algorithms and other computational agents are only able to crawl the textual documents for simple instances of identified keywords of criteria, i.e. there is nothing that identifies Barack Obama as the 44th President of the United States apart from the co-occurrence of those discrete terms within a set of texts, and any attribution of the presidency to Barack Obama is done by the user themselves based on their knowledge of U.S. politics. So, while "Barack Obama" and "44th President of the United States" may appear in the same text, the link between the two is never made by the computational agent. A Web search

for “Who is the 44th president of the United States?” would search for each of those terms and present information in the form of a series of hyperlinks to the user via some relevancy algorithm. Any determination of fact or interpretation of underlying meaning is left to the user, despite the algorithm’s likelihood of presenting one set of results over another. That is, the determination that Donald Trump is the current president, and not Barack Obama, is never made by the computational agent in this case.

The current Web environment is much different than the one just described, however. Search algorithms, and other capable computational agents, engage in semantic search. These agents parse an additional layer of site markup that converts “raw” text into semantic data. The technology that makes this “raw” textual information machine-readable – semantic data – relies on providing a definition of terms, properties, types, and a formal statement of relationships, all of which become marked and interlinked by Web developers. There are a number of requirements for this to occur. First, there needs to be a schema, or set of schemas, for describing types and how they relate to other things as well as the set of properties they have and inherit. Next, there needs to be a semantic structure to determine how those properties connect things to one another and to specific instances. Instances, things, and their properties all need to be uniquely identifiable to a machine through the use of International Resource Identifiers (IRIs)[2]. Lastly, the collection of statements needs to be represented in a

[2] IRIs are a more general version of Universal Resource Identifiers (URIs) that draw on a Unicode typeset and allow for international characters, where URIs draw on ASCII which do not. Both IRI and URI are a more general type of identifier than something like a Universal Resource Locator (URL), which only points to a resource’s location and not its name.

formal set of relationships, called an ontology or vocabulary[3]. These ontologies define the rules of representation and establish all permissible relationships, property attachments, and instance values. That is, they construct the very existence of the world to a machine.

2.1 How Does It Order the World?

Crucial to understanding the development of these ontologies is the type of data model used in their creation. The semantic web relies on a graph theoretic model called the Resource Description Framework (RDF). The RDF conceptual model uses nodes to model subjects and objects and edges or arcs to model properties and relations. Each of which is identified specifically in the ontology

Figure 1. Graph Model of Presidential Succession in Terms

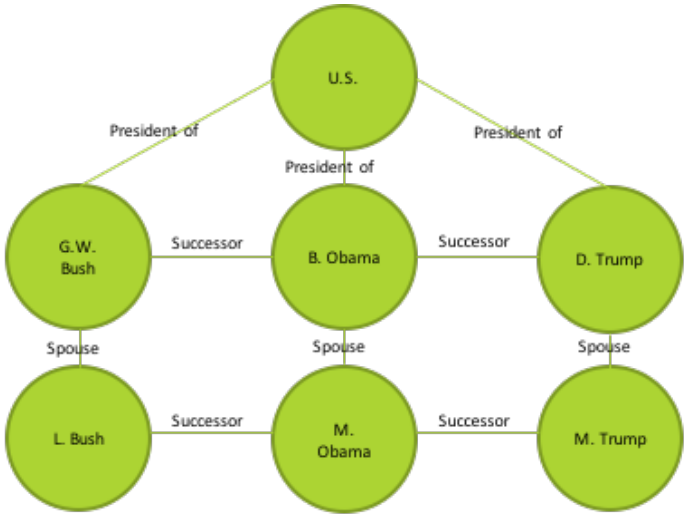


Figure 1. provides an example. This example shows a graph of recent presidential succession

[3] Ontology and vocabulary can be, and are often, used interchangeably. In this study, I do use the two terms interchangeably for the purposes of sentence structure and readability.

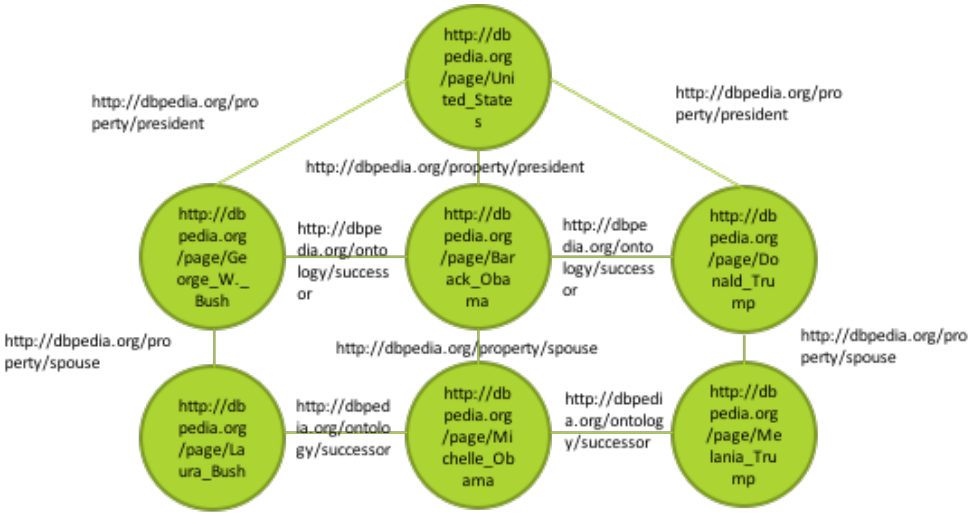
and the set of spousal relationships. Within Figure 1., each node is a subject and/or object and each edge or arc is a predicate or property[4]. An agent parsing this graph could begin at any point and divine information about any other point in the graph. For example, one could ask "Who is George W. Bush's successor's spouse?" and an agent could parse the graph and tell you that it is "Michelle Obama". Moreover, this very simple graph could be much more complex. In this current case, the graph actually would include the entire expanded set of presidents, spouses, presidential sitting dates, and where each identified president sits in the line of presidential succession. A graph of this sort would also include an entire set of additional subjects and relationships. If asked the name of the current President of the United States, the agent could look at this graph and tell you that it was Donald Trump[5]. If drawing on the expanded graph, you could also ask an agent "Who is Laura Bush's spouse, when was he the President, and who did he succeed?". The agent would parse all the relevant property and subject positions in the graph to provide "George W. Bush, 2000-2008, and William J. Clinton". This principle functionality is one of the technologies that underpin artificial intelligence systems like Google Assistant, Siri, Cortana, Alexa, and Hound among others. It is so pervasive

[4] The designation of subject or object or the determination of edge or arc depends on the way that a relationship is framed and the ability for a relationship to be interpreted in reverse. For example, Donald Trump is the *successor* of Barack Obama, in this example. This reads from right to left as Donald Trump as the subject, Barack Obama as the object, and *successor* as a directed arc. We could simply reframe the statement as Barack Obama *hassuccessor* Donald Trump. Here Obama is the subject and Trump is the object, with *hassuccessor* being a directed predicate. Modeling the inverse in graph theory is simple, but the actual application of inverse properties relies on a markup standards supporting reverse parsing and/or the presence of specified inverse properties in the ontology. In this case, if the Dbpedia vocabulary and the chosen markup both permitted inverse interpretation, the arc would be an undirected edge.

[5] This is exactly the approach Google's Knowledge Graph takes.

that it is also embedded in smaller systems like Gmail, Microsoft Outlook, and various calendar applications, among others.

Figure 2. Graph Model of Presidential Succession in IRIs



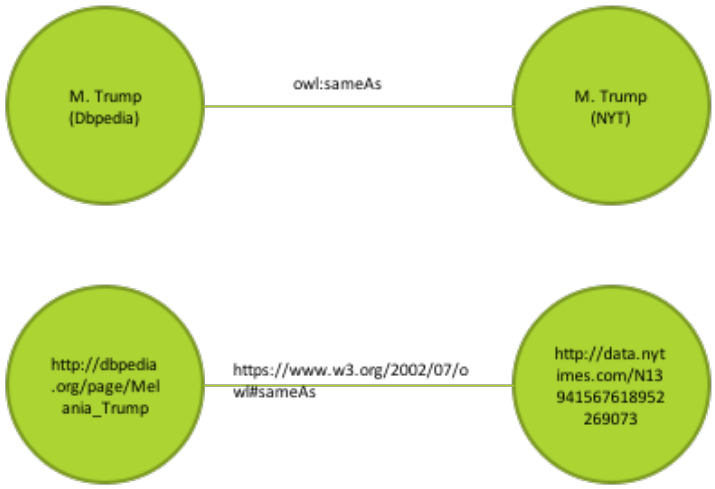
However, Figure 1. is not how a computational agent views semantic data. These are terms that ultimately have no meaning to an agent. They are just literal values. In other words, this graph has no machine-readable context. Figure 2. shows the same graph as an agent would see it[6]. Here we are identifying each node and edge with a unique IRI. In this particular case, each IRI points to a specific location in the DBpedia[7] database. We can ask the same questions as before, but now the agent knows where and what to look at for each node and edge. We have provided the agent and the graph the context from DBpedia. Returning to the original

[6] Agents do not see graphs in this same visual sense, they parse markup that is embedded into webpages, emails, application program interfaces (APIs), etc. Showing the graph form is simply for clarification.

[7] DBpedia is a semantic version of Wikipedia. It is both an ontology and a database, which is the collection of applied semantic content to a domain or domains.

query, the agent would now know who Donald Trump was and what the President of the United States is and how the two instances relate to one another. Furthermore, by providing the context from DBpedia, the agent can abstract the entire context that DBpedia provides. Meaning, it can see that Melania Trump has one son, and know who he is. In this actual case, it would also know that Melania was born in the former Yugoslavia (now Slovenia), what the capital city was, and that she was born under Tito's regime[8]. This is all to say that by providing the necessary context, any set of properties or objects that are connected to Melania Trump to any degree, can be understood by an agent.

Figure 3. Graph Model Connection Multiple Databases



[8] The relationship between the former Yugoslavia and the current Slovenia is over determined and occluded in DBpedia. In Melania Trump's page, there is no machine-readable designation that permits an agent to understand that relationship. Furthermore, the capital city in this case is Belgrade, not Ljubljana, the current capital of the country. This highlights a central ontological problem with the semantic web.

This specific use of the RDF model is one application of the semantic web; however, it only represents a closed system. All of the subjects, objects, and predicates in this graph have only referenced DBpedia, and so a computational agent can only know what is contained in that ontology/database. Much of the development of the semantic web, as well as its initial imagining, rests on the principles of linked data. Linked data contextualizes those instances and properties, by supplying more information on that data, providing a way for machines to more easily understand information by both traversing class hierarchy paths and linking of bits of meaningful data across domains and uses (Berners-Lee et al. 2001; Coyle 2008; Giri 2011; Halford et al. 2012). As it actually happens, DBpedia is a linked database. Continuing with the previous example, an agent only knows Melania Trump to the extent that DBpedia explains her and her set of relationships. Furthermore, an agent only knows that specific version of Melania Trump, not the many other and potentially different representations of her. That is, it cannot understand that the Melania Trump of DBpedia is the same Melania Trump of a *New York Times* article or an *Us Weekly* article. Linked data allows us to tell the agent that the various instances of Melania Trump are referring to the same person. The semantic web community developed a technique for marking this type of equivalency. The Web Ontology Language (OWL) has a property – *owl:sameAs*[9] – which publishers can use to tell a parser that one instance is equivalent across actual use cases. Figure 3. shows a graph theoretic model of this example[10]. In actual usage DBpedia does use *owl:sameAs* to link its database with the *New York*

[9] Italics are used to represent language statements, class, types, and properties.
[10] Figure 3. creates a third dimension to the graph model where each node has additional depth in their sets of relationships. This new relationship has been separated from Figures 1. and 2. for ease of interpretation.

Times. This equivalency means that the whole set of relationships in which Melania Trump is embedded in DBpedia is applicable to the Melania Trump instances found in the *New York Times*. Furthermore, it means that the set of relationships in the *New York Times* are applied to the set in DBpedia as well.

The interlinking does not stop between two ontologies, however. In this limited example, the Melania Trump entry in DBpedia links not only to the *New York Times* but also to Wikidata, marking exact equivalences between all three[11]. Practically, this means that all the sets of relationships that each ontology carries for those applied to Melania Trump are applicable across all vocabularies and use cases that are marked up in this way. The underlying idea here is that ontologies do not need to be redundant. They can simply link to other ontologies that already cover the content domains that they need, allowing ontology developers to focus on just their area of expertise or interest. However, in actual practice, ontologies are often redundant, built differently, with different approaches to the application of the RDF model, and at different levels of complexity. That there are no universal ontologies, fully agreed upon standards, and different philosophies regarding ontology depth and breadth means that there will be practical issues in markup and epistemological issues in representation. The problem of drawing on a number of different ontologies, or ontologies in general, is that they can sometimes serve to occlude differences, and make equivalency statements between instances that may or may not be valid (Halpin et al. 2010; Poirier 2015). This problem is exacerbated when two or more linked ontologies treat an instance differently.

[11] This interlinking across ontologies can be used to apply any number of ontologies to an instance.

2.2 What Are Its Consequences?

One often noted issue among the semantic web community stems directly from the use of the *owl:sameAs* statement. Many developers and ontologists are reluctant to use *owl:sameAs* because it has the necessary consequence that any statements about one instance are true about the others. While the statement is helpful in delineating equivalencies across Web domains when they are not textually referred to in the same ways, it obscures real differences in aboutness, and identity (Halpin et al. 2010). For instance, Poirier (2015) notes that in DBpedia, content about Caitlyn Jenner is *owl:sameAs* content about Bruce Jenner despite the very important and relevant identity transformations and political problems that exist in such a statement. The linked nature of the semantic web makes it so that any link to existing data sets with *owl:sameAs* makes a statement about that instance applicable to both, independent of the truth of the statement. This presents clear issues relating not only to accurate representation but also classification. DBpedia makes a similar false equivalence between Yugoslavia and Slovenia in the example above.

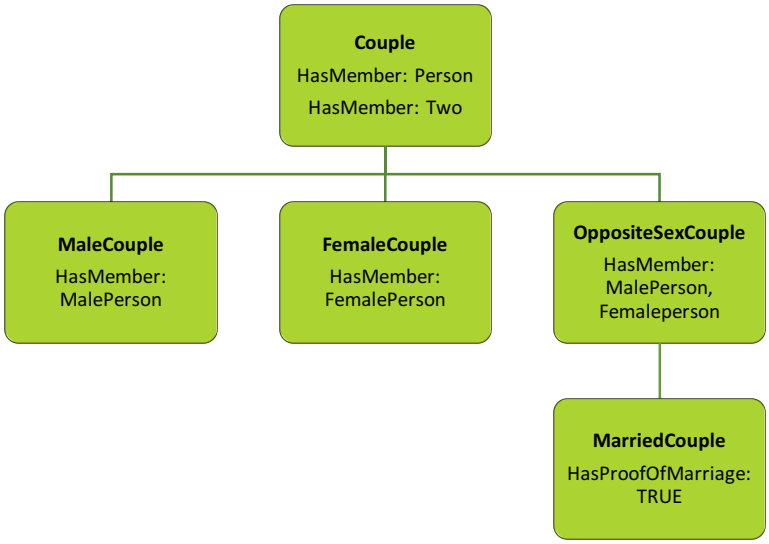
A second issue arises when there are contextual differences. Using *owl:sameAs* to link instances may not be appropriate across use case contexts. Making such an equivalency erases the distinctions that contexts add to identity and definition. For example, the identity a developer is trying to cultivate for Bill Gates the father is not the same as Bill Gates the philanthropist, or Bill Gates the former Chief Executive Officer (CEO) of Microsoft. Here an IRI does mark the same referent and the properties describing referent hold across contexts, but it isn't necessarily appropriate to re-use an IRI when it is out of context. In that case, the properties linked to that IRI do not matter, as the ontology itself works to recreate context. This

point will be discussed in more detail in Chapter six. A third issue is the confusion of identity, representation, and properties of things. Properties of things, or that describe things, are not the same as representations or identities of those things, though equivalency statements are often made between the properties and the thing itself. Furthermore, representation is not the same as identity. URIs – pictures, email addresses, social security numbers, etc. – can stand in as signifiers of a person, referring to a representation of a person, but are not reducible to the person itself and in no way addresses the complexities of their identity. The question of representation is a very real problem confronting the semantic web. However, this question of representation has long been debated in philosophy and will not be the focus of this work. The aim here is to move past the intractable philosophical problems of possibility to the empirical problem of actual practice.

Another major criticism of semantic web technologies comes from their treatment of ontological indeterminacy. Ontological indeterminacy is "the inescapable fact that two or more incompatible conceptual systems can often be applied to a domain of interest with equal empirical adequacy." (Ginsberg 2008:1). This can occur when two communities of practice approach the same concept differently and hold different meanings for it. These terminological boundary objects lose their indeterminacy when coded into ontologies. Ontologies, as taxonomic orderings of the world, are fully determinate classification and representation systems. What can be known and/or represented in them is bound completely by the terms set out in the various ontologies that are drawn on. That is, it permits a predetermined range of claims that can be made against it. Natural language, on the other hand, is not fully determinate. Differences can occur in actual use, and disagreements may arise about the

applicability to particular contexts (Ginsberg 2008; Waller 2016). The translation processes involved in converting the indeterminacy found in natural language to the fully determinate language of the semantic web results in a state of ontological over-determination. I adopt Ginsberg's (2008:7) definition that ontological over-determination applies to "situations in which the use of a formally defined [semantic web] term ipso facto commits one to accepting certain implications or consequences that one could refrain from accepting in natural language while still using the term in the same way".

Figure 4. Marriage Defined in Opposite Sex Ontology

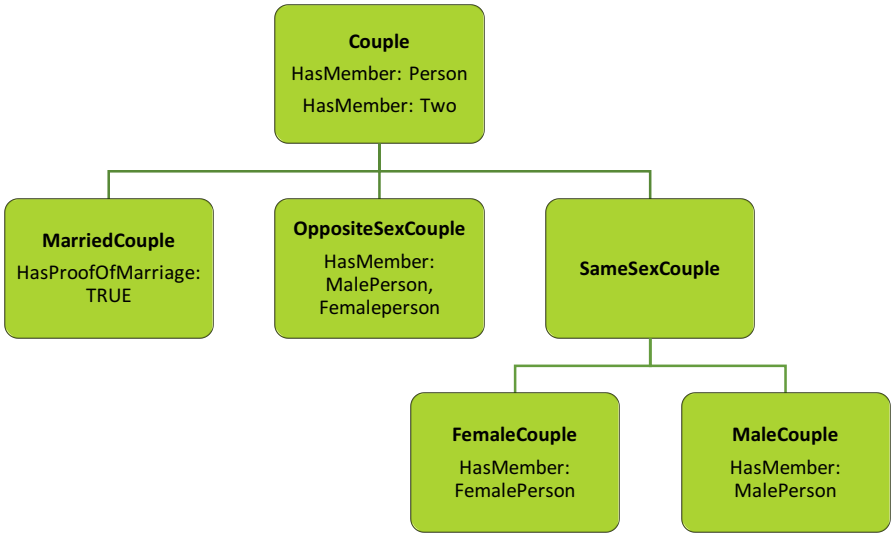


Source: Ginsberg (2008:8)

Ginsberg (2008) shows the practical effects of ontological over-determination as it applies to two states and their acceptance/treatments of gay marriage. To say that homosexual couples can marry is an ontologically indeterminate statement. At the time of the article's writing (2008) gay marriage was only legal in two states, but its very legality in those states

means that the statement is both empirically true and false. The author compares two ontological descriptions of couples and the ability to marry. In one ontology (Figure 4.) the class "couple" has subtypes "male couple", "female couple", and "opposite-sex couple". The subtype "opposite-sex couple" has its own subtype "married couple" that is not a shared subtype of the other two types of couples. This constructs the knowable world as one where marriage only exists as a possibility for heterosexual couples. In programming logic, a decision is made based on the evaluation of presence. Presence indicates the truth of a statement, where absence marks falsity. Here, however, the ability to make a claim of marriage extends beyond simply an evaluation of a binary true/false input and instead forecloses on any alternate possibilities by eliminating the need for the binary decision rule. The realm of knowledge here is prealgocratic, meaning that what can be represented and known is decided before an AI or computational agent can even act. In this way, meaning spaces are premediated (Aneesh 2017).

Figure 5. Marriage Defined in Ontology Regardless of Sex



Source: Ginsberg (2008:8)

In the second ontology (Figure 5.), the class “couple” has the types “same sex couple”, which itself contains “male couple” and “female couple”, “opposite-sex couple”, and “married couple”. In this second ontology, marriage is independent of the couple’s sex determined makeup. The problem arises when a married homosexual couple enters the state that is using the first ontology. In this ontology, marriage is ontologically over-determined. Despite being legally married, a computational agent in the state where gay marriage is illegal has no ability to comprehend the couple’s marriage. The two ontologies are incompatible but not in disagreement because, as Ginsberg (2008:10) argues, they “cannot be referring to the same concept”. To an AI in the illegal state, gay marriage is not illegal, it simply does not exist. This would result in the complete erasure of the married couple from any decision rule that used marriage as a factor in that state, despite its legality in the state where the couple is a citizen. This second ontology, as are all ontologies, is prealgocratic, only it is more flexible and expressive in delimiting the boundaries it has constructed.

This example brings up a related, but distinct issue with ontologies, that of ontic occlusion. While not used in the study of the semantic web, Knobel’s (2010:3) definition is helpful here. He defines ontic occlusion as a:

> mechanism by which representational differences exert control over discourse. That is to say, one representation of an idea, situation, or event can take precedence and occlude, or block, another representation. Thus, the elements of the occluded representation do not enter into the discourse and are left without legitimate role in shaping the narrative.

Returning to Ginsberg’s (2008) gay marriage example, in both ontologies marriage is a sub type of couple, the only difference is based on the sex and implied sexuality of the couple involved. This makes a number of statements about the world and the nature of sex, sexuality, and

coupling. First, it creates a world where polygamy does not exist. While this might make sense in many places like the United States, where polygamy is criminalized, it does not make sense when considering the many countries where it is legal or recognized. Furthermore, even in places where polygamy is criminalized, it still occludes the actual fact of its existence. Second, it reinforces a binary view of sex and sexuality while reducing sexuality to a derivative of sex. It does not begin to allow for a more complex and nuanced understanding of sex and sexuality that might be more closely aligned with actual individuals involved. Here, the choice to represent marriage in the couple form, between two men, two women, or a man and women, is an act of ontic occlusion, rendering alternate states of being invisible. It thus becomes essential to critically examine the process by which this pre-data text becomes semantic linked data. To do so requires that we also examine the data assemblages that organize, and provide the means for sharing and consuming that data.

Thus far, little work has engaged with this topic. While scholarly interest in the semantic web has been growing in recent years, there are no systematic and empirical studies of its development or the conditions of its use. Ford and Graham (2016) have convincingly shown some of the consequences for our understanding of place and space in semantic representations. They argue that semantic content obscures data provenance, topical nuance, and the ability for users to interrogate content as information seeking agents. Waller (2016) correctly argues that the semantic web is seeing a consolidation whereby a few major ontologies are coming to represent the whole of the semantic content on the Web. Like Ford and Graham's (2016) work, the major sociological implication here is that people are being placed in a new set of trust relationships with information providers, particularly given the

recent shifts to mobile and semantically driven AI. The unstated consequence of their work is that semantic web ontologies ontologically order the world for machine interpretation (McCarthy 2017). Problematically, these authors engage only with the interpretive capacity of individuals at the front end of these relationships. I echo their concerns and in principle, I agree with their attributed causes, but their work neglects any discussion of how those causal mechanisms work. Specifically, they do not provide any insight into how these ontologies are developed or how they are actually deployed, let alone the interactions various additional components contribute to these processes. This is the task undertaken here.

3. Semantic Assemblages: Their Ontologies and Their Components

The semantic web is a data assemblage. It creates a specific type of data environment for machine readability. In so doing it enacts a set of specific states of being with each ontological state dependent on the boundaries set by the relevant semantic ontology. To understand how ontological states are declared by these ontologies we need to examine how the world is translated in a way that in amenable to machines. Thus far, literature is lacking in this area. While work on translating tacit knowledge and expertise is helpful, the question extends beyond the simple translation of knowledge between experts in a field. As semantic web ontologies are themselves assemblages composed of various systems of thought, standards, infrastructures, practices, and understandings of data, the literatures surrounding each of those components can help to situate an understanding of this translation process and thus an understanding of how semantic data assemblages are created and deployed to construct worlds.

At the outset, I should state that this study is not making the argument that representation is impossible. That argument has been debated by philosophers of language and science for some time. I argue from the position which assumes that the problem of representation is intractable. Instead, this is a study of interactions, of process, and of making, or as recent work argues, empirical ontology (Coopmans et al. 2014; Sismondo 2015; Woolgar and Lezuan 2013, 2015). In what follows I outline an assemblage based theory of ontology that argues ontological states are dependent on their activations in assemblages. Rather than engaging with the nature of being, it focuses its attention on those empirically activated existences. Next, I examine the literature on data and data assemblages showing that data is a

specific effect of the interactions between components of the data assemblage and that the contexts of those interactions remain under studied.

Following that will be a discussion of work that examines knowledge representation specifically as it pertains to knowledge engineering in expert systems and information infrastructures. While this work details the production of knowledge and the transfer of expertise between humans, as well as the interpretation of machine output by humans, it does not attend to the translations from human to machine and the resulting machine to machine communication that ontologies imply. Since this problem is not only just a process of translation and explication, but also a problem of classification, the next section will examine the ways that classification systems and standards serve as one way of creating terminological states. Classification systems implicate standards in ways that are often imperfect. I extend recent work on this matter arguing that these rough interactions with classification systems and standards require that practitioners negotiate with them in multiple ways. Lastly, I return to the discussion of ontology as it pertains to the empirical enactment of ontologies in practice.

3.1 An Assemblage Theory of Ontology

Derrida argues that the problems with ontology begin at the terminological level (Derrida 1967[2016]). Any attempt to articulate a manner of essential ontology only ever pushes that essence away, as those terms only serve to implicate the very things they are trying to obscure or differentiate from, thus subverting the effort to make meaning inhere within an object. Words are simply too slippery to refer to anything in such a positive way. Views such as this reflects an abandonment of ontology all together, as attempts to establish some sort of

fixity belie the impossibility of such projects. More directly, such deferments cannot be modeled precisely, as terms cling to their overly determined subjects or objects and the inherent meanings that their connections produce.

A similar problem exists at the level of relationships. Two things cannot be said to simply relate in such and such a way, as that just relocates the identification of two things to their relationship, reducing difference to positive identity. In this sense "things are different because they fall into different categories. They are different precisely in not being identical or the same." (May 2005:77). While Deleuze focuses on the relations of identity, analogy, opposition, and resemblance, for the semantic web we might focus on *type*, *sameAs*, *action*, or *category*. They each require that we have not only the understanding of the terms of relating, but the stability of the terms that are to be related. That is, our predicate determines the relationship between our subject and our object where the subject is always the subject, and the object always the object (Deleuze 1968).

While some distance themselves from the ontological project for reasons relating to its impossibility, others embrace the ontological project from a different perspective. In the continental tradition, thinkers have developed an ontology that attempts at an explanation to account for what there is. Deleuze relates this approach to the quest for finding a solution to a question. The problems of being, identity, and definition are problems that look for a particular answer, they expect an answer of certain definite type which renders the problem posed in the question obsolete. May (2005) calls this approach an ontology of discovery. Deleuzian ontology, however, is an ontology of creation. Rather than declaring these problems of definition solved, we accept that a definition is but one of many within a field of possibilities, whose limit is not

knowable (Deleuze 1968; May 2005). This is creative in that the question acts as a prompt to explore a particular expression or solution, not determine it. The assignation of meaning, identity, distinction, etc. is a particular form of activation in this ontology. The form activated, is the result of a set of individuation, or distinguishing, processes occurring in the field. For Deleuze, the field is the plane of intensities, for Luhmann, the environment, for the Schema.org, the Web (Deleuze 1968; Deleuze and Guattari 1987; Luhmann 2012).

This form of ontology does not to say that things have no existence apart from their interpretations, in fact it obviates the need for such constructivist understandings. The question of "what is?", either essential or constructed, misses the point entirely. The answer to this question demands an understanding of "how it is?", but hides that fact just as all claims of ontology hide their underlying instability (DeLanda 2006; Latour 2005; May 2005). Following from this, a Deleuzian approach to ontology, is an approach which tries to understand becoming, as opposed to being. It is an attempt to understand the activations and distinguishing processes that express certain forms and not others, affording certain interpretations while constraining others. This approach of multiple ontologies, like similar approaches by Actor Network Theory, are about the processes of connection, interaction, and assembly (DeLanda 2006; Deleuze and Guattari 1987; Latour 1987, 2005; May 2005).

Working backwards, uncovering how a given state is declared becomes a matter of exploring the processes of connection and interaction between components of the semantic data assemblage. We need to examine the ways in which those components specifically enact and create contextual fields within which meanings, identities, definitions, and forms come to take shape. However, assemblages are not composed of perfectly integrated parts. DeLanda

(2006) calls the relations between these components "contingently obligatory" (DeLanda 2006:10). The actual fact of integration between the various components of any given assemblage are only a matter of their empirical co-evolution. They exist together, but also apart, changing both as co-related elements and as elements involved in many other sets of relationships. As these relations are defined as relations of exteriority, they are not specifically beholden to the particular forms called forth in any given web of connections. Given this, we would expect to find a semantic data assemblage to be filled perturbations and inconsistencies. So, to understand the declaration of contextual states through attempts at representing a non-deterministic world to deterministic machines, we need to also investigate how those perturbations are negotiated and how the relations between those components push and pull in service to and against the assemblage at large.

3.2 Data and Data Assemblages

Studies of data have taken center stage in many different areas in recent years. Work in this area has primarily focused on big data, investigating its epistemology, biases in its model construction, the unequal distribution of data and the skills to utilize it, and the consequences of its use for privacy, identity, social sorting, and development, among other concerns (Andrejevic 2013; Arora 2016; Barocas and Selbst 2016; Bowker 2013; boyd and Crawford 2012; Cheney-Lippold 2011; Choudhury et al. 2014; Friedman and Nissenbaum 1996; Gillespie 2014; Leonelli 2014; Noble 2013; Sweeny 2013; Vaidhnayathan 2011; Warf and Grimes 1997). While this emerging interdisciplinary field is producing large volumes of valuable work on data and their infrastructure, the focus thus far has been dominated by big data, its use in sorting

algorithms, and its ethical implications. Missing from this growing body of scholarship are studies that investigate alternate forms of data, their production, and their assemblages. Echoing Seaver (2013), I argue that the components of algorithms, how they are constructed, and how they are classified are an essential part of understanding code-based systems. He argues that social scientists are operating with a flawed understanding of algorithmic systems. Reverse engineering studies and algorithmic audit studies are premised on the existence of a singular algorithm of study. Not only does he argue that algorithms are not singular, but also that they are not stable or unified decision-making tools. For example, at any given moment, any interaction with Bing's recommendation engine could be drawing on a large number of algorithms, and those algorithms are likely to be different for different users at different times. Companies, such as Microsoft, often run experiments with various algorithms to optimize user experience, a fact neglected by the experiments employed in audit studies. The transparency and expertise allegedly required for critical analyses of algorithms overlook the fact that algorithms are built by many people over periods of time who often are not even fully aware of the range of inputs and decision rules that contextual "real world" deployment brings about. Likewise, a sole focus on transparency of inputs obscures the potentially more important variable of data and feature creation. This focus on the "what" and not the "how" is often the concern of code studies. As Seaver argues while discussing Facebook's EdgeRank algorithm,

> If we are interested in talking about algorithms' cultural effects, then it is not enough to know that something called "affinity" is included. Our questions should be more ambitious: What is affinity? How is it defined? What experiences do the engineers producing affinity scores draw on as they attempt to formalize it? How might these formalizations differ if we started from different assumptions and experiences?... We need to examine the logic that guides the hands… choosing particular representations of data, and translating ideas into code. (Seaver 2013:9)

Here, the author argues that despite the publicly available knowledge of many of the components used in EdgeRank, absent information about how those components are calculated and the data involved are created, we know, and can learn, very little about the EdgeRank algorithm. To this point, studies of data making practices in socio-technical systems are lacking. While researchers engage in the process of data construction every time they do research, there is a deficiency in sustained empirical analyses that capture the ways in which data and their assemblages come to be. Moreover, the work that does either directly or tangentially engage with data making often treat data as given, not something enacted (Mol 2002; Vis 2012).

Floridi (2012) argues that what it is to be data is dependent on the position one speaks from. As Kitchin (2014b) summarizes, "from an epistemic position, data are a collection of facts, from an informational position, data are information, from a computational position, data are collections of binary elements that can be processed… form a diaphonic position data are abstract elements" (Kitchin 2014b:4). This sentiment captures the uncertain nature of data, as it relies on decision logics, perspective, context, and purpose among a great many other things. The central feature of this argument is that much like code based systems, data are not neutral or objective things (Friedman and Nissenbaum 1996). Gitelman and Jackson (2013:2) argue that "data are always already 'cooked'". An understanding of data must move beyond a simplistic understanding of them as "merely being the raw materials of information and knowledge" (Kitchin 2014b:185). Data are wound up in webs of socio-material relationships; they are subject to systems of thought, discourses of governmentality (Lyon 2007; Raley 2013), science (Ribes and Jackson 2013), privacy (Nissenbaum 2004; Solove 2004), and economy (Brine and

Poovey 2013; Gandy Jr. 1993; Mackenzie 2008), among others. Data do not exist in a vacuum as they are both subject to the field-specific discourses in which they are a part, the practices of users, and the specific forms of the standards and infrastructures that enable them (Bowker 2005). Any understanding of data must mind these contexts, systems of thought, infrastructures, standards, classification schema, technologies, communities of practice and institutions that are involved in their creation, collection, warehousing, and analysis. In other words, the study of data must also be the study of the data assemblage.

These data assemblages are complex systems that are grounded in specific modes of thinking, classification, infrastructures, standards, practices, institutions, policies, and technologies. Tables 1. & 2. detail the elements of data assemblages, outlining the various ways that different components "frame what is possible, desirable, and expected of data" (Kitchin 2014b:24). A consequence of this assemblage based approach to the study of data is that they are not simply the result of stable forms of classification and recording, they are dependent on the specific interrelations between and amongst elements of the data assemblage. While the list of an assemblage's components shows the various inputs into what makes data, work in this emerging area argues that data and their assemblages actually co-constitute one another. That is, while the assemblage produces data, the data act back upon the assemblage to modify its form and function (Kitchin 2014b; Kitchin and Lauriault 2014; Ribes and Jackson 2013). A consequence of this reflexivity is that the data and the assemblage are in a constant state of potential flux; that data and the assemblage both, are purely contingent. The interactions between components and the interpretations of the data produced open the possibility of all

states of being, but not the necessity of any. The particular state and the way that it is determined is the question to consider.

Table 1: The Apparatus and Elements of a Data Assemblage

Apparatus	**Elements**
Systems of thought	Modes of thinking, philosophies, theories, models, ideologies, rationalities, etc.
Forms of knowledge	Research texts, manuals, magazines, websites, experience, word of mouth, chat forums, etc.
Finance	Business models, investment, venture capital, grants, philanthropy, profit, etc.
Political economy	Policy, tax regimes, incentive instruments, public and political opinion, etc.
Governmentalities and legalities	Data standards, file formats, system requirements, protocols, regulations, laws, licensing, intellectual property regimes,
Materialities and infrastructures	Paper/pens, computers, digital devices, sensors, scanners, databases, networks, servers, buildings, etc.
Practices	Techniques, ways of doing, learned behaviors, scientific conventions, etc.
Organizations and institutions	Archives, corporations, consultants, manufacturers, retailers, government agencies, universities, conferences, clubs
Subjectivities and communities	Of data producers, experts, curators, managers, analysts, scientists, politicians, users, citizens, etc.
Places	Labs, offices, field sites, data centers, server farms, business parks, etc., and their agglomerations
Marketplace	For data, its derivatives (e.g., text, tables, graphs, maps), analysts, analytic software, interpretations, etc.

Source: Kitchin (2014b, pg. 25)

Table 2: The Apparatus and Elements of Schema.org

Apparatus	**Elements**
Systems of thought	Graph theory, resource description framework, entity property representation, generalized application, pragmatic modeling
Forms of knowledge	Research texts, manuals, websites, experience, W3C best practices, W3C community forum, Schema.org
Finance	Financed by sponsor companies
Political economy	Emerged in vacuum left from original semantic web vision as a way to improve search and content embedded in sponsor applications.
Governmentalities and legalities	Microdata, RDF and its serializations, JSON-LD, version control systems (Git),
Materialities and Infrastructure	Computers, databases, server farms, etc.
Practices	Bricolage, generalization, acontextual mapping, ontology modifications, etc.
Organizations and institutions	W3C, Google, Microsoft, Yahoo, Yandex, independent contractors, major corporate users
Subjectivities and communities	Schema.org ontologists, users, community members, academic participants
Places	Virtual spaces via Skype, conference calls, GitHub, Google docs, W3C message boards and IRC chats
Marketplace	Collection of use cases

Source: adapted from Kitchin (2014b)

The dilemma of context is also a prominent critique in recent studies of coded systems and big data. Work here argues that they enact a new brand of heightened technological

positivism that constructs a particular epistemic viewpoint where there is only what appears in the data. In this construction, not only do numbers simply speak for themselves, but there is also nothing to be found outside of the numbers (Andrejevic 2014; Bowker 2014; boyd and Crawford 2012; Floridi 2012; Kitchin 2014a; Leonelli 2014; Seaver 2015; Thatcher 2014). One of the charges against this techno-positivism is that it strips away any context from problem formulation, data collection, the algorithm, and the results. Thus, this removal of context renders the entire data processing operation meaningless (boyd and Crawford 2012; Seaver 2015). Critics argue that this is especially the case with aggregated datasets or data processed automatically through machine learning algorithms (Bowker 2013; Busch 2014; Mittelstadt and Floridi 2016). Busch (2014:1734) calls this process "layering". In short, layering involves the process of factor reduction whereby, in our collective efforts to produce greater precision, clarity, and objectivity, those aspects of things that are not amenable to numerical or statistical analysis—that situate particular phenomena—are systematically downgraded or removed from consideration" (Busch 2014:1735). Here the contexts of data collection, processing, analysis, and any other ontological positions that are implied by those methods are made invisible in the process of creating large datasets. That is, big data obscures the various ways that phenomena and the resulting data are enacted (Mittelstadt and Floridi 2016).

Seaver (2015), however, counters this trend, arguing that debates about the big data revolution stripping context from data miss the point entirely. He argues that context is exactly the battleground. Context is imputed to the instance, by the community, practitioner, or interpreter in every encounter. It is everywhere and always a matter of selection and investigation. Big data work establishes a context in the particular selection and mobilization of

data. This occurs in each stage of a problem solving such as target variable definition, data collection and labeling, feature selection, and model interpretation. This is not to say that critics of the big data revolution are incorrect or misguided in their critiques, but rather that social science's critiques of big data and algorithms are looking for a different type of context that big data's epistemological stance cannot ever satisfy. That is, data analysts and social scientists are using two different types of context. In this literature those types are positivistic and interpretive. In the positivistic set of contexts, "context is considered to be a stable container for activity: one's context can be described as an accumulation of data points such as location, weather, the people nearby, or the time of day" (Seaver 2015:1105). This type of context is a matter of pure construction from the data themselves. In the interpretive mode, context is performative and built up in the process of interactions, not something that can be reduced to a simple quantitative representation in a dataset. While it is clear that the two disciplinary views of context are speaking past one another, the major takeaway for Seaver and this study is that context is everywhere and at all times a matter of making decisions and statements about the state of the world. This means that context "is a necessarily political project" and as a result becomes a question of is creation (Seaver 2015:1106). So, while the development of semantic ontologies purportedly strips away context from Web text, it actually only ever displaces the creation of context. This ontological displacement, occurs at the level of the ontology as it sets the boundaries within which the world can be spoken of. This construction and its effects have not yet been investigated.

As mentioned, the role that context plays in the semantic web is complicated. Ontologists develop their ontologies to be acontextual and able to be applied across dissimilar

use cases. However, as I will show in later chapters, context plays a central role in their development. Likewise, the actual use of ontologies by Web developers never occurs outside of a specified context. Waller (2016), drawing on Heidegger (1973) and Lakoff (1987) shows that the definitions of concepts and their attachments to properties always necessarily rely on social context. Indeed, as Bakhtin (1981) argues, all claims are inextricably embedded in specific contexts. As we will see, this dilemma of context will create problems for creating and applying semantic data, as there is no way to establish or reconcile differences in the intention of development and intended application.

In her study of research use of the Twitter data, Vis (2012) takes a step to understand and critically evaluate the process of data making through an analysis of the application programming interfaces (APIs), researcher questions, data selection practices, and the computer programs they use. The author argues that APIs can expose or occlude certain types of data depending on access, sentiments echoed by work on the "big data divide" (Andrejevic 2014; boyd and Crawford 2012; McCarthy 2016; Mittelstadt and Floridi 2015). Additionally, this work finds that APIs often present data where data provenance is left in question. In this particular example, the Twitter API often displays imputed data that has its provenance hidden since it was imputed prior to its availability through the API. Relatedly, Ford and Graham (2016) find similar issues with geopolitical information in semantic search applications. In their earlier mentioned study of semantic representations of the city of Jerusalem, they show that representations of the city, its history, and entanglements with global politics are blunted, and the sources of the underlying data are obscured. Obscured data provenance is significant because, our concerns when studying the sociology of information should not just be about

which components are included or not, but where they came from, and how those components were created in the first place (Seaver 2013). A clear picture of data's provenance is essential to those aims. For example, Twitter location data are created from processes of investigation and imputation. While they may be accurate, they also may not be. In the large majority of cases, location data are imputed. These imputations may be based on neglected profiles, poor secondary data, or overly broad, thus inaccurate, location predictions. Regardless of the particular reason, the data are used and are applicable in analyses and decision making.

In addition to data's unknown origin, how they were developed is a central concern. Vis (2012) shows that researcher data making is a complex process of decision making. She argues,

> Decisions about what to collect (what is in, what is out), from which API data is collected, for which period, including which metadata, including an awareness of how this collected data is itself created by APIs, are important stages in the data making process. (Vis 2012:*no page*).

Here the author highlights just a few of the many decisions that researchers need to make in the process of constructing their datasets. Moreover, these decisions are made in the context of specific systems of thought, theoretical debates, and from specific ontological and epistemological positions (Boellstorff 2013). As Gitelman and Jackson (2013) argue, to think about data, we need to understand how researchers envisage their objects of research. Data in many cases are modeled. Here I do not mean used in modeling techniques, though that is certainly a common use of data. Instead, I mean to say that data are modeled as a part of their creation in a process of establishing their ontological state. These data producing models have a co-constitutive relationship to the data they model. They are created based on the disciplinary understanding of the world, itself an interpretation of sets of data and used to produce new data through transformation.

Superficially, data would seem to occupy a first order position in the knowledge of one's object of study. This orientation suggests that data is the necessary condition for conceiving of one's subject. However, data's real relationship to a discipline inverts that perspective. How one understands their object of study has a mutually constituting relationship to the data the one creates and mobilizes toward that understanding (Brine and Poovey 2013; Gitelman and Jackson 2013). In their analysis of Irving Fisher's contributions to data-driven economic theory, Brine and Poovey (2013) show the ways in which the understanding of economics comes to influence the understanding of data. Their analysis of Fisher's Appreciation and Interest shows that economist's understanding of value and financial modeling contributed to the transformation and creation of data to better fit the calculations and theory at hand. The process of transformation for theory matching is indicative of the distance between data and what they allegedly represent. Their works suggest that the particular understanding of economic theory has a role in conditioning the data mobilized in support of that economic theory.

Edwards (2010) remarks on a similar process with regards to the use of climate data. Discussing the infrastructure involved in climate data, Edwards argues that what we know as climate data has already undergone filtration, interpretation, and modeling by computer systems. Amassing the large stores of climate data is not a matter of collection. Since the existing climate infrastructure was developed over extended periods of time with shifting aims, the measurements produced by the available instruments were diverse, scattered across the globe in less than systematic ways, uncalibrated, and not always relevant to the new tasks of measuring global climate. This existing set of measurements were processed to become data by

climate scientists. In effect, the newly emerging understandings of global climate helped produce the very data used in the empirical study of global climates. Scientists first had to integrate, normalize, and convert the disparate sources of existing data to reflect the preexisting models. So instead of drawing on available sets of climate specific data, climate science was a matter of using computer models to transform disparate forms and locational measurements into representative, homogenous, and global climate datasets (Edwards 2010).

Indeed, data production as a function of the modeling and interpretation of existing forms of data is commonplace. These same processes occur in options pricing models, credit scoring, risk evaluation, debt rating, and many other practices (Krippner 2011). This work does a good job establishing that data are often managed and pre-processed before they become or are considered as "data". This implication has three main consequences. First, as has already been established, data are dependent. They are dependent on how a focus of investigation is conceived, the instruments used in measurement, the practices of experts, and the technologies involved in the collection, management, and use. Second, what counts as data, is a matter of specific perspective. What the work from Brine and Poovey (2013) as well as Edwards (2010) clearly indicates is that existing data is not always sufficient for the task at hand. Moreover, as Stanley (2013) echoes, data may be insufficient for the task which can exclude it from being considered data at all. That is, data, like technology, establishes a standing reserve of possibility which allows it to be understood as data or not. The specific contextual relationship that the data have to the set of interpretations to be applied to it, mark its ontological status as data. Vis (2012) argues similarly that big data analyses and the choice of data based on difficulty, convenience, or technical ability can serve to exclude certain types of

potentially very relevant and important data from analyses. She uses the analysis of images as an example. Until recently, and now only in limited ways, techniques used to analyze big data were insufficient for analyzing images. While images were an important potential source of data, they were not included as data due to the inability for automated analysis. Third, what follows is that data assemblages are enacted. The specific set of technologies, practices, and orientations brought to bear on data call them into being as data of specific types and allow them to declare certain things and not others. Tools and practices both create and display data in certain ways and not others. Moreover, the complexity of these tools, their actual programming, and the decisions built into them are not transparent or standardized.

What the above make clear is that data depend on a wide range of inputs. They are the product of many different people, framings, and goals. Work has only recently begun to investigate these data assemblages and their outcomes. Indeed, researchers in this area have highlighted the need for more work that engages in the sustained empirical investigation of data infrastructures, standards, modeling, and creation practices (Kitchin 2014b; Kitchin and Lauriault 2014; Kitchin et al. 2015a, 2015b). This work calls attention to the diverse forms of data and data assemblages, as well as the need to investigate the components of these assemblages and their interactions. For all the recent work on data assemblages, there is still a dearth of work that investigates how a data assemblage's components interact as part of its expression. This study seeks to fill that gap by investigating how specifically the interactions between components not only enact the semantic data assemblage, but the particular state of being it engenders.

3.3 Information Infrastructures and Expert Systems

Related to the set of tools and practices are the infrastructures that come to bear on data. Data infrastructures are "the institutional, physical, and digital means for storing, sharing, and consuming data across networked technologies." (Kitchin 2014b:32). There are a wide range of types and examples of data infrastructures including data holding systems, archives, catalogs and directories, repositories, and cyber-infrastructures. Of significant note, however, are studies that show the particular effects of the way infrastructure, and cyber-infrastructures in particular, are designed, maintained, and developed.

Data infrastructures are technologies for interoperability between various communities of practice. These boundary objects are approached with sometimes divergent interpretations of purpose, form, and function specific to those communities (Bowker and Star 1999). The use of agreed upon standards in these infrastructures can reduce this tension, allowing for smooth interoperability. However, mobilizing the set of standards on offer is not always as simple or predictable as an infrastructure's background operation would suggest. In information infrastructures, particularly in scientific e-infrastructure, these tensions, or "data friction", can result in the breakdown and failure of work processes and the underlying infrastructure (Edwards et al. 2011). In their study of two distributed environmental science projects, Edwards and colleagues (2011), show that standards, rather than being stable, predictable, and uniform means for enabling interoperability are often mobilized in ad hoc ways. Scientific practice is often messy and encounters novel situations that defy the sort readily handled by the existing configurations of the infrastructure and its standards. This work shows that standards, particularly metadata standards, can be mobilized in creative and unpredictable ways to suit

the task at hand and navigate the contradictory pressures required of standardization and flexibility (Hanseth et al. 1996). However, those mobilizations can come with the unintended consequences of affecting the form of data, its availability, and sharing.

In addition to the use specific characteristics and contexts, the development of information infrastructures has substantial effects on the form and content of their related data. Infrastructures are long duree systems, meaning that the focus on development needs to be cognizant of long-term problems, needs, and possibilities. However, the developers of these infrastructures are not always in agreement about the challenges and direction that the infrastructure is, or will be, facing. The tensions surrounding the goals, purposes, and motivations of participants affect the design, development, and implementation of these infrastructures (Ribes and Finholt 2009). These effects have far-reaching consequences. The specific infrastructural states and practices affect the possible ways that data can be stored, which affect the type of data that will be stored, which affect the type of data to be collected. Much like classification standards, information infrastructures have a jussive power over the data and their possible uses. They define and implicitly impute values to what is stored and what counts as data (Bowker 2005). Moreover, the data stored in these infrastructures, their uses, and their users, shape one another in different ways. Users of the Social Science Data Archive (SSDA) adjusted its data collection efforts to respond to the increasing scholarly demand for different forms of data, while at the same time, old and new audiences for SSDA data engaged in new types of analyses and new research partnerships with different governmental and institutional participants, as well as altering their outlook on their field (Ribes and Finholt 2009; Shankar et al. 2016). All of this is to say that the specific form information

infrastructures take and the specific decisions made by the developers of those infrastructures have a significant bearing on the types, amount, and uses of data connected to those systems as well as the range of possibilities that those infrastructures, and their data, engender.

Work in the social sciences that investigates ontologies as cyber infrastructure is limited but growing. One early study by Ribes and Bowker (2009) offers some initial inroads to understanding their development. In their study of the geoscience network (GEON) ontology, they examine the processes of translation between domain experts and ontology developers. This work shows the problems that occur between two different expert groups who share no domain knowledge and have different epistemic cultures (Knorr-Cetina 1999). While the geoscience community held a great deal of geological subject matter expertise, they had little computer science expertise. The reverse can be said of the ontologists charged with converting geoscience domain knowledge into a format able to be represented in a linked data ontology. Thus, the fundamental problem that Ribes and Bowker identify is one of "reapprehension", or how domain experts approach their knowledge bases in new ways (Ribes and Bowker 2009:200). Much domain specific knowledge is tacit and held by the communities of experts in those domains. As such, a great deal of that knowledge is taken for granted and rests in the background of practice (Ribes and Bowker 2009; Schutz and Luckmann 1973, 1989).

Domain expertise is such that experts spend their time deploying information and interrogating its quality, but little time thinking about the underlying organization of their background knowledge in its own right. As Ribes and Bowker put it, "the term reapprehension is intended to emphasize how they come to see data and knowledge anew as a question of informational order, and then seek to act on its organization as information." (Ribes and

Bowker 2009:215). In the GEON project, the interactions between domain experts forced both sets of experts to approach their tacit knowledge bases in new ways, not only to render them explicit, but also enable communication across disciplinary boundaries. In their case, this involved community discussions to triangulate and establish a community wide consensus about the best ways that their tacit knowledges could be translated as faithful representations of expertise.

As instrumental in understating the self-reflection involved in translating knowledge as this work may be, there are some causes for concern, however. While the translation between experts and ontologists is a necessary step in trying to represent specific forms of knowledge in machine-readable ways, it obscures the fact that this process is inherently political, as knowledge engineers, experts, ontologists, as so forth, are all engaging in processes of selection and modification as they grasp at bits of meaningful information, while pushing away those bits that are seemingly unimportant. Furthermore, even small subsets of the disciplinary knowledge are far too complex to represent and so those representations tend to over-determine concepts. While they recognize the many contested meanings both across and within disciplinary boundaries, there is little recognition of the difficulty or problems of trying to capture that disagreement in such determinate systems. In essence, they do not provide an understanding of how the translation between humans and machines is performed, which prohibits understanding the crux of ontologies, which is the translation from machine to machine.

We can look to past research on expert and formal systems to engage with these omissions. This work points to a similar problem regarding the translation of expert knowledge

across domains called the knowledge acquisition bottleneck (Collins 1990; Forsythe 1993). This concept defines a problem where knowledge engineers attempt to extract knowledge from domain experts. This process is time consuming and fraught with complications, as much of that domain knowledge is not amenable to modeling in computer systems. For Collins (1990, 2010) the problem is one of socio-cultural understanding. Collins (1990) stresses that deterministic systems, such as AI, can only approach the forms of expertise and knowledge that most closely resemble the formalisms of computers (Berg 1997). That is, formalization is only ever possible where society itself has been constructed to resemble the processes of machines. While rule based knowledge – tacit, heuristic, or otherwise – can often be modeled here, they have trouble grasping many other forms of tacit and socio-cultural knowledge, which includes the ability to understand, use, and interrogate those knowledges, a point echoed by Ager (2003). Collins argues that the success of an expert system vis-à-vis knowledge acquisition rests on the availability for interpretation since all knowledge cannot easily be transferred through the system. Forsythe (1993) agrees here, noting that certain domains are more amenable for knowledge extraction and modeling than others. However, there is some reason for skepticism, as it obscures the fact that ambiguity exists at all levels of language, regardless of domain. Problematically for the semantic web, a consumer's interpretive capacity is often stripped away as data provenance is obscured (Ford and Graham 2016). Furthermore, while I agree with both Collins (1990, 2010) and Agre (1995, 2003) that agents must always make sense of communication in every instance, as interpretation is situated activity, this does not deal with the fact that situated context is the thing to be negotiated.

Forsythe (1993) echoes Collins and Agre, but argues that the problem of representation is also a problem of the approach to knowledge that the engineers have. She argues that engineers approach knowledge and the acquisition problem in two ways. First, they are like Star's (1994) naïve formalists, only viewing knowledge as something codified and formalized. She argues that knowledge engineers do not consider tacit of socio-cultural knowledge as knowledge. Rather, knowledge, in their view, is something that is able to be made explicit in texts, as something to be extracted or transferred. The second understanding is in the way they evaluate knowledge. Here, she argues that engineers evaluate knowledge introspectively, not through seeking empirical evidence. These understandings of knowledge lead her to argue that knowledge modeling in expert systems is a matter of deletion, noting that a "notion of straightforward transfer completely obscures their own role in the selection and interpretation of what goes into a knowledge base." (Forsythe 1993:463). Furthermore, she argues that while sometimes these knowledge engineers are aware of what they are doing, this selection process is bracketed from direct consciousness. Elsewhere, Agre (1995) argues that knowledge engineers viewed the formalization process as creating a sort of precision that the original knowledge and language lacked where "the vagueness and ambiguity of ordinary language are repaired through mathematical definition" (Agre 1995:no page) accordingly these engineers did not care much about the bias and semantic inaccuracy that resulted.

These authors deal well with the problems of translation between experts, and the intractability of converting certain types of knowledge in a machine-readable way. However, the actual processes of translating that knowledge for computer specific consumption is not well covered. While past work investigates the specific natures of certain type of knowledge,

the orientation to experts, and the approaches to knowledge that engineers have, they lack any discussion of the interactions with other important components of technical systems. That is, they do not examine the data modeling approaches, methods for classification, or the standards used in creating machine-readable data. It is my contention that only an assemblage based approach that explores those interactions between components can truly help us understand how human-machine translation occurs and how specific representations of the world become codified in technical systems.

3.4 Classification and Standards

Semantic web ontologies, like all classification systems, are more than just assigned tags to existing phenomena. They are more than just reflections of some alleged underlying reality. Ontologies and all classification systems are socially constructed systems that play a role in constituting the realities that they purportedly represent (Bowker and Star 1999; Westbrook and Saperstein 2015; Zerubavel 1996). Our attention, then, must be directed to ways those systems are constructed, the realities they construct, and the consequences of the institutionalized classification schemes used in these systems. In this case specifically, we must look at the construction, output, and consequences of semantic web ontologies.

In sociology, studies of classification systems have been primarily in organizational, neo-institutional, and cognitive domains. Organizational studies of categories have directed their focus on the various ways category construction and adoption mark similarities and differences between companies and their products (Zhao 2008, Zuckerman 1999, 2000). Neo-institutional work on classification has primarily focused on the contests surrounding institutionalized

meanings of categories and the institutional isomorphism that arises from shared understandings of institutional fields (DiMaggio 1987, DiMaggio and Powell 1983, 1991). Lastly, cognitive sociology has investigated the ways in which people draw on, and construct, categories in "shaping how we perceive and organize our realities" (Brekhus 2007:450). While approaching the topic of classification systems from two different vantages, cultural cognition and neo-institutional approaches in particular share an outlook that classification systems provide the social substrate from which actors can meaningfully engage with the world through a set a stable and shared interpretive schemas and rules (DiMaggio and Powell 1991; Meyer and Rowan 1977; Zhao 2008).

Recent work in survey design research questions how best to structure a given classification system, showing a degree of self-awareness in the problematic way that classification systems may be built. This recent work criticizes the running first order assumption that surveys and/or classification systems can more or less accurately and without fundamental faults, represent the world, rather than constitute it. As such, within this work, there has been very little critical attention to the construction identities as a consequence of the classificatory sorting that happens when categories are chosen and/or ignored (Westbrook and Saperstein 2015).

In their study of large social survey designs, Westbrook and Saperstein (2015) being to fill this void by showing that surveys create serious political problems where they represent sex and gender. In their work, they find that surveys largely limit sex and gender choices to two normative dichotomous states foreclosing on other possibilities. While steps have been taken to expand the range of representations, particularly in popular culture, work still remains

(Oremus 2014; Westbrook and Saperstein 2015). In addition to the limitations presented by overly broad classificatory categories, surveys also tend to conflate sex and gender, obscuring the significant conceptual differences between the two things. This form of reduction in classification systems leads interpreters to ignore both the between and within-category differences to the effect of not only valorizing the choices on offer but also overcoding diversity in categorical distinctions (Zerubavel 1996). For Deleuze, this represents the failings of continental ontology as it asserts positive identity at the expense of difference (Deleuze 1964; May 2005). In the process of reducing the anxiety surrounding ambiguous categories, the realm of possible states and interpretations is foreclosed. The political issues potentially involved in binary classification systems are not new. Neither are the implications of exclusion and inclusion of categories or conflating ontological states with one another, as category boundaries are rarely neatly drawn or equally valued (Lakoff 1987; Wittgenstein 2009). We have long seen the effects these classification issues have had on subjectivity, legitimate forms of knowing, and world building, particularly as they have been described in post-structuralist, feminist, and post-colonial theories (Bowker and Star 1999; Derrida 1967[2016]; Haraway 1991; Spivak 1999).

Many scholars have theorized the consequences of classification. Foucault, in the preface to The Order of Things, writes while referring to the oddities of animal classification in a Chinese encyclopedia, “In the wonderment of this taxonomy, the thing we apprehend in one great leap, the thing that, by means of the fable, is demonstrated as the exotic charm of another system of thought, is the limitation of our own, the stark impossibility of thinking that.” (Foucault 1973[1995]: xv). Here, Foucault is making a critique on the analogous Western system

of classification and its supposed objectivity. Rather than capturing any real connection to the world, classification systems are only ever specific versions and attached to specific cultural contexts. Where a classification system exists, an alternative might always be found. For his part, these classification schemes are the discursive regimes that order our interpretations of phenomena and act to construct our subjectivities (Foucault 1973[1994], 1977[1995]; Haraway 1991). The ontological, or state of being, in this view is only ever historical contingency in disguise.

In debt to Foucaultian archeology, Hacking (1986, 1999) developed an analytical framework for studying the act of "making up people". In this framework, he examines the way in which scientific classifications create new ways of thinking about people that reflexively act back upon the very people classified. Lauriault (2012) draws out five features in this process, classification, people, institutions, knowledge, and experts. In this "looping effect", people are classified and made to fit into modified or altogether new categories and they then come to interpret those categories differently, often self-identifying with those categories. As these categories gain familiarity and internalized acceptance, they begin a process of institutionalization. Once institutionalized, they often go through refinements, and reproductions as groups of experts enact those classification systems, beginning the process again. While Hacking (1986, 1999) uses the looping effect to describe the classification of people, and Lauriault (2012), the classification of space, this framework can apply to the study of semantic data, specifically at the point where the loop turns back upon itself.

No doubt inspired by this work, contemporary studies of big data echo Hacking's findings. These scholars argue that modern data analytic practices can have major ramifications

for a data subject's self-determination (Cheney-Lippold 2011; Gillespie 2014). As categories are inferred from data streams and applied to people, algorithms are being used to automatically determine individual and group level traits and characteristics. These classifications, however, are simultaneously being recreated and their meanings are being newly re-determined. The products of those algorithms create the new conditions in which those same algorithms operate. As Cheney-Lippold argues, "We are effectively losing control in defining who we are online, or more specifically we are losing ownership over the meaning of the categories that constitute our identities." (Cheney-Lippold 2011:178).

The implication of their thought is the act of classification is not an act of naming, it is an act of creation that constitutes the very thing that it names. While this may seem overly constructivist, it bears more similarity to the a Deleuzian perspective as well as those aligned with Gibsonian affordances (Introna 2007; Lauriault 2012; Parchoma 2012). Hacking explains thusly, "Who we are is not only what we did do, and will do but also what we might have done and may do. Making up people changes the space of possibilities for personhood… our possibilities, although inexhaustible, are also bounded." (Hacking 1986:114). The importance of this sentiment is that essence and existence are artifacts of contextual interpretation, that classification classifies and creates, but not completely and not immutably. The space for interpretation here is bounded by the particular activations that make up people.

Other work in the area has directed its focus on classification systems' their taken-for-grantedness and their embedding in standards and infrastructures (Bowker and Star 1996, 1999). This line of work investigates classification systems along four primary themes. First, classification systems are ubiquitous and pervade all aspects of our lives. Moreover, these all-

pervasive classification systems are not discrete, they are interwoven with one another. However, that they are ubiquitous and interwoven, does not mean that they are intertwined in a completely fluid or stable way. As classification systems vary by internal logics, divergent perspectives, and applied context, some components may be antithetical to one another despite other shared similarities. This last complication has major significance for semantic web ontologies, as their ability to supply semantic content and to function as part of a web of linked data relies on the smooth interoperability and a shared set of perspectives on a given topic.

The second theme is that classification systems have materiality. While investigations in this area tend to focus how classification systems are embedded into built environments and offer certain affordances to material structures, the central insight is that classification systems do actual work in the world and are not just virtual. While this study recognizes the various server farms, companies, personal and organizational resources, etc. that materially comprise the semantic web, it does not investigate its material aspects in this way. While these concerns are valid in their own right, a focus on this aspect of the semantic web draws attention away from my central questions.

A third thematic approach is that classification systems order the past in indeterminate ways. This orientation to classification systems seeks to uncover the ways that seemingly universal and natural orderings of the world come to be constructed and how alternate perspectives are made invisible. This is particularly important to my investigation into how semantic web ontologies declare ontological states. Attention drawn to the particular orderings of classification systems serve only to obscure the what is left aside, or ontologically over-

determined and occluded. The last major strand of this research involves unpacking the politics of classifying. As Bowker and Star (1996:4) note:

> There are two aspects of these politics: arriving at categories and standards, and, in the process, deciding what will be visible within the system (and of course what will thus then be invisible). The negotiated nature of standards and classifications follows from indeterminacy and multiplicity that whatever appears as universal or, indeed, standard, is the result of negotiations or conflict.

Much like code-based systems, all classification schemes are constructed by and for someone. Someone or some people must decide the nature of the categories, what the decisions rule of inclusion in a category will be, and the way that the classification will be applied in a standardized way. The crucial question then is to understand what that seemingly invisible code is, how the system is developed, how categorical incoherence is negotiated, what consequences emerge from inclusion/exclusion criteria, and what is rendered invisible (Epstein 2007).

These analyses ultimately understand categorization systems as fields for the exercise of power and politics (Bowker and Star 1996, 1999; Epstein 2007). A major focus of this work is on the development and deployment of standards, particularly classification standards. As standards often rest in the background, their power and the outcomes they enable are not always obvious. The jussive power that standards hold has inspired a large body of research into questions about who benefits from them, how they are implemented, what happens when they come into conflict, and what outcomes they promote, among others (Timmermans and Epstein 2010). The particular investment in a standard can have a transformative effect on the environment it is applied to (Becker 1992). Standards do work that shapes the world (Bowker and Star 1999; Star and Lampland 2009). Thus, the particular form, application, and underlying

decision logics in which a given standard is utilized are likely to have differing effects depending on the niche that they are being applied to (Epstein 2007). For this study, this is particularly relevant as it pertains to classification standards and their deployment. Scott (1998) reminds us that not only are the particular configurations of classification standards central to shaping outcomes, but those standards and outcomes privilege a particular point of view. The development of a standardized system of permanent last names, cadastral surveys, taxation, and agricultural metrics shaped the way that individuals related to larger institutional bodies as the subject of a state. Moreover, we see that the application of a classification standard does not necessarily bear any resemblance to an actual given ontological state, but does serve to create it when applied.

The form a classification system takes depends on the particular schema mobilized in its construction and the particular form of institutional or cultural logic adopted in the field (Douglas 1986). Zhao (2008) investigates the two different wine classification systems used for Californian and French wines. He notes that Californian wines adopt a system based on the type of grape, wine making process, and location in the state, with a horizontal form of classification where inputs are not explicitly more valuable than others. French wines, on the other hand, use a system based on a wine's terroir. In this system, a wine is explicitly ranked in a hierarchy. The significance of this work is that "under diverse classificatory schemes and structures based on distinct institutional logics, similar categories may have different significances in consumers' valuation of wines across these two industries" (Zhao 2008:169). While this study does not adopt a neo-institutional framework, it does take this essential argument to be important. The particular schemas and rationales used to develop a classification system have inevitable

effects on the uses and interpretations of artifacts that draw on that classification system. For the present case, the particular schema and decision-making affect not only what and how semantic data can be represented, or which domain actors can mobilize a given classification scheme, but also how those ontologies can interact.

Bowker and Star (1999) refer to the things, where an instance does not fit the delimited borders of one category or meaning set, as boundary objects. Elsewhere, Zuckerman (1999) notes that when products are affiliated with diverse categories, they are structurally incoherent, meaning that they break with the structural organization of the classification system that holds categories as mutually exclusive, even if that represents a near impossibility (Bowker and Star 1999). Zhao (2008) argues that a boundary object's coherence is a function of the importance and specificity of the categories the object is transgressing. Where a boundary is subject to greater contests of significance, or more firmly established inclusion/exclusion criteria, a boundary object is less likely to occupy a transgressing role. The existence of these boundary objects depends, in large part, on the particular decision logics and classification schema at hand.

In practice, we know that classification systems are often not mono-schematic. That is, they are often integrations and compromises between multiple schemas and multiple social worlds. This is particularly the case where large-scale technical systems are being developed and the very development depends on coordinating multiple communities of experts (Bower and Star 1996, 1999). As Timmermans and Epstein (2010:84) argue, "standards transform by coordinating disparate elements, but the outcomes that standards achieve depend on the specific standards and the circumstances under which they are made to work.". This argument

is inspired by a number of studies into the workings and deployment of standards. First, these specific circumstances are relative to specific social worlds. While standards bridge communities, there are specific sets of users and communities of practice that surround any individual standard, and the adoption, and mobilization of a standard is impacted by the specific target, user, or institution involved. For instance, research has shown that the specific standards applied to different bodies for biomedical research and insurance classification depends on the specific body in question (Epstein 2007, 2009; Lengwiler 2009). In other work, researchers found that the type of metadata standards, the opinions about them, and the future-oriented outlook on their success used in scientific research varied based on the specific scientific community interacting with those standards (Edwards 2011 et al.; Millerand and Bowker 2009). In these cases, holding favorable opinions and successful outlooks depended on one's proximity to the standard's creation and the level of knowledge one had of the intricacies of the standard. Their findings suggest that the way one interacts with a standard is contingent on one's location in the social world surrounding the standard.

In addition to standards being relevant to specific users and communities, they are partially dependent on the specific technological infrastructure of which they are a part. Millerand and Bowker's (2009) work shows that the successful implementation of a metadata standard for classifying ecological research depended on tools used to implement the standard, the standards compatibility with existing infrastructure and work practice, and the data storage and management techniques and technologies in use. This is in line with Bowker's (2005) earlier work that suggests that way we understand information depends on the "memory practices"

that we utilize, including the technical means for storing, the practices of labeling, and other components of recording information.

As standards are embedded in larger sets of relationships, interactions with them are not always smooth and they do not always work to enable interoperability. Like Latour's (1987) useless facts, standards sometime get in the way. While Bowker and Star (1999) argue that standards tend to have stabilizing effects, they are in actuality often disruptive things to be negotiated. Some work has noted these negotiations where practitioners work around the limitations presented by tools (Berg 1997; Whooley 2010). Using electronic medical records as an example, Berg (1997) shows how nurses work around the limitations of software. Nurses in a medical ward were tasked with recording patient conditions in computerized software. Problematically, though, when a patient is unstable or when only incidental changes occur, frequent updates can either be too cumbersome or unnecessary. Additionally, poorly designed records systems require can sometimes inhibit efficient work practices. Berg's nurses often times write patient conditions on paper, to be entered later, or use a second terminal to work around limited information visible in a given screen. Likewise, Whooley (2010) shows that psychiatrists often attempt to assert their professional autonomy in diagnosing mental illness by creating workarounds to standardized diagnosis typologies like the Diagnostic and Statistical Manual of Mental Disorders (DSM). He argues that working around the DSM involves drawing on alternate diagnosis typologies, manipulating illness codes on formal documents, and negotiating the extent of diagnoses with patients. Koehne et al. (2013) have similar findings in their study of diagnoses of borderline personality disorder in adolescents. Halpin (2016) builds on this work, but disagrees arguing that psychiatric professionals work within the boundaries of

the DSM, as opposed to around it. He argues that even though, professionals negotiate DSM diagnoses, they still do so in line with the DSM. They deploy the manual strategically in different ways depending on the professional and institutional context, noting that its influence is felt across research, clinical, and institutional domains. The major difference is that the way the DSM is used to enact diagnoses depends on that setting.

While much past work has located standards within communities of practice, work still remains to explain the way that standards' affordances and constraints are negotiated (Sandholtz 2012). These omissions beg researchers to situate standards' use in practice and question how these blockages are negotiated and how the restrictions created by standardization are relaxed while still enabling the interoperability they are designed for. While scholars have been exploring the tensions between organizational level practices and the homogenizing effects of standards for some time, work that both locates standards at the level of professional communities and that investigates the negotiations of barriers is only just begging to emerge (Epstein 2007, 2010; Halpin 2016; Heimer 2001; Whooley 2010, 2014; Whooley and Horwitz 2013). As two different components of the semantic web data assemblage, addressing this gap through developing an understanding of how communities of practice negotiate with the limitations and affordances of standards is paramount.

3.5 Empirical Ontology in Practice

What is lost in Hacking's (1986, 1999) looping model, but implicit in an assemblage model is that contextual interpretation is cultural. The specific "epistemic cultures" (Knorr-Cetina 1999) that bear on a given assemblage impact each point in the loop. Knorr-Cetina

(1999) studies the ways that scientific cultures come to act on various disciplines – high energy particle physics and molecular biology – to create a disunity in the scientific community. Her goal isn't to show how knowledge is created differently, as much as it is to show how disciplines themselves are enacted differently and how the focus of research is enacted differently based on the practices of scientists in those disciplines. Much as there are many different epistemic cultures, there are many different ways of approaching interpretive schema and classifying phenomena (Friese 2010).

Practice theory views practice as not only activity but also the knowledges and capabilities that enable it. Furthermore, all activity and the knowledges, capabilities, and meanings behind practices are more or less organized in a field, which can be demarcated and subdivided ad infinitum. This type of thinking holds that "the social is a field of embodied, materially interwoven practices centrally organized around shared practical understandings" of the world (Schatzki 2001:12). Practices structure and are structured by social location (Bourdieu 1977), discursive regimes (Foucault 1977[1995]), embodied experiences, technology (Latour 2005; Law 2010), and the immediate physical environment, among other things (Goffman 1959). Accordingly, determinism has little place in practice theoretical accounts, as practice, while durable and habitual, is emergent and unfolds as community members interact with the various conditions of their practice.

Practice's durability is often a major concern of social theorists, as Knorr-Cetina (2001:184) notes "conceptions of practice emphasize the habitual and rule-governed features of practice… agreeing that practices should be seen as recurrent processes governed by specifiable schemata of preferences and prescriptions.". This view of practice, she argues, is ill

equipped to explain the sphere of activity implicated in non-routine problems and the relationship to objects that are entirely open-ended, that generate new lines of questioning, new forms of interaction, and new possibilities of being. In her analysis of knowledge work, she shows that practice moves between rule-bound, habitual procedures and emergent intra-active interpretation. Her theory of objectual practice argues that objects of inquiry, or what she terms "knowledge objects" (Knorr- Cetina 2001:185) only ever become partial objects, as they are first simulated and represented and then subject to evaluation, correction, and improvement. They never completely shift into the background of habitual practice. These objects and the relations in which they are implicated are never complete; they are always emerging and transforming based on the intra-actions between objects and the individual(s) engaging with them. Investigation and interaction with them produce greater levels of complexity rather than a reduction in it (Knorr-Cetina 2001). "The signifying force of partial objects resides in the pointers they provide to possible further explorations. In this sense these objects are meaning-producing and practice generating; they provide for the… extension of practice." (Knorr-Cetina 2001:192). Knorr-Cetina taps into Deleuzian notions of molarity and molecularity here (Deleuze and Guattari 1987). As one attempts to tame and manage complexity, features always evade attempts to be fully understood. The complexity of things is inimical to our development of classificatory schema to apprehend those things. Here, capture and escape are inextricably related (McCarthy 2017).

The concept of informatic practice borrows from Knorr-Cetina's theory of objectual practice as French (2014) conceives of information, its relationships, and objects as never complete. In his study of Big Data processing in the Canadian health industry, he aims to dispel

the hyperbolic discourse around data surveillance. Advancing the concept of informatic practice, or "the assumption that information has a material basis in the spatio-temporal milieu of everyday life." (French 2014:230), he details the ways that patient data is collected and coded and embedded in material practice. An understanding of immaterial information requires attention to that mundane performance of routinized work – filling out forms, categorization, filing, and communicating effectively. For French (2014) and others (Albrechtslund and Lauritsen 2013), practice contains human error and deviations from expected performance. This material breakdown contributes to an immaterial breakdown where data does not accurately represent the thing intended. This study adapts and improves this line of thinking in at least two distinct ways. First, it makes semantic ontologies the target of objectual and informatic practice. In doing so it focuses attention on the practices of interaction and construction as opposed to strictly the practices of deployment as does French (2014). Second, it extends Bowker's (2005) work on cataloging and memory practices by examining the data assemblage as a necessary and inseparable component of data creation and gathering.

In recent years, many science and technology studies scholars have shifted their focus on the ways in which phenomena are brought into being, or enacted. (Coopmans et al. 2014; Mol 2002; Woolgar and Lezaun 2013, 2015). This "turn to ontology" (Coopmans et al. 2014:2) reflects the philosophical move away from terminology that posits any stable and concrete entities of representation, and the framing of objects of investigation as independent of the investigations themselves. As many scholars argue, the object and its means for investigation co-constitute one another (Barad 2007; Introna 2005, 2007; Kitchin 2014b; Kitchin and Dodge

2011; Latour 2005; MacCormick 2013; Mackenzie 2005; Manovich 2013). Barad (2007:818) notes, “The primary ontological units are not ‘things’ but phenomena – dynamic topological reconfigurings / entanglements / relationalities / (re)articulations. And the primary semantic units are… material-discursive practices through which boundaries are constituted.”. In other words, in this intra-actional approach, subjects, objects, and discourses have no ontological priority; they are one another’s common history. Observers, instruments, and objects of investigation do not exist as such independently of their entanglements. Moreover, these intra-actions exude a certain potentiality of being, affording and proscribing certain outcomes. Introna (2007) shows that the intra-actions of the plagiarism detection phenomenon enable and constrain certain uses, forms, and meanings of technology as well as the identities and practices related to plagiarism detection. The intra-actions involved in the use of word processing, electronic scholarly databases, Wikimedia, Turnitin detection software, the commodification of education, fairness, and intellectual property rights all combine to re-constitute what it is to plagiarize, to write, to instruct, and to detect plagiarism.

The shift in focus from the epistemological concerns of representation to the ontological concerns of enactment means that concern is no longer simply on how well science understands its object, but how it assists in creating it. Similar to work on assemblages, studies of enactment investigate the ways that phenomena are called into being through practice (informatic or otherwise), engagement, social interactions, and artifacts (Coopmans et al 2014; French 2014; Knorr-Cetina 2001; Law and Lien 2012; Mol 2002; Moser 2008; Sismondo 2015; Woolgar and Lezaun 2013, 2015). Objects can, and must, be understood as manipulated, created, destroyed, and made meaningful through these practices. Mol (2002), shows the

various ways that atherosclerosis comes to be enacted in different ways through different practices, arguing that the enactment of the disease is contingent upon the various practices engaged in any intra-action. Variations in practice lead to the enactment of multiple realities for a given disease, such that the meaning of atherosclerosis depends on the practical engagement with the whole of the wider actor assemblage. For instance, atherosclerosis as pain in the leg is enacted through the practice of a patient consultation whereas the disease is differently enacted as arterial blockage through post-mortem or post-amputation investigations in pathology departments. In this case, the creation of the objects of investigation is inseparable from the investigation itself. Reference and representation refer only to the reality as constituted.

Similarly, Alač (2013) shows how results of fMRI scans acquire meaning through enactment. The fMRI visual output acts as an intermediary between an embodied individual and knowledge of that individual. Often times lab technicians are unable to see the physical patient and must rely solely on the output from the scans. So too, secondary research has no ability to see a patient and must interpret brain signals as representations of a given patient. In these scans, patient motion creates brain scan nonalignment and thus interpretive problems in the visuals for technicians and researchers. One task of these practitioners is to spot those nonalignments and decipher their origin. Of course, this is problematic given that the observer is never actually able to see the patient or their alleged movement. They rely on data from scans that display “hints” that an informed practitioner can decode as specific types of movement. Through the specific intra-action between instrument, researcher, and the

mobilization of specific cultural knowledge those hints allow the practitioner to enact movements in any given nonaligned scan.

Drawing on this "turn", Kallinikos et al. (2013:357) argue that digital objects have something of an "ambivalent ontology". Like Knorr-Cetina's (2001) knowledge objects, digital objects – semantic web ontologies – are incomplete and always changing. For the semantic web the issues is not so much about representation, because while practically very important, we know that "real" representation is not possible. Following Woolgar and Lezuan (2013) I do not begin by assuming that there is a picture of the world that ontologists are getting right or wrong, but rather stipulate that they are engaged in the "practices of depiction" (Woolgar and Lezuan 2013:324), but only in so far as those practices are bound up in a larger assemblage. As Latour (1987) once argued, to understand the ready-made, we first need to understand the making. To "get the most efficient machine" we first need to "decide on what efficiency should be" (Latour 1987:9). The interactions between components of the semantic web are not as simple as the black box implies. They are a set of negotiations with the affordances and constrains standards, design philosophies, other people and the existing state of the ontology. This new Janus-faced project needs to decide how meaning is to be made and how translation is to occur before simply making the translations.

3.6 Conclusion

The problem of translating between humans and machines posed by semantic web is bigger than just the translation between different sets of experts or making knowledge explicit. Instead, it is a problem of creating a distinct world of understandings and meanings among an

entire fluid universe of them. It is a process where decisions about representation are made and codified. It is a process by which the affordances and constraints of information infrastructures, standards, and classification systems are negotiated by communities of practice. Lastly, it is a process of co-creation between data and their assemblages. Where science studies research once investigated the use of machines to translate the natural world for human consumption, I now invert that relationship and investigate the way that humans translate the world for machine consumption (Knorr-Cetina 1999; Latour 1987; Vertesi 2015). In so doing, I make contributions to re-emerging work on AI and ontologies, negotiations with standards, and data studies. Furthermore, I answer Woolgar and Lezuan's (2013, 2015) call to open new provocations for empirical ontology.

4. Researching Schema.org and the Semantic Web

How does one unpack the semantic web's black box? As code based systems tend to be, its operations are hidden within complicated computer languages, behind the walls of trade secrets, and embedded within other larger technological systems. Similar to other information infrastructures, these features make the semantic web more or less invisible and taken for granted. Much of this is by design, as ontologists purposefully work to make the semantic web as seamless and simple to use as possible, so that it can be the substrate of larger more visible technologies. Research on information infrastructures argues that such features – size, ubiquity, embeddedness, and community specialization – work to render these systems invisible, but has provided a number of means to expose them, such as ethnographic observation, examining systemic breakdown, and specialized techniques like infrastructural inversion (Bowker and Star 1999; Bowker et al. 2010; Edwards 2010; Ribes and Finholt 2009; Star 1999).

My research problem confronts two main methodological quandaries. First, what case is large enough to draw significant insights from, but exposed enough that it has not completely disappeared into the Web's architecture, and thus permits investigation? As I explain below, Schema.org is such a case. Second, following from my chosen case, what methodological approaches will allow for an examination of a system whose work processes are dispersed, natively digital, and will fit with an assemblage based approach to answering my chosen research questions? This chosen methodologies will by necessity attend not only to the communities of practice, but the standards they encounter, the systems of thought they bring to bear on their work, and the ontologies they interact with, among other things.

The ensuing chapter resolves these two quandaries though a justification and explanation of my case, as well as the methodologies deployed to answer my research questions. This dissertation relies on the in-depth case study of the users, developers, vocabulary, code standards, and digital artifacts directly associated with the Schema.org project. As both the objects of investigation and their impact on my research questions are diverse, multiple methodological approaches drawn from recent work on digital technologies are employed. This chapter is divided into five parts. In the first part, I will describe the case that guides my research. This section will involve a description of Schema.org, the major institutions involved with the project, the communities of practice that are most closely involved in its development and its use, the vocabulary that constitutes the project, and the primary code standards implicated in its creation and use. The second section will describe the primary sites in which my research occurred. The third section will describe both the data collection processes and the preparations for analysis. The fourth section of this chapter will discuss the two main methodological techniques used in this study. The fifth and final section of this chapter will describe the limitations of my methodology and include a statement of my standpoint as a researcher and the biases that such a standpoint and approach may bring about.

4.1 Schema.org

I selected Schema.org as my case to study the processes involved in constructing data assemblages for a few main reasons. First, the project is unique in that it represents a nearly even participation between industry competitors. The four major search engines, Google,

Microsoft (Bing), Yahoo[12], and Yandex that are involved with the project all contribute freely and openly to Schema.org's development despite being competitors in search, and in the cases of Google and Microsoft, competitors in search based artificial intelligence, a major use of semantic web technology and structured data. This collaboration is significant because it represents not only a consolidation of semantic web ontologies, but also a strong symbolic statement in favor of Schema.org being the "best" ontology. My research indicates that this statement of support makes this particular case the de facto semantic web standard in eyes of users and industry experts alike.

The second reason for selecting Schema.org is that it is a public and transparent[13] process. Anyone who has the desire or motivation to contribute to the project's development is welcome to participate. In practice, participation is rather limited as there is a substantial amount of expertise, experience, and knowledge required to contribute at a high and productive level, or to be regarded amongst the core developers as competent. Nevertheless, there are many contributions by less central community members. These contributions occur in the development of both the core and extended ontologies across both of the two primary work sites. The community of developers places a high degree of importance on transparency in how decisions are made. This is evident in both the community facing, everyday discussion, and the more selective steering group discussions, which are published in various formats.

[12] At the time of writing Yahoo is in the process of being broken up. This included the sale of many parts of its core Internet business to Verizon, the resignation of its CEO, Marissa Mayer, and other major forms of restructuring. It is uncertain, at this point, what will become of Yahoo's remaining businesses, Verizon's commitment to search, or either company's involvement in Schema.org.

[13] This assumes that one is both aware of the project and how to navigate the work sites.

Moreover, all of this work and discussion regarding development and use happens in public[14]. All discussion is cataloged on either the World Wide Web Consortium or GitHub's websites. These features make Schema.org an ideal case for study because there is a precise and detailed record of exactly what discussion happened, when they happened, the results of those discussions, and the underlying rationale behind the community's decisions.

The third reason for selecting Schema.org is that it is a wide-ranging and general ontology. Self-described as a set of schemas, the project aims to cover diverse content domains, rather than specialize in a single or a few more limited domains. This means that Schama.org is more directly trying to represent the known Internet, rather than explicitly limiting itself to niche areas, as is most often the case with semantic web ontologies. This directly suits the interests of the sponsor companies, particularly Google and Microsoft, and has consequences discussed later. Lastly, Schema.org is one of the semantic backbones behind Google, Microsoft, and Apple's search based artificial intelligence systems. While these systems are all varied and have additional proprietary knowledge bases, one of their major functions is to serve as an intelligent interface between users and search. While the specific implications of this interfaced layer are beyond the scope of this study, investigating how those AI systems come to understand what they are parsing is of direct relevance, and of high sociological importance.

As mentioned above, Schema.org is a semantic web ontology developed in collaboration between four sponsor companies – Google, Microsoft, Yahoo, Yandex – independent

[14] Excepting the conference calls between steering group members. This is a limitation that I address later in this chapter.

contributors, and the W3C. The central stated aim of this project is "to create, maintain, and promote schemas for structured data on the Internet, on webpages, in email messages, and beyond." (Schema.org 2017a). Schema.org is an open source and peer produced project that is attempting to create a single unified place for webmasters to go to learn about and apply semantic markup to their web domains. The project is particularly focused on defining content domains and types that are of value to search engines and the sponsor companies' search based properties.

Unlike most other semantic web ontologies, Schema.org is continuously developed, extended, and modified. As Schema.org is an effort to bolster search and search based artificial intelligence, it takes a general approach to content coverage. While there are other broad form ontologies, Schema.org is among the largest and most used, covering over 10% of existing web content, as it is employed on over 10 million websites and integrated into many applications including, but not limited to, the *New York Times*, *The Guardian*, IMDB, Monster.com, Pinterest, LinkedIn, Yelp, Zillow, and the Google, Microsoft, Yahoo, and Yandex search based ecosystems (Guha 2014). As of writing, Schema.org has 258 contributors from private industry, non-profit standards groups, academia, and independent origins. Currently, the core vocabulary contains 583 types, 846 properties, and 114 enumeration values. The core vocabulary covers a wide range of content including, but not limited to: creative works, events, people, places, products, organizations, and actions. In addition to the core vocabulary, Schema.org also has three[15] distinct hosted extensions – automotive, bibliographic, and health-life science – one fully

[15] There are actually five hosted extensions, but two of those extensions are tools for the development process and not extensions for website developers to use for content markup

developed and mapped external extension – GS1 – and one extension in the process of being mapped and considered as an extension – Wikidata[16]. Within these hosted extensions there are an additional 114 types, 217 properties, and 153 enumeration values. In addition to those operational extensions, there are a number of other Schema.org community groups that are working on extensions to the core ontology. These include such content domains as, educational courses, archival information, tourism, finance, sports, legislation, and meat[17]. In addition to creating semantics for simple Web content, Schema.org is also beginning to branch out into semantically linked relational databases, the Internet of Things, and civic participation, though these aspects of the project are in early stages of development.

Schema.org is organized into two primary groups. The first is a steering group comprised of paid employees of the sponsor companies, a few independent contractors, and a representative of the W3C. The steering group oversees the development of the project, as well as the approval and implementation of new proposals. The second, community group, is a group open to any interested persons, who's role "is to propose, discuss, prepare and review changes to Schema.org, for final review and publication by the steering group." (Schema.org 2017b). One member of the steering group chairs the community group, and all members of the steering group participate in and monitor the developments and discussion in community

[16] There is an ongoing process of mapping Schema.org and Wikidata to one another. Currently, there are debates occurring within the two communities to consider Wikidata as an official extension of Schema.org. Thus far there is no resolution on the proposal and the individuals involved are unsure whether or not to consider it an external or hosted extension.

[17] Of note is that many community members report being involved in creating their own external extensions. The details about these extensions are not readily available and their development is not recorded or publicized for scrutiny. These external extensions, much like the GS1, happen outside of the purview of the broader Schema.org community.

group. Apart from the paid employees of the sponsor organizations, contribution to Schema.org is voluntary. However, many of the frequent contributors are consultants who are paid to implement semantic markup for their clients, and so while their contributions are formally voluntary, they are often in service to their paid professional work.

While all large Web-scale technical projects like Schema.org rely heavily on a vast array of standards and protocols, it would be counterproductive, and potentially impossible to explore the full range of them in this study. Discussions of many of these and other protocols, standards, and their effects are discussed in detail elsewhere (Abbate 2000; Coyle 2008; DeNardis 2009; Galloway 2004, 2006). However, there are standards that are of central importance to the development and use of Schema.org that I will discuss in detail. These markup standards all apply the RDF data model, or something that approximates it to varying degrees of sophistication, and enable the Schema.org ontology to be applied to Web content. Thus, they are central to the process of enacting semantic data. Schema.org supports the use of three markup standards in particular, Microdata, the Resource Description Framework for attributes (RDFa), and Javascript Object Notation for Linked Data (JSON-LD). This marks a point of departure from other semantic web ontologies, as most do not actively specify a preferred standard or set of them.

4.2 Research Sites and Sources

My research took place in three primary sites, the W3C community message boards, Schema.org's GitHub repository, and Schema.org's own website. In addition to those three sites, I conducted additional research on the three markup standards which included the W3C

specifications pages for Microdata, RDFa, and JSON-LD. While not research sites, per se, these documents were instrumental in understanding not only how the project is implemented, but the affordances and constraints presented by these technical standards. Below, I will describe each site and its operation in additional detail.

The W3C message boards are the public discussion forum for the Schema.org project. They are publicly accessible from the both the W3C's website and Schema.org's main page. The message boards are similar to an email list server, recording each post and response in a threaded fashion. The message boards are used for the proposal, discussion, and debate of "all updates, changes, and improvements to Schema.org" (Brickley 2016). However, in practice, the message boards also include large amounts of questions and answers from confused or uncertain Web developers who are trying to apply the Schema.org vocabulary. Additionally, the detail and sophistication of the discussion and debate that happens on the W3C message boards is far less active and detailed than the what happens on GitHub. Often times the discussion and debates on the message boards are simply a springboard for inclusion on the project's GitHub site. The community group chair[18] is responsible for managing the workflow between the W3C message boards and GitHub, but in practice, other steering group members and core contributors do the same.

My main site of research for this project occurred on GitHub. GitHub is a platform for coordinating software development projects amongst multiple people, that draws on the distributed version control system, Git. It allows for distributed and non-linear work and is particularly useful for hosting and developing open source and peer-produced code systems.

[18] Also a steering group member in Google's employ

Like many professional disciplines, code work on GitHub has its own specialized jargon that will be important to have as a reference going forward. Such a reference is available in the appendix. Among the platform's many features, there were some that were particularly important to this project. These features included, housing project documents and the master repository, project Wikis, labeled issue tracking, feature and pull requests, commit histories, and workflow documentation.

GitHub is the site of most of the work on Schema.org. While many proposals originate from the W3C message boards, many bypass that forum and are proposed directly on GitHub. Moreover, all actionable work, regardless of where it is proposed happens and is discussed on GitHub. Each work item commit and discussion is logged, linked to other related issues, and tagged with metadata on the GitHub page. The type of participant on GitHub displays a higher degree of understanding and sophistication regarding semantic web standards, ontology development, and markup. Communication and work here are generally from a smaller core group of a few dozen community members.

While GitHub houses the complete documentation for Schema.org, the project's own website (Schema.org) contains much of that information in a more easily digestible form. Likewise, GitHub contains the actual vocabulary, but this file is designed for machine consumption, which does not lend itself to human investigation. Schema.org's website contains the entire vocabulary in an accessible and interlinked form. This allowed me to traverse type, class, property, enumeration connections with relative ease. With this I could see the full range of properties, expected values, and instructions relating to any given markup. Additionally, the project's website has an unpublished mirror of itself (pending.schema.org) that allows one to

view pending changes before they are actually added to the vocabulary. Both sites also contain a detailed chronological list of additions and modifications to the ontology along with links to the specific issues on the GitHub issue tracker. Lastly, the site contains a number of helpful resources for Web developers on using and understanding markup and the Schema.org workflow.

My final data for investigation came from the specification sheets for the three markup standards supported by the project. While not all three standards are all W3C recommendations, the W3C contains the specification sheets and best use practices in their Web domain. These specification sheets define the basic terminology for each standard, provide a detailed overview of their syntax and data models, instructions for use, how the syntax is interpreted by computational agents, their general affordances and constraints, and their relationships to other markup standards. Problematically, they suffer from a high degree of technicality and as general standards and specification documents, they do not refer specifically to their use with Schema.org[19].

4.3 Data Collection and Preparation

Data collection took place at two different junctures. The initial collection stage began in September 2015 and took place over the course of two weeks. During this time, I manually scraped all posts organized monthly, by subject, in threaded form, from the W3C message boards. The W3C message board data dates as far back as April 2015 when all community group

[19] In the JSON-LD development community group, there is substantial overlap between the developers for the syntax and the core community members of Schema.org. This has had the effect of making JSON-LD align well with Schema.org's needs.

work on the Schema.org project shifted to their servers. Scraping the posts in threaded form allowed for the best way to keep track of posts and their direct responses. However, since the message boards are a continuously operating work platform, there were many instances where a response to a thread happened in a different month than the thread originated in. This was particularly the case near the beginning and end of months. Where applicable, those responses were added to the original discussion. This left me with a series of documents that tracked the history of all discussions on the community boards in order. In total this accounted for 278 distinct messages across 66 subjects. The initial phase of data collection from GitHub was similar to that of the W3C message boards. Schema.org's GitHub page allows one to view all issues, regardless of metadata tags, in both open and closed statuses. Much like the W3C data, I manually scraped all content from each thread, then organized them by subject and issue number. GitHub extended further back in time than the data from the W3C, with coverage beginning in April 2014. In total this accounted for 345 distinct issues. The W3C markup specifications were also scraped into their own separate documents. Data from Schema.org's website was referenced in situ to preserve the benefits offered by the digital objects embedded in both the ontology and the history of releases.

Following data collection, all textual data was stored in both an encrypted portable flash drive and on Dedoose's servers. Once uploaded to Dedoose, I added metadata tags to capture the source, the number of participants, and the length of discussion. During this process, I also added the metadata tags that were originally included in the GitHub data but were lost when the data was first scraped. These tags designate the type of issue are assigned by the steering group members. They include administrative, extension tracking, for Steering Group attention,

guidelines documents and examples, standards and organizations, vocabulary, rough, small, exact, and large proposals, and tricky problems among others. Next, I marked each issue by its status as open or closed.

To capture the ongoing work on the project, as well as the improvements, changes, and new releases since my initial collection phase, I completed the second phase of data collection in July 2016. The process of collection was similar to the initial phase, except for the added complication of there being additions to the already collected material. For the W3C data, this was not an issue because the message board's platform allows sorting by date. While GitHub also allows sorting by modification date, it forced me to append each original issue that had recorded changes from the initial data collection phase. This meant that many of my originally captured issues were closed or merged with other issues. Fortunately, GitHub readily provides that information so making the amendments was relatively simple. This second phase of data collection brought my data totals to 682 W3C messages across 164 subjects and 971 distinct GitHub issues.

While both platforms are publicly open, contributors are primarily Web developers looking to apply Schema.org to their sites or prominent community members contributing to the day-to-day development and operation of the project. Nearly all participation is from Western collaborators, the vast majority coming from people located in the United States, England, and Germany. There is substantial variation in the length of issue threads and message board chains, as well as in how detailed those data are. They range from zero to forty-two responses, averaging six responses per issue, with each response at a paragraph or more in length and often including lengthy examples of existing or proposed Web markup.

Code and Web markup were not scraped and uploaded to Dedoose but were included in the analysis. On GitHub the specific changes to a code are highlighted in the code itself, most of which is too large to be manageable in Dedoose, while the important data for my analyses are just these highlighted segments, removing them from the context of the larger code set removes potentially important information. Additionally, this particular way that code is formatted and presented is important to preserve because the specific formatting and characters used to display and organize scripts serve important purposes in programming languages. Copying these bits of information to a text document compromises that formatting and at times the characters involved. Fortunately, all of this information is preserved in the correct form on GitHub.

4.4 Analytic Approach

This study adopted two main methodological techniques. The first followed the netnographic approach developed by Kozinets (2010) on conducting digital ethnographies of online communities. The second methodological technique was a variation on multimodal critical discourse analysis. The first was used to investigate the development and use of Schema.org, and the second was used in combination with the first to understand the affordances and constraints imposed by the three markup standards.

Ethnography has a long history in the study of science and technological systems. Ethnographic laboratory research emerging from the field of science and technology studies (STS) have shown the central roles of communities of practice, technological artifacts, and their intra-actions in the production of scientific facts (Knorr-Cetina 1999; Latour and Woolgar 1986;

Latour 1987; Mol 2002; Traweek 1988). Additionally, similar methods have been used to study the development of expert systems by early AI researchers (Forsythe 1993a, 1993b; Suchman 1987). These studies have been instrumental in showing the ways that the seemingly black boxed inner workings of scientific, medical, and computer science communities come to produce their objects in active and dynamic ways (Knorr-Cetina 2001; Mol 2002). Moreover, the ethnographies that have come out of the STS field, in particular, have given us license to interrogate the intra-actions between human and non-human actors with the same levels of priority while assuming the same levels of potential agency. While research setting and cultural artifacts have always been a central component of ethnographic research, the affordance-based perspectives that are both explicitly and implicitly adopted in contemporary ethnographies provide a stable footing for understanding the ways that communities of practice interact with technologies in varied, changing, and emergent ways. Ways that are inseparably connected to the socio-material contexts in which those intra-actions occur. This form of agentive realism reflects an ontological and epistemological commitment to take the two as unsettled and incomplete achievements (Barad 2007; Mol 2002).

Now old discussions of digitally mediated interactions enabled by ubiquitous computing and the increasing role that digital media play in community creation, have coincided with digitally oriented forms of ethnography. Often termed, digital ethnography, virtual ethnography, or cyber-ethnography, these methodological approaches mirror the grounded concerns of lived experience, interactions, gestures, and situated action of traditional ethnographies. Research in this area has studied the construction of identity in online communities, digital interactions as extensions of face-to-face interactions, impression

management in social media, and the construction of publics in discussion forums among other aims and digital contexts (Boellstorff 2010; Hine 2000; Markham 1998; Springer 2015; Turkle 1995).

While many of these digitally inspired ethnographies view the digital as only one component of the total social experience, this study orients differently. Natively digital artifacts, like software and code systems, require a focus that is also natively digital, particularly in the present case where the boundaries of practice are coterminous with the boundaries of the digital (Rogers 2013). These types of digital technologies and their interfaces reconfigure social interactions and "this reorganization changes how empirical knowledge is produced and transmitted among users in a digital environment. It also restructures how we as ethnographers observe, analyze, and categorize empirical knowledge." (Hsu 2014:1). Thus, to study a natively digital site, such as GitHub, one must adopt a methodology that is sensitive to these reconfigurations that mark distinctions between online communities and communities online (Kozinets 2010). While this study recognizes the false dichotomy of material/immaterial that often underpins sensationalized studies of digitality, this study follows Hsu (2014) and focuses on the digital as a pragmatic decision for the sake of methodological and analytic lucidity. However, this comes with the acknowledgement that work occurs only through the benefit of a wide-ranging set of material artifacts and practices.

Kozinets (2010:60) defines the practice of netnography as, "research based in online fieldwork. It uses computer-mediated communications as a source of data to arrive at ethnographic understanding and representation of a cultural or communal phenomenon." Netnographic practice allows for a situated analysis of community engagement available in

online forums. While I abstained from participation to avoid being disruptive, netnographic research, like traditional ethnographic methods, necessarily includes a range of techniques including archival and documentary analysis, the tracing of event timelines and narratives, and in-depth case study all as deployed here. Springer (2015) argues that this provides a set of observational data similar to, but more reliable than field notes and artifacts from traditional ethnographic methods. It additionally provided an increase potential for researcher reflexivity, as I was able to easily revisit data in its original pre-coded form (Springer 2015).

As netnography is a necessarily diverse set of methods, my approach to the data was also diverse. The first stage in my research process was to familiarize myself with the community and their sets of practice. This involved the daily monitoring of both primary research sites to situate myself in the day to day operation and development of the project. This occurred in two ways. First, I constructed a running set of field notes that described the situation on the ground in its own terms. This involved a simple summary of what happened and why based on the information directly available. Given my interest in the specific ways this system is brought into being, I was focused on how new developments were proposed, who proposed them, why they proposed them, as well as the process and results of the debates that surrounded any given issue. Of crucial importance was uncover the rationales behind both the initial proposals and proposed solutions.

Once familiar with the process, I was able to look back at older data and make sense of the issues, discussions, and resolutions of the past. By systematically following the thread of debates, proposed and acted upon solutions, commits, and interactions between the assemblage's components I was able to construct a working history of the project from the

perspective of work practice, rather than project releases, and one that contained multiple timelines based on specific concurrent issues. However, having the timeline of actual releases at hand was helpful in determining the types of issues that make it to the release stage, and the types of issues that get stuck in production for whatever reasons. This allowed me to uncover patterns in work priority, coverage areas, treatment of certain forms of proposals, rationales for decisions, and the power dynamics of the immediate community. Importantly, it also allowed for reoccurring issues to emerge in the data in a more meaningful way. The construction of a narrative timeline allowed these problem areas to emerge through reoccurrence over time and also through tracking issue assignments and mergers.

In addition to the common ethnographic practices surrounding field notes, I engaged in a more formal textual analysis that followed the approaches of Berg (2004) and Strauss and Corbin (1990). Continuing from my desire to approach the case in a grounded manner, I began this coding process by establishing codes derived from the language of the actors themselves. This had the benefit of providing a view towards the meanings that actors gave to their actions. At this stage, I had read all of the data so I was able to look for the linkages between data points (threads, issues, commits, etc.). As a step in the data analysis process, I allowed in vivo codes and categories, as well as previously undefined narratives and codes that relate to my research project, but were unspecified, to emerge from the data. Much of these codes related to efficiency, pragmatics, and coherence across the applied data model. Next, I established grounded categories where I placed this material into context with the narrative and action specific conditions of the texts – e.g. limiting complexity because of across standard variations or intentional ambiguity to account for missing coverage. In this stage, a number of unexpected

categories and themes emerged. Notably, these included the tendency to opt for simplicity over semantic accuracy and decoupling the context of creation from the contexts of use.

The second stage of the research involved revisiting selected data in light of sociological constructs drawn guided by the themes of my research and the concerns of the relevant literature. Here I reconstructed my "field notes" in a way the a drew out the ontological issues of over-determination and occlusion, as well as other issues of position taking, representation, and negotiating affordances of the technological environment. Much of this portion of the research involved a close critical reading of the ontology and the markup standards used by Schema.org. This study takes the view that the ontology and the relevant markup standards are forms of computer language based digital objects. Kallinikos et al. (2013) offer a useful definition of the concept. Digital objects, they argue, are editable, interactive, reprogrammable, and distributed digital cultural artifacts. As cultural artifacts, they contain the implicit assumptions, norms, and values that exist in the immediate cultural landscapes from which they emerge (Lupton 2014). As language-based standards they contain the ability to shape the means for representation, everyday practice, and our interactions with them (Busch 2011; Fiormonte et al. 2015). As mentioned earlier, however, these digital objects are not deterministic, nor are the entirely within the realm of free play. Like all cultural artifacts, standards contain certain affordances and constraints that shift and reshape as contexts change. Additionally, as cultural artifacts, they can be interpreted as texts.

To analyze these digital objects, I adopted an approach frequently used to study affordances of technological artifacts (Introna 2005, 2014; Introna and Hayes 2011; Lupton 2014). This approach has similarities to multimodal discourse analysis (MDA), a recent

methodological technique that allows the researcher to investigate meaning in objects beyond the simple confines of language. It takes other objects like images, symbols, actions, audio, and I argue code based markup standards as semiotic resources (O'Halloran 2011). MDA has been applied in a number of different areas to articulate the meaning potential in visual design (Kress and van Leewuen 2006), artwork (O'Toole 1994), mathematics (O'Halloran 2005), and televised news (O'Halloran 2011) among other areas. The specific approach of MDA is a similar approach to the approach used in other forms of discourse analysis that assume:

> that discourse, subjectivity, and practice are densely interwoven, and that discourse is primary to subjectivity/practice through its constituting or framing powers. This means that dominant and widespread discourse shapes both how to talk about a subject matter and the meaning that we develop about it… It is not the details of the account and its context as much as a perceived general tendency that is deemed significant to use (Alvesson and Karreman 2000:1138)

This approach allows a researcher to identify and interrogate the ways that broad discursive patterns help to shape affordances and prohibitions, the ways of being and the constraints that might be placed on being, the ways of doing and the constraints placed on practice, and the ways of representing and the constraints on possible claims making. The constant conversation that I had between the data in my field notes and the affordances of the ontology and standards allowed me to understand specifically, how certain forms can and cannot be represented, how representations can serve to over-determine or occlude others, and how the ability to make semantic data as a practice is enabled and constrained by the particular forms of the standard and ontology drawn on in a given context.

As a part of this second stage of research, the textual analysis involved coding texts based on constructs and questions guided by themes relevant to my research questions. Given my primary focuses the interactions between components and human-machine translations, I

reread the data in light of these focal points. Here I identified statements that dealt directly with negotiations with markup standards, divergent interpretations of semantics and their application by developers and users, design best practices, and decision making and rationales. These additionally included issues surrounding conflict, consensus, and proper classification, which itself was broken down into instances where false equivalencies were made and recognized, where properties, representations, and identities were conflated, and when instances evaded proper classification within the ontology. This phase also included coding of the actors and components involved in the assemblage. Here I coded actors by reported country, affiliation, and role. Additionally, I coded actors' stated involvements with other ontologies.

The final stage of the formal coding process involved drawing connections between codes from both stages of the coding process. In this phase, I modified and collapsed codes where appropriate and the placed them into more direct conversation with the study's research questions and the contexts they implied. This allowed me to draw connections between themes that may have initially been treated separately and apply coded information directly to the study's focal interests. For example, that usability and practicality took precedence over accuracy in representation and classification when divergent interpretations of semantics and their application occurred between participants.

4.5 Researcher Reflections and Limitations

Such and undertaking has ethical implications that must be considered and dealt with accordingly. This research was conducted in full consideration of the American Sociological

Association's (ASA) code of ethics and complied with all Internal Review Board (IRB) prescriptions. This document follows all ASA recommendations for professional competence, integrity, professional and scientific responsibility, respect for people's dignity, and social responsibility. As such, I undertook only the activities in which I was professionally competent. My multiple statements of research were clear and honest in intention, and despite the at times, critical nature of this study, I do not seek to endanger my research subject's professional well-being. Throughout the research process, appropriate measures to protect research subject's confidentiality and anonymity, despite the public nature of the two primary research sites that were used. Additionally, all research subjects were informed of their right to be omitted from the research, in accordance with ASA and IRB guidelines (American Sociological Association 2008).

This study is about the process of constructing and enacting data as well as the creation of ontological states through human-machine translations. As such, the irony of conducting qualitative research and making specific enactments of data to solve my questions is not lost on me. The same biases, rationales, and patterns of interactions that I attribute to my case as an ostensibly clairvoyant outside body can be alleged to this very investigation with equal legitimacy. As I claim that Schema.org's ontologists are enacting a particular ontology, I am doing precisely the same thing. This illusion of purity could be said of all social science research particularly when that research is engaged with knowledge work (Latour 1991).

To mitigate the bias introduced by my personal choices of questions, case, and methods, I am left with only a few possible options. The first of which was to follow the strategies of the researchers before me. My interest in this study follows from an established body of research,

drawing on theories, concepts, and approaches that have withstood the scrutiny of peer-review. This work has shown that researcher reflexivity is key to undertaking ethical and valid social research. The second way I mitigated my own personal biases was to allow the data to speak in their own terms. In this sense, I attempted to ground my research in the site and community where it occurred, drawing on their language, and their understandings of what they were doing before adding my own. Of course, this strategy is not fool-proof and still involves interpretations and assignations of cause and rationale on the part of the researcher. To address this nagging concern, my third strategy was to allow the data to speak for itself as much as possible throughout this document and my analysis. To do so, context will be supplied by both description and data where possible. While I interpret the data, and draw conclusions from it, I have tried to make those interpretations and conclusions follow logically from the data I present.

The relative lack of social science research on the semantic web and on data enactments suggests that a grounded empirical analysis is needed. However, as a case study, this research only investigates one semantic web ontology and one main form of data, despite its various enactments. While the case under investigation is among the largest and most used example available, this has the potential consequence of limiting some generalizability. However, I address the specific features of Schema.org that make it unique and contribute to its specific importance as a case for research and increasingly representative of the semantic web in general. Additionally, this analysis is not a comparative analysis on the enactment of other types of data, which are part of other data assemblages, so will necessarily have different enactments. However, this concern is marginal to this study as it represents one small step in

the emerging study of data. One contention of this and other work (Kitchin 2014b, 2014c; Kitchin et al. 2015a, 2015b) is that disparate types of data and their assemblages need to be studied in their own right.

Additional limitations are presented by the data available to this study. While the data used contain all of the public discussion on how to best alter and apply the ontology, it does not contain any backchannel communications either in person or via email. The data indicates that there are occasional back-channel discussions between a few community members on certain issues. Most of the time those discussions are disclosed to the group with a summary of the results, but the actual content of the discussion is lost. This may be fine for continued work on the project from the perspective of a developer, but from an analytic perspective, there could be information important to my findings, for or against. Additionally, since the data are entirely textual and interviews were not performed for this study, there is no possibility for follow up or clarification questions that are not in the text. The effects of these limitations seem negligible, however, since all modifications are public and the data of primary interest are the discussions and debates about the ontology's coverage.

5. Negotiated Standards

In *Art Worlds*, Howard Becker (1982) famously, shows the ways in which material standards constrain the production and display of artwork. Among other things, he shows that building standards impact the size of artwork as well as how and where it can be displayed. Likewise, Klose (2009) explains how the adoption of the Twenty Food Equivalent Unit for shipping containers has helped to not only shape the course of globalization but also the way we now think about organization and modularity. Furthermore, we know that classification standards work to afford certain ways of knowing a thing while constraining others (Bowker and Star 1999, Epstein 2007, Scott 1998). This work shows that while standards often exist to reduce user friction and to make things interoperable, they do so at the expense of competing understandings and states. The investment in, and deployment of, a specific set of standards helps to set the boundaries of practice and its results as those communities must work within the affordances set by those standards.

We also know that the creation of standards and the process of adopting them is fraught with conflict. Much time, money, and effort go into creating and adopting ascendant standards, and for creators and early adopters, there may be substantial costs associated with switching between them. Work has shown that the standard that prevails is not always the most technically proficient, capable, or compatible standard, as a wide variety of external forces affect their adoption and use, as was the case with Betamax and VHS (Genschel 1997; Yasunori and Imai 1993). Additionally, standards development and deployment does not happen evenly across contexts, as certain actors and domains are either not exposed to, or reticent to adopt standards (Noble 1984).

More recently, work has examined cases where standards do not work (Bowker and Star 1999; Epstein 2007, 2010; Heimer 2001; Whooley 2010, 2014; Whooley and Horwitz 2013). Even when accounting for a standard, or set of standards', propensity to afford and constrain, or their bumpy and contested development and adoption, the accepted standards do not always work as supposed or desired. In these cases, a standard, or set of standards, actually may work to inhibit practice and its smooth operation, or to constrain results in ways that work against the standard's supposed function. While Bowker and Star (1999) argue that standards tend to stabilize systems, rather than just tending towards stabilization, they acknowledge that effects can be varied as standards are artifacts to be negotiated. While work in neo-institutionalism repeatedly show that decoupling is the most common outcome when standards conflict with practice, this work lacks due attention to when decoupling is not possible (Meyer and Rowan 1977; Sandholtz 2012; Thévenot 2009; Zbaracki 1998). These omissions beg researchers to situate standards' use in practice and question how these blockages are negotiated and how the restrictions created by standardization are relaxed while still enabling the interoperability they are designed for. While scholars have been exploring the tensions between organizational level practices and standards' homogenizing effects for some time, work that both locates standards at the level of professional communities and that investigates the negotiations of barriers is only just begging to emerge (Epstein 2007, 2010; Halpin 2016; Heimer 2001; Whooley 2010, 2014; Whooley and Horwitz 2013).

The following section engages with these omissions. Below I will show how the markup standards described in previous chapters are deployed across the two primary fields of practice implicated in Schema.org. In line with existing work on standards and classification, I will show

the ways in which they set certain affordances and constraints on the development and use of the Schema.org ontology and thus the affordances and constraints placed upon the enactment of the data assemblage. I will also show the ways that these standards work against the thing they are supposed to enable, and how the communities of practice negotiate those restrictions. I echo recent work, showing that practitioners work around those limitations, but also extend that work arguing that most often, practitioners work with and through the standards in their negotiations (Berg 1999; Epstein 2007, 2010; Heimer 2001; Whooley 2010, 2014; Whooley and Horwitz 2013). Furthermore, I will detail some of the unintended consequences of those negotiations and how they can lead differential enactments, or ontological states.

Specifically, I will examine the three markup standards adopted by Schema.org and its users. Each standard plays a major role in shaping the way that the Schema.org ontology is developed and applied in actual use. Importantly, each standard conditions that development and use in subtle but different ways that constrain and enable the enactment of the assemblage and the way that information can be encoded for machine readability. This chapter details those affordances and constraints in both the general semantic web and Schema.org specific contexts. These markup standards are built on different computer languages with each containing different grammatical structures and syntactical rules regarding their use. As many of the affordances and constraints that these markup standards put onto ontologists and Web developers are rooted in these technical grammars, this chapter pays close attention to the details of syntax as they serve to create protocological control (Galloway 2004, 2006). The consequence of this is that the ensuing discussions are by necessity, technical. Much of the affordances and constraints discussed will reference specific artifacts of these markup

standards that evade simple description and so each standard will be analyzed via reference to empirical examples from figures in the text. Following a technical discussion of the standards, this chapter introduces some of the specific ways that those technical features affected Schema.org's development and use, as well as some unintended consequences of those effects. While there are three standards in use, much of my attention is paid to the two standards that feature most frequently in the data and are most important to Schema.org, Microdata and Javascript Object Notation for Linked Data (JSON-LD), the former being the previous preferred markup syntax and the latter the current preference.

5.1 Microdata

Microdata is a syntax that allows machine-readable data to be embedded in Hypertext Markup Language documents. It provides a way to annotate Web content with sets of name-value pairs in sequence with the rest of the HTML document. This means that Web developers can encode machine-readable data in the same segment of the HTML script as the non-semantic content. They do not need to insert additional code or append documents in any way beyond the Microdata markup. However, this implies that Microdata is being used at the onset of page development, as already existing HTML script would need to be amended, sometimes with significant difficulty, to include the machine-readable content. Even still, the machine-readable content can be added in parallel to the existing HTML script (Hickson 2013).

Figure 6. Microdata Markup of a Government Permit

```
1.      <div itemscope itemtype="http://schema.org/GovernmentPermit">
2.        <span itemprop="name">NYC Food Service Establishment Permit</span>
3.        <div itemprop="issuedBy" itemscope itemtype="http://schema.org/GovernmentOrganization">
4.        <span itemprop="name">Department of Health and Mental Hygiene"</span>
5.        </div>
6.        <div itemprop="issuedThrough" itemscope itemtype="http://schema.org/GovernmentService">
7.        <span itemprop="name">NYC Food Service Establishment Permit Service</span>
8.        </div>
9.        <div itemprop="validIn" itemscope itemtype="http://schema.org/AdministrativeArea">
10.       <span itemprop="name">New York</span>
11.       </div>
12.       <time itemprop="validFor" datetime="P1Y">1 year</time>
13.     </div>
```

Source: Schema.org (2017d)

Name-values pairs are akin to a set of tags that can be attached to Web content where a group is referred to as an item, or *itemtype* in syntax, and where each name-value pair is a property, or *itemprop*, and the context for the description is supplied by *itemscope*. The empirical sample in Figure 6. provides an illustrative example. In Figure 6., Schema.org is being used to describe “A New York City Food Service Establishment Permit” as issued by the “Department of Health and Mental Hygiene” and in effect for one year. We can see the *itemscope* and initial *itemtype* being set by “http://schema.org/GovernmentPermit” (Line 1), which is being used to provide the markup’s context by specifying the ontology’s location, the class, its supra-classes, its properties, its inherited properties, properties for which instances of “GovernmentPermit” can appear, as well as relevant definitions, and the affordances and constraints on value inputs among other things. This canonical IRI refers a computational agent to “http://schema.org/GovernmentPermit” for the context in parsing the ensuing markup. The entire ability for an agent to understand the marked information is delimited by what appears in that context specification. In other words, the realm of possible understanding is bound by the ontology specified. Here the assembling of the Microdata standard and the Schema.org

ontology combine to enact the meaning space for food service permits, government permits, administrative areas, etc.

The first name-value pair, “name – NYC Food Service Establishment Permit” (Line 2) establishes a description of the permit in plain text. Note, that while a computational agent will understand that the “name - NYC Food Service Establishment Permit” pair is referring to the “http://schema.org/GovernmentPermit” IRI and all its associations, in this markup it does not understand any of the content found in the description since it is just a text string, it merely presents that as part of the HTML script. However, this description could itself be subset and marked up to indicate things like “Food Service” and “New York City (NYC)” as further semantic detail specifying the exact context and form of the permit issued, though only by a single specified type and through a single specified context file. The decision to do so depends on the aims and skill of the Web developer and the availability of coverage in the ontology.

Within the same *itemscope* and *itemtype*, a subset with a new *itemprop* “issuedBy” is defined by a separate *itemscope* and *itemtype* “http://schema.org/GovernmentOrganization” (Line 3), which itself has a new property, “name”, establishing that the government permit of the previously established name was issued by the government organization “Department of Health and Mental Hygiene” (Line 4). The markup then begins a new code loop under the original *itemscope* and *itemtype* to establish the property “issuedThrough” with yet another new subset scope and type “http://schema.org/GovernmentService” (Line 6) with the property “name” and the value “NYC Food Service Establishment Permit Service” (Line 7). This process is further repeated to describe the valid administrative area (Line 9) of New York (Line 10) and the one-year validity (Line 12). This process can be repeated and further elaborated on for any

content that has corresponding coverage in the ontology. Problematically for Microdata, as discussed more below, this process can only be done within the bounds of a single ontology and type specification.

Microdata was created for human readability and webmaster usability and emerged as a markup syntax aimed at providing structure to the World Wide Web. While not a W3C recommendation, meaning it lacks the institutional support and development of other standards, Microdata is part of the HTML5 specification so that any parser that can understand HTML5 can understand Microdata markup, making it broadly applicable. Additionally, its embedding in HTML5 increases the likelihood of Web developers being familiar with its basic structure, adding to the chances that they apply markup to their content. Indeed, this is one of the stated reasons for Schema.org's initial support for Microdata, and likely to be a significant part of the reason for its continued popularity despite its often bemoaned limitations (Guha et al. 2015, Schema.org 2017c).

Microdata benefits from its similarity to HTML, its human readability, simplicity, and broad support by browsers and Web content consumers, however, it has a number of limitations for creating semantic data. First, Microdata can only be used in HTML5, which precludes its use in any other computer languages, such as Javascript, which commonly used to add interactivity to websites or browser specific functionality. This is significant because it limits the applicability of ontologies, restricts the possible amount of semantic data, and blunts a machine's understanding of the Web domain. Additionally, Microdata was specifically designed so that content is structured using the context of a single ontology, making it difficult to integrate different ontologies. While workarounds exist, their use is rare and they add

substantial complexity and run the risk of creating competing sets of data for different consumers, as some may not be compatible or able to read the content included in the workaround. This limitation creates the issue where a Web developer has the choice between semantic consistency and semantic hegemony. Drawing on a single ontology has the benefit that its class/type/property configurations will likely be in agreement. In the other case, a developer is bound by the specific construction of the world as found in the adopted ontology. So, while you may consistently describe your domain, the limits of how that human-machine translation occurs is bound by the chosen ontology. Relatedly, Microdata does not accept multiple types for a single object across ontologies, or trans property use across types in a single ontology, further restricting its classification within a single class/type/property matrix. This means that an artifact with ambiguous, but equally valid, semantics will be ontologically over-determined in Microdata markup (W3C 2011, WHATWG 2016). As the example above indicates, each listed *itemscope* is bound and determined by the single deployed *itemtype.*

Another limitation is that the Microdata specification does not support inverse properties. Not being able to natively reverse property relations means that an agent's ability to capture relationships is constrained to only one direction. In the example from chapter two, an agent would not be able to understand that Barack Obama was Donald Trump's predecessor, only that Trump was Obama's successor. To create inverse data for use in the Microdata syntax, ontologies must be specifically designed with pairs of properties to capture reverse relationships creating additional problems for Web developers and ontologists alike. Generally speaking, such additions and modifications run counter to the ontologists' preferences, as they do not like adding extra complications to the ontology. However, such

inverses can be important signals of directionality and intent. They currently advocate for using alternate markup standards or the experimental *@itemprop-reverse* syntax, though there are disagreements in the Microdata community about using syntax that is not a part of the official specification. Not only do these affordances and constraints affect the development of ontologies and the in-situ markup, but they also contribute to forms of semantic hegemony whereby ontological states are applied to semantic data in the form of a single context file and/or Web content is constrained to a single type despite having an indeterminate nature. These types of semantic hegemony detailed above can include the two forces of ontological over-determination and ontic occlusion outlined in earlier chapters.

5.2 RDFa

The Resource for Description Framework in Attributes (RDFa) is similar to Microdata in that it is a markup syntax designed to convert a Web environment initially designed for human consumption into one consumed by machines (Herman et al. 2015). Much like Microdata, RDFa can be implemented in parallel with HTML scripts to provide structured semantic data to Web content, only it does so with different sets of values to designate attributes. Schema.org supports RDFa but only a limited version called RDFa Lite. This is due to the level of complexity the RDFa Core standard has for developers, as it would run counter to the stated approach of the Schema.org developers.

Figure 7. RDFa Markup of a University Alumnus

```
1.      <div vocab="http://schema.org/" typeof="Person">
2.        <span property="name">Delia Derbyshire</span>
3.        <link property="sameAs" href="http://en.wikipedia.org/wiki/Delia_Derbyshire"/>
4.        <div property="alumniOf" typeof="OrganizationRole">
5.        <div property="alumniOf" typeof="CollegeOrUniversity">
6.        <span property="name">University of Cambridge</span>
7.        <link property="sameAs" href="http://en.wikipedia.org/wiki/University_of_Cambridge"/>
8.        </div>
9.        <span property="startDate">1959</span>
10.     </div>
```

Source: Schema.org (2017e)

Figure 8. RDFa Markup of a Government Permit

```
1.      <div vocab="http://schema.org/" typeof="GovernmentPermit">
2.        <span property="name">NYC Food Service Establishment Permit</span>
3.        <div property="issuedBy" typeof="GovernmentOrganization">
4.        <span property="name">Department of Health and Mental Hygiene"</span>
5.        </div>
6.        <div property="issuedThrough" typeof="GovernmentService">
7.        <span property="name">NYC Food Service Establishment Permit Service</span>
8.        </div>
9.        <div property="validIn" typeof="AdministrativeArea">
10.       <span property="name">New York</span>
11.       </div>
12.       <time property="validFor" datetime="P1Y">1 year</time>
13.     </div>
```

Source: Schema.org (2017d)

Figure 7. and Figure 8. show two examples of RDFa markup. Much like the Microdata example in Figure 4., both RDFa examples begin by specifying the context from which each object draws its information. Note that in both figures the markup draws on "http://schema.org/" for this context. Figure 7. describes a type of person (Line 1) named "Delia Derbyshire" (Line 2) who is the *sameAs* the person referenced in the Wikipedia entry at the given URL (Line 3). This person is an *alumniOf*, which is defined as a type of *organizaitonRole* (Line 4), with the specific designation that Delia is an *alumniOf* a specific form of organization "CollegeOrUniversity" (Line 5) that has the specific name "University of Cambridge" (Line 6)

which is the *sameAs* the referenced Wikipedia entry (Line 7). And finally, it specifies that Delia was an alumnus of Cambridge in 1959 (Line 9).

Figure 9. RDFa Markup Using Multiple Ontologies

```
1.  <p vocab="http://schema.org/" prefix="ov: http://open.vocab.org/terms/" resource="#manu"
    typeof="Person">
2.   My name is
3.   <span property="name">Manu Sporny</span>
4.   and you can give me a ring via
5.   <span property="telephone">1-800-555-0199</span>.
6.   <img property="image" src="http://manu.sporny.org/images/manu.png" />
7.   My favorite animal is the <span property="ov:preferredAnimal">Liger</span>.
8.  </p>
```

Source: Sporny (2015)

Interpreting this markup is a straight forward affair, but an important one to discuss because of the underlying difference in the data model used by RDFa and Microdata. Microdata creates semantic data by simply structuring information through attaching specific points of relationships via reference to the Schema.org vocabulary. It does not use the RDF data model to create semantic data. RDFa, on the other hand, is a specification of the RDF data model and thus applies the full subject – object – predicate format to content. In doing so, RDFa provides a direct format for linking data and supplying additional context in the form of referenceable subjects, objects, and properties all with attendant relationship sets as defined by the contextual ontology. In Figure 7., semantic data is created by supplying not only the set of formal relationship as determined by Schema.org, but also by linking the content being marked up with information found in the Wikipedia entries for Delia Derbyshire and the University of Cambridge, each with a unique identifier. Recall that each component of the RDF data model needs to be uniquely referenced by IRI or literal values. In this case the "*name*", "*sameAs*", "*organizationRole*", etc. predicates and objects are referenced implicitly by their associated IRIs

within Schema.org – “http://schema.org/name” or “http://schema.org/GovernmentPermit“ (Figure 8.), but do not require the individual specification of the root namespace each time, as Microdata does. One can specify context at the onset, and all of the following code will draw from that specified root to determine context and set the boundaries of the meaning space. Figure 9. makes a deviation from the models supposed this far (Figure 7. and Figure 8.) by integrating a second ontology (Line 1). The interpretation is the same as the one from Figures 7. and 8., but for the added caveat that predicates and objects that begin with the prefix “ov” reference the Open Vocab ontology for their sets of relationships. This is a significant departure from the Microdata specification, as it allows the RDFa syntax, and the Schema.org ontology to enter into connections with other data assemblages, in this case the Open Vocab ontology. When unspecified by a prefix, the parser reverts back to Schema.org in this case since it is pre-defined in the standard itself. In practice, this can become very complex as managing compatibility and working across ontologies can lead to the issues discussed in earlier sections, where types, classes, and definitions are treated or understood differently.

While compatible, RDFa is not built into the HTML5 specification as is Microdata. It works in a far wider variety of languages and situated contexts, but it relies on a parser being able to decode the RDF data model, not just HTML5. Additionally, since it is not directly integrated into HTML, one of the most widely used Web standards, one can reasonably expect that it will be less familiar to Web developers, as they would need some additional training in building semantic Web architecture. Indeed, the data seem to indicate that developers are more familiar with Microdata than RDFa. However, RDFa does not suffer from the same problem with modeling reverse relationships and so does not constrain ontology development

and use in the same way. Similarly, RDFa can support multiple types for a single object and so some of the problems of semantic hegemony discussed above are mitigated with regards to markup standards.

5.3 JSON-LD

Javascript Object Notation (JSON) is a lightweight and human readable syntax for sending, receiving, and storing data on the Web. Since data transmission between servers and browsers is limited to text only, JSON provides a compact means for transmission between Javascript objects on the server and a browser which can then be embedded into HTML and other formats and languages (W3C 2017). JSON-LD is a W3C linked data specification of the JSON syntax. Being JSON based, JSON-LD can be parsed by any agent that reads the JSON syntax, not strictly those that read HTML such as Microdata, or those that can parse the RDF data model such as RDFa. Moreover, JSON-LD is more compact and easier to code than the other two syntaxes. Rather than revising the HTML markup in Web content, JSON-LD markup can be embedded into HTML or application environments as discrete objects making it easier for content developers to adopt semantic technologies (Sporny et al. 2014). JSON-LD is also developed openly, and so anyone with the inclination and ability to contribute to its development and modification, can. However, that openness can result in its own set of problems and complications. Furthermore, JSON-LD is a good mechanism for adding semantic markup to content when the information environment and vocabulary is changing, as it is relatively easy to update. However, there are limitations here that are discussed below.

Figure 10. JSON-LD Markup of a Government Permit

```
1.      <script type="application/ld+json">
2.      {
3.        "@context": "http://schema.org",
4.        "@type": "GovernmentPermit",
5.        "issuedBy": {
6.          "@type": "GovernmentOrganization",
7.          "name": "Department of Health and Mental Hygiene\""
8.        },
9.        "issuedThrough": {
10.         "@type": "GovernmentService",
11.         "name": "NYC Food Service Establishment Permit Service"
12.       },
13.       "name": "NYC Food Service Establishment Permit",
14.       "validFor": "",
15.       "validIn": {
16.         "@type": "AdministrativeArea",
17.         "name": "New York"
18.       }
19.     }
20.     </script>
```

Source: Schema.org (2017d)

Figure 10. shows the same original markup from the Microdata example (Figure 6.) and the RDFa example (Figure 8.), only in JSON-LD. Of immediate note is the parsimony in the actual markup. This snippet of code would be embedded in a standard HTML document or other code script, as offset by the specification of the script type and curled brackets (lines 1-2). As with the other markup standards, we see that context is initially supplied by Schema.org (Line 3). Furthermore, each ensuing part of the initial markup example is clearly and cleanly encoded by the appropriate Schema.org types, classes, and properties. In JSON-LD the ampersand indicates to a parser to deference the contextual IRI followed by the specific add-on in the ensuing quotations. These code forms following each ampersand are digital objects that enact a sort of protocological control that specifies and permits only certain expected values (Galloway 2004, 2006). The specified IRI is then used to establish the rest of the semantic string. For example, "*@type*": "GovernmentOrganization", indicates that the agent should parse "http://schema.org/GovernmentOrganization" to establish is subject, object, predicate relation set. The "GovernmentOrganizaiton" subject, has the expected property, or predicate, "name"

which expects a literal value, in this case “Department of Health and Mental Hygiene”. This simulates the general RDF data model used in the semantic web, without actually being the formal RDF model. Of note, however, is that one could add an additional property to the string by adding the expected property, “URL”, with an appropriate value. This would effectively create an RDF triple where each subject, object, and predicate was uniquely referenced by an IRI. As we will see below, however, simply adding additional properties that deference IRIs, can conflict with the context file and invalidate markup.

Figure 11. JSON-LD Markup Using Multiple Ontologies

```
1.  [
2.    {
3.      "@context": "http://example.org/contexts/person.jsonld",
4.      "name": "Manu Sporny",
5.      "homepage": "http://manu.sporny.org/",
6.      "depiction": "http://twitter.com/account/profile_image/manusporny"
7.    },
8.    {
9.      "@context": "http://example.org/contexts/place.jsonld",
10.     "name": "The Empire State Building",
11.     "description": "The Empire State Building is a 102-story landmark in New York City.",
12.     "geo": {
13.       "latitude": "40.75",
14.       "longitude": "73.98"
15.     }
16.   }
17. ]
```

Source: Sporny (2014)

Like RDFa, JSON-LD is mobile and can also be used with multiple vocabularies. Figure 11. shows an illustrative example where the context for “name”, “homepage”, and “depiction” are supplied by the fictional IRI specified in the initial *@context* declaration (Lines 3-6). It shows a second context used to supply “name”, “description”, “geo”, “latitude”, and “longitude” from the second specified *@context* declaration (lines 9-14). This would likely be done to remedy gaps in semantic coverage from the first contextual ontology. This ontology about “persons”, would be unlikely to cover locations, their descriptions, or their geographic coordinates.

Likewise, the second contextual vocabulary about place, would be unlikely to cover people and their descriptions. Where semantic ontologies are very limited in scope this can be useful strategy, made all the easier with JSON-LD, as it allows context to be defined local to the syntax and from multiple namespaces. However, in this example, there are overlaps between the "name" property. While in this fabricated example drawn from the W3C specification document, "name" would not likely be defined in drastically different ways across the two ontologies, it sets up a point of potential conflict in the markup strategy. For instance, "name" in the place vocabulary, may specifically indicate that it is looking for the name of a location as defined by a standards organization or it may expect an IRI or set of G.P.S. coordinates as a value, not a simple text string. Moreover, recall that the @syntax, while a mechanism for establishing IRI indication, is also a control protocol that establishes acceptable conditions. These protocols must be in alignment with the expectations of each ontology separately and each ontology combined for the multiple semantic data assemblages to successfully connect. These differences can, and sometimes do, invalidate markup and create representations that deviate from their modeled reality. In the JSON-LD specification, when item overlap occurs, the parser defaults to a "last in" override. In this case, a parser would not use the context file from "person" to define "name" (Lines 3-4), but would instead use the context file from "place" (Line 9) to define both "name" specifications (Lines 4 & 10). This means that a developer must account for the full coverage, the structural relationships, expected values, and underlying semantics of any ontology that they include in their markup. This complication increases in its affect as ontologies consolidate, become broader based, and grow larger. Of further note, this

context specification (Lines 3 & 9) can all occur on the same initial line, i.e. both person and place context files could appear on line 3.

Of course, JSON-LD, while the preferred markup standard by Schema.org, contains additional limitations that constrain the development and use of the vocabulary. Embedded JSON objects add extra file size for each page, and this makes parsing it more computationally expensive. This added expense can create performance problems for smaller and older processors, or internet connections with lower bandwidth. This added expense is another factor that can contribute to Schema.org's de facto status as the standard semantic web ontology. Not only is it easier to ensure semantic coherence when one is drawing on a single vocabulary, but it is also less computationally expensive to only have to download one context file each time a webpage is parsed.

5.4 Effects on Ontology Development

Schema.org's ontologists must continuously grapple with the affordances and constraints provided by the three markup standards. Moreover, they also must balance these affordances and constraints against their approach to ontology development and their specific data model. These conflicts and limitations most often rest in the background and do not visibly affect development and use. However, sometimes, these affordances and constraints work against their modus operendi, their applied data model, their ability to make the world machine-readable, and ultimately their ability to construct a world of semantic data.

Until somewhat recently, Schema.org over-determined all service provision, operation, production, and performance in explicitly commercial terms. Their way of representing the

provision of services, goods, performances, and the production of creative works, among other things, was to subsume these types under the seller property. As mentioned, this had the unintended effect of capturing non-commercial activities under a commercial rubric, with all of the attendant assumptions and properties that the seller property brings to those semantics. As a solution, one of the steering group members proposed that they adapt the provider property to cover non-commercial relations, notably, for this example was coverage for library holdings. Problematically, this change would alter the inheritance paths for some other existing properties and markup, as well as increasing the complexity of adopting the markup in further use cases. Ultimately, the group settled on another steering group member's suggestion that they recycle the existing property *makesOffer* as a property type between the types *person* and/or *organization*, redefining it as "a pointer to products or services offered by the organization or person."

There are three substantial issues that arise from this change. The first issue relates to the application of the Schema.org data model and increased complexity, while the second issue is a complication stemming from the first. Both of these two issues will be discussed in the next chapter. The issue most relevant here, however, is in the directionality of the *makesOffer* property. As a solution, *makesOffer* was an elegant way to resolve the initial problem, as the ontologists only needed to encourage its use by developers and to modify the recommendations and wording on the property's page to fit the new use context. However, *makesOffer* is unidirectional. A *person* or *organization* can make an *offer* of type *offer* of any large number of types, products, services, etc. each with their own modifying terms and properties. An agent might interpret this and display a list of associated services and functions

but have no ability to see the inverse relationship unless an inverse were allowed or specifically specified. This means that if an agent were to parse the thing offered, it would not be able to traverse the relationship graph to determine that it was an offer and who or what was making the offer. Potential use cases include severing the link between outpatient services and the hospital that offers those services, as well as searching for a book title but not being able to connect the search to the library in which the title was located, among many others, as this property has between 10,000 and 50,000 use domains.

As discussed earlier, RDFa and JSON-LD both natively contain the ability to represent inverse properties. This means that an agent can parse any content in any direction when that content is marked up with either one of the two standards, providing that the agent can parse those syntaxes. Microdata, on the other hand, does not include that ability. Microdata's inability to encode reverse properties comes with a good deal of consternation among the ontologists, as two contributors note, "I'm all for reverse property mechanisms but as far as I'm aware @itemprop-reverse isn't part of the html5 specs… [m]eaning microdata is left behind with this solution" and "well, RDFa defines @rev and JSON-LD defines @reverse so it's really only microdata hanging out there". While some contributors echoed familiar refrains about beginning the process of deprecating Microdata like, "and anyway, JSON-LD is the new king in town. RDFa will retire and go to the Guggenheim Foundation for its abstract beauty and Microdata to NASA's Houston Space Center for its practical impact", others make note of its seemingly unfortunate continued relevance to their work, despite its limitations. Microdata is the most frequently used of the markup standards, a fact that some contributors point to as a counter to deprecation noting, "[u]nfortunately, no matter how much we denigrate other

serialization approaches, a significant proportion of the adopters of Schema.org are wedded to Microdata" and "I should check with [name] or [name] before claiming Yandex is satisfied, given the relatively high amount of microdata in [R]unet".

Here, we can see the frustration that ontologists have over the constraints that markup standards place on their work, in this particular case, Microdata. To resolve this issue, ontologists had to break with their design principles and data model to create an additional and separate reverse property, *offeredBy*. This same problem pervades much vocabulary development where the ontologists are trying to model inverse relationships. Additionally, by simplifying the markup to the *Organization or Person -> makesOffer -> Offer -> itemOffered -> Product* pathway, they are creating added markup for Web developers that use Microdata. Not only will Web developers need to add a separate property to model the inverse relationship, they will also need to specify a new *itemtype*, "*href*", and their associated unique IRIs since each offer would now require the code to model the inverse as a separate feature of the markup. The impact of this constraint is potentially severe, as the data indicates that not only is Microdata the most common markup standard, but it also generates the most questions from Web developers. In each case, they are confronted by three options. They can create an additional property, breaking with their stated desire for simplicity and parsimony. They can refer Web developers to alternate markup standards, a complication discussed later. Alternatively, they can refer users to the experimental *@itemprop-reverse* syntax.

This last option deserves additional discussion. Recently, advice has been to use the experimental syntax, rather than encoding reverse properties in the ontology. This experimental syntax is not a part of the official specification and thus sees more limited use

across the Web. Presented with the continual problem of encoding reverse relationships in markup, some of Schema.org's ontologists worked through the limitation and enacted an unofficial change in the markup standard to suit the designs of the ontology. If and when this syntactical change occurs at the official level, the already developed reverse properties will eventually be deprecated, invalidating existing markup. While this change would be in the best interests of both users and ontologists alike, it is a complication for enacting machine readability. It also makes a questionable assumption that Web developers keep up with the latest Microdata specifications.

Another major issue mentioned to earlier, is that Microdata does not allow for multiple types for a single object where that object exists across ontologies or when property use is not the same for each type in the multiple type entity. This adds a major constraint to both ontologists and Web developers alike, ultimately leading to both ontological over-determination and ontic occlusion. For example, in Schema.org, many classes are listed as subclasses of product – any class that could be viewed as a product to be offered, sold, performed – even though they may also be represented in non-product contexts. This limitation is significant because the inheritances, expected types, properties, and values of the product class do not include the same ones relevant to other classes and will cause validation issues when the expected classes, which are not applicable or available in a non-product context, are not satisfied. That is, the real-world characteristics of non-products are not necessarily the same as those of products, and cannot legitimately be coerced to be so in the markup. In one broadly applicable and often cited issue relating to the problem of commercial over-

determination and multiple types, the Schema.org ontologists indicated the following needed changes to completely address the issue:

- We must establish it as a standard that products are multi-typed entities in examples etc.
- All major consumers of schema.org markup can process multi-typed entities.
- Move all subtypes of Product up to Thing or a branch thereof.
- Move up all properties of Product that are not tied to the Product role up to that subtype(s) of Thing (e.g. weight).
- Update descriptions and examples.

This solution requires that the ontologists move all of the subtypes and properties of *Product* to the level of *Thing*. This seemingly simple solution is actually a fundamental reorganization of the way that the ontology is ordered. The *Thing* class is the root level of the ontology, it is the most basic class with the most basic set of types and properties. All artifacts in the ontology inherit the types and properties of the *Thing* class. Moving these product subtypes to this level would not only invalidate a large amount of existing markup, but it would also either over-determine content as a product and in the Microdata contexts, occlude all other possible types, or require that all subtypes contain the same set of product related properties, even in non-product contexts. Unsurprisingly this potential solution to the multiple types problem was met with resistance and lament for Microdata's continuous effect on development, with some finding "it very unfortunate that we seem to keep struggling with limitations of Microdata." In this example, markup standards are directly inhibiting the ability to semantically encode information. Using the Microdata standard, one is left with the options of over-determining content as a product, thus occluding other more accurate data or not creating semantic

content, and where all search is now semantic search, this creates an entirely new set of problems for a Web developer.

There are times where seemingly small constraints imposed by markup standards have major reverberations throughout ontology development. In one issue, a steering group member noted a general guidance problem with respect to encoding languages in markup. The problem is that, in some cases, guidance directed users to use the formal language and in others informal abbreviations, a problem complicated by the diverse understanding of language that Schema.org needs to attend to (addressed in the next chapter). Another steering group member made the following proposal:

> To be frank, I think the handling of language information was much better in the strict RDF worlds where any plain literal could have a language tag, so it was straight forward to represent alternative texts for the same property depending on the language. Of all relevant syntaxes for schema.org, only Microdata lacks this feature AFAIK. Why don't we simply recommend RDFa or JSON-LD for use-cases that require language meta-data? That seems much better to me that introducing a mechanism at the level of the vocabulary.

Here, this member notes the limitations of the Microdata standard in adding meta tags to tell the parser what language a given set of content is in. Rather than hard coding complexity and, perhaps more importantly, constraints into the ontology, this member proposes that the ontologists shift the responsibility for language tagging to the users. For this particular suggestion, that responsibility shift is substantial. The data indicates that using semantic markup is challenging and fraught with confusion. While selection bias and other issues of representation in the data do not allow me to generalize, certain standards seem to be more confusing than others. The challenges of markup, and drawing on semantic ontologies in general, are often noted in the data for reasons of understanding the underlying application of

the RDF data model, the depth of class-type-property hierarchies, differences in expected values, and the affordances and constraints imposed by markup standards. So, a suggestion for a shift in responsibility to the user is an uncommon occurrence, one that actually goes against Schema.org's collective development philosophy of making the ontology as simple as possible for Web developers. Moreover, suggesting that in particular use cases, users work around the problem and adopt different markup standards in different use cases is problematic.

Technically speaking, there is nothing the prohibits a developer from using all of the syntaxes on the same page or even to describe the same thing. However, differences and the cross-compatibility issues described earlier, mean that there will likely be information loss or confusion on the part of whatever parser is looking at the markup. This is particularly the case if one were to use multiple syntaxes to mark a single piece of content. In fact, both Google and Microsoft's documentation for building semantic data into a webpage is very clear about this (Bing 2017; Google 2017). Moreover, the added time, knowledge requirements, and complexity means that the likelihood of error is increased and the ability of a given computational agent to parse the content is reduced. So, assuming that one needs to mark their content, in this example the constraints imposed by Microdata make it so that one either shifts their usage patterns, which is difficult and time-consuming or adopt a second standard, which can create compatibility issues and information loss. Alternatively, the ontologists can restructure the ontology around the combined affordances of the standards or operate with those known restrictions. I find that the ontologists most often either try to negotiate with those restrictions by either developing around the constraints and allowing generalized and bricolage type approaches to markup, or by allowing the issue to persist pending further work.

In this particular case, the constrains imposed by Microdata opened up further issues relating to the encoding languages. As one steering group member notes,

> [Y]ou're missing the leading use case for the Language type: the many other contexts in which it is useful to mention and name languages, beyond annotation of sections of textual markup. For examples see the incoming properties section at bottom of http://schema.org/Language:
> - http://schema.org/availableLanguage (of a ServiceChannel, ContactPoint)
> - http://schema.org/inLanguage (of a CreativeWork, Event, etc.)
> - http://schema.org/programmingLanguage (on SoftwareSourceCode)
> - http://schema.org/subtitleLanguage (Movie, ScreeningEvent, TVEpisode)
>
> In all these cases we might have an informal name for a language, or (for human languages) a code from https://tools.ietf.org/html/bcp47 or perhaps a Wikipedia link.

While using different markup standards to encode language metadata will solve the original issue, this member is noting that the suggestion to simply advise users to use alternate markup standards does not fully solve the problems of language. They point to the leading use case of language markup as being relevant to a wide subset of different usage domains, not simply marking the natural human language of a given set of content. Here, it is evident that language markup can refer to that language used on a webpage, that language a service or content is available in, the language used in a creative work or event, the computer language used in a piece of software, or the language used or available for subtitles, among other language uses that are in development.

While many of these differences can be resolved by specifying best practices and clarifying guidance, some cannot. One member summarizes the issue thusly,

> Regarding programmingLanguage, which has as its expected type Language, I believe this is an awkward conflation of "language" in the conventional sense of "conventional human languages" with computer programming "language", "a system of signs for encoding and decoding information". They're clearly quite different things, which is why one won't find C++ in BCP 47… IMO the expected type for programmingLanguage

> should *not* be Language, but Text. (Absent, say, an enumerated value or code for programmingLanguage - but the former is unwieldy and has extensibility issues, and for the latter no standard, AFAIK, exists. This is not the case for human languages, which are unlikely to be extended and *are* supported by standards like BCP 47).

This member is drawing attention to the fact that *programmingLanguage* expects a language type as allowed by the adopted standard, which is a problem where the ontology expects a link to the Internet Engineering Task Force (IETF) BCP 47 standard in order to be compliant. As the community member notes, C++ and other computer languages are not included in natural language standards like IETF BCP 47. Relatedly encoding programming language either overcodes differences in things like HTML which are not really considered programming languages, but rather a representational script, or omits them from the ontology altogether. At the moment those pseudo-computer languages are omitted from markup coverage due to the complexity of modeling the distinguishing between representational script and programming languages. Here such differences are occluded.

The problems ontologists and Web developers encounter are not limited to Microdata, as another limitation to resolving this issue is that JSON-LD does not permit language to be declared by an IRI, as would be common practice in semantic web markup. Instead, JSON-LD's *@language* specification expects text only. This means that the *@language* type could not be meaningfully used for something like computer language or programming language as declared in the Schema.org ontology, as those are links to both ontology contexts, and to specific use cases. This also means that language for JSON-LD has limited machine understanding since a parser cannot use JSON-LD markup to point to a relevant or equivalent case as defined by an IRI. In the Schema.org context, the *@language* feature of the JSON-LD specification prohibits

translating the variety of languages in a way that is machine-readable if one is trying to declare a language specification.

To negotiate these limitations, ontologists created a new language type called *computerLanguage*, updated the language term in the ontology to note that former uses may have included *programmingLanguage* (in effect requiring old markup to be corrected), updated the *programmingLanguage* to expect both the new *computerLanguage* link and text strings (to account for JSON-LD's limitations), and added the IETF BCP 47 standard to the expected natural language types with the added property value *alternateName* to include deviations from the IETF BCP 47 standard. This works around the limitation by creating a new type to avoid the association of parsed language in use, indicated by *@language*, and language used by marked content. Some of these changes are significant in that using markup to indicate text strings is not creating semantic data. It merely creates a structured presentation mechanism for a parser. The computational agent has no understanding or reference for that content.

While many semantic web standards allow for the bricolage of different vocabularies, they have trouble when terms overlap. As mentioned earlier, in JSON-LD you can supply multiple namespaces for vocabulary context. However, it uses a "last in, first out" method of handing cases of term overlap. This means that if a publisher were to mark content with the specific semantics of the first namespace and other content with the second if that second namespace contained the type or property used to mark the instance from the first namespace, it would overwrite the markup that was originally declared using the first namespace. This would happen regardless of the semantic differences between uses in the two vocabularies. While in some instances, this is a non-issue, in other cases this might result in more serious

problems. In either case, however, the standard puts the interoperability between ontologies at risk, and thus endangers a major feature of the semantic web. It also risks reducing the semantic coherence between a user's intent and computational agent's interpretation. The risk is furthered where deep understandings of JSON-LD are missing, and/or sophisticated knowledge of the various ontologies is lacking. While the data is likely to have some bias towards misunderstanding, given that those who understand are less likely to ask questions, queries about markup standards and ontology details are commonplace, and the ontologists frequently discuss the problems that the complexity of ontologies and markup standards bring to their proper use. Data also indicates that in many cases, even sophisticated core community members have trouble understanding proper class type usage and inheritance paths.

While the standards implicated here, particularly RDFa and JSON-LD, are linked data standards, their specific implementations and promotions through Schema.org obscure that ability. My investigation revealed almost no mention of linking to other ontologies, no examples of best practices for interlinking, or in the advice to that effect on the community board. The main exceptions were when linking to directly associated extensions. This includes the various internally developed and hosted extensions, the externally hosted, but compatible GS1 extension, which was developed with input from the Schema.org community. The other exception was when linking to the Wikidata database, which is externally hosted and developed but is in the process of being mapped for cross compatibility with Schema.org.

It is important to note at point that this is not attributing any form of intent on the part of the ontologists or their associated organizations, but that certain unintended consequences can emerge from the limitations these markup standards contain especially when those

standards are used across large scale ontologies that are likely to have a large amount of overlapping coverage. One consequence for users is that one can either stay strictly within the approved Schema.org universe or manage the potential complexity and incompatibility of multiple vocabularies. This also means that ontologists looking to extend or map their existing vocabularies to Schema.org need to do so nearly exactly. Furthermore, any of these compatibility mappings need to harmonize across external extensions, internal extensions, and the core vocabulary. This all makes creating and using ontologies very difficult and in many ways, counter to the original aims of the semantic web (Berners-Lee at al. 2001).

Recall that markup standards either directly deference an IRI for each individual subject, object, and predicate, or set an initial context(s) through which they establish the definition of terms, properties, and relationships. Particularly for JSON-LD, that means that all context needs to be supplied in each case, or available from a publicly hosted server requiring that any ontology have the ability to host a file in a way that can sustain potentially significant levels of Web traffic, as each time a piece of markup was parsed, the computational agent would request the hosting page. Indeed, this was an early concern among some of the Schema.org developers while JSON-LD was being created. As it stands now, Schema.org does host a context document. This context file also includes the entire range of other vocabularies that have compatibility with Schema.org. This effectively gives Web developers two choices: provide IRIs for each context file, dealing with the added complexity and room for error that such an approach entails, or use Schema.org's context file and associated vocabularies. This can have the unintended effect of implicitly forcing the semantic web community of users to abandon the more complicated, but potentially more expressive, and diverse set of ontologies in favor of

a smaller set that have become a de facto standard. This also has the effect of forcing ontologists to contribute to Schema.org's set of ontologies, develop their own in accordance with Schema.org's extension guidelines, or risk disuse and conflict with one of the most widely used and supported platforms. As the de facto semantic web standard, their choices become everyone's choices providing that a Web or application developer wants search engine exposure or the ability to parse the largest set of semantic data. As one community member put it "for me and my coworkers, Schema.org is the semantic web."

Ontologists and developers alike use validation tools to ensure that their markup conforms to the dictates of the relevant standards and will be legible to content consumers. Problematically, however, these validation tools can come to disagreements in how they parse content. This issue is most relevant to JSON-LD markup. The underlying problem is not specifically a problem of the standard, the parser, or the ontology, but emerges from the interaction of the three. The parser downloads and interprets the relevant context file at the first stage of the JSON-LD syntax. From there it uses the context file to interpret the terms, objects, and values given by the JSON-LD markup. These items are the ones in the syntax with ampersands preceding them. They indicate to the parser what to expect as an input based on the dictates of both the context file, which dereferences the ontology, and the syntax rules of JSON-LD. A steering group member explains one problem as such,

> Currently the context file… has this: "namedPosition": { "@type": "@id" } ... because an URL is a possible value. However, text is also a possible value, and currently more likely. The problem is that the JSON-LD context forces the property value to be interpreted as a (possible relative) URI reference, hence in http://json-ld.org/playground/ the value shows up relative to the site the data's on: We could over-ride this, e.g. using: "namedPosition": { "@value": "Quarterback" } Or we could change the context for this property (and others?), so that literal values are the default. But then we'd need to use

> (something like) this notation for controlled values: "namedPosition": { "@id": "http://sport-vocabs.example.org/Quarterback" }

This steering group member is explaining that in the context file for the given property (there are many more affected than indicated in this excerpt), the *@type* tag is expecting an IRI preceded by the *@id* tag. In JSON-LD, the *@id* in *@type* objects must always be an IRI or a term from the ontology that is implicitly expanded to an IRI via the function of the ampersand in JSON-LD syntax. However, to create maximal coverage, Schema.org allows many properties to expect both IRIs and text. Problematically, neither the context file nor the *@id* tag permits text strings for these properties. As this member mentions, the group could change the context file so that the default expectation is *@value* rather than *@id* so as to indicate that a parser should expect a text string as opposed to an IRI. However, to keep the value as semantic data, users would have to add additional separate notation to point to a specific IRI. Currently, this issue is still unresolved because the two solutions presented above create different additional complications. First, overriding the context file is both unintuitive and creates complicated added markup for developers. Second, these specific overrides fail validation on two of the three primary validation tools, notably the widely used Google Structured Data Testing Tool (SDTT). This means that not only will Web developers be unable to see if their markup actually works, but also they will be told by the SDTT that it is not compliant for Google's parsers. Third, by changing the context file, ontologists potentially invalidate markup that already uses IRIs instead of text strings. This data would then be lost until Web developers recoded their content. At present, the current guidance is to override the context file by being specific about both *@value* and *@id* where applicable.

This issue is significant because it shows how markup standards interact with other components of the data assemblage to enact data in specific ways. Ontologists develop ontologies in specific ways that categorize and classify the world and its relationships to create semantic data. This involves complex negotiations with how and why to represent things. These decisions are bound by the affordances and constraints of the markup standards adopted. These standards impose constraints on the ways the data assemblage can be enacted, but they do not do so absolutely. The ontologists and Web developers negotiate these affordances and constraints to both work around, through, and with the markup standards. Problematically, these negotiations occur within other contexts, levels of expertise, and means for validation that bear upon data creation.

5.7 Conclusion

This chapter detailed the general ways that markup standards shape the development and use of the semantic web and the specific ways that they condition Schema.org and its use. Enacting semantic translations is not just a matter of classification. Classification work tells us that representation necessarily leads to forms of over-determination and occlusion complexities are blunted and marginal cases are made to fit into the classification schema at hand. Schema.org and semantic ontologies are no exception here. The process of translation from human to machine is also one of translating through code standards. At times markup standards block these translations as they prohibit certain directionality, plurality, and representational forms among other things. The Web content to be encoded may be of a certain static type, but ultimately, the particular markup standard that one deploys is actually

what calls the semantic data into existence, and it does so in potentially different ways as it translates between the work of ontologists, Web developers, and computational agents. At times the standards force ontologists to cater to their constraints, while in other times ontologists and users both, create work arounds where they avoid direct engagement with the standard and those constraints. Sometimes this means that they deploy a different standard to account for one's limiting features. In others, they adjust their approach to modeling and encoding content to better match the dictates of the standard. However, these actors also work with and through these markup standards, engaging directly with the standard itself. They work with the standard by accepting its limitations and making subtler modifications to their practice to enable other parts of the markup standard to function in ways other than the standard's and ontology's original intent. Here there the negotiation with standards is one of strategic compromise where the affordances and constraints of practices and standards shift to enact the translation. However, at times practitioners work through the standards, changing the standard to fit the needs of practice, even if in unofficial, or not yet sanctioned ways.

The resulting situated markup for a given part of the environment will potentially differ in both form and function for different markup standards depending on the content to be encoded and the way affordances and constraints are handled. This has an effect such that the specific ways that ontologies and standards are adapted to fit their respective allowances and prohibitions can result in the differential enactment of semantic web assemblages and the semantic data the enable. This is especially the case as ontologies grow in scope and size and themselves take on the role of a standard (Waller 2016).

In addition to affirming past work attesting to standards' jussive power, this chapter begins to address the lacuna in the literature on standards by showing the ways that these standards sometimes work against the very stability and interoperability that they are supposed to enable. Furthermore, I situated the standards in the communities of practice that negotiate those restrictions, showing that contrary to the divergent claims of prior work, practitioners work around, with, and through standards to negotiate their failings (Halpin 2016; Whooley 2010, 2014; Whooley and Horwitz 2013). However, I conclude that most often practitioners worked with and through the markup standards. Here that took the form of adapting the ontology to better accommodate the strategic use of the standard, overriding context files, or deploying markup in ways that over-determines contextual content. These practitioners negotiate the limitations of different standards differently depending on the standard implicated, the possibility of affecting large amounts of markup, the complexity of the solution for the ontology, and the added complication of a fix for users. Lastly, I detailed some of the unintended consequences of those negotiations and how they can serve to over-determine and occlude certain understandings of the world.

6. Practices, Rationalities, and Communities

As noted earlier, the shapes that data assemblages take are dependent on the interactions between the assemblage's components. Thus, the study of semantic data must necessarily engage with the questions relating to processes of creation and the interactions between the various components. Here I am concerned with the specific apparatuses that include, systems of thought, practices, and communities that work together to enact the ontology (Kitchin 2014b; Lynch 2013; Woolgar and Lezuan 2013, 2015; Sismondo 2015). Within these apparatuses are the conceptual models, rationalities, techniques, conventions, and the communities involved in applying those systems of thinking and forms of practice to the process of human-machine translation (Kitchin 2014b:25). As such, this chapter examines the interactions between those various elements of the Schema.org data assemblage. In this section I concentrate on the two main communities of practice relevant to Schema.org and the design philosophies, habits of logic, and rationales they deploy. These two groups are the ontologists charged with the vocabulary's creation and maintenance, and the Web developers who deploy the vocabulary to convert their content into semantic data. This examination will reveal the ways that Schema.org ontologists adapt the RDF data model to their specific needs, their underlying philosophical approach to modeling semantic data, the problems they encounter in doing so, and the consequences of the ways in which they negotiate those problems. Additionally, it will show how interactions with the user base affect those issues and negotiations.

This chapter contributes to gaps that exist in our understanding of AI systems. I fill the gap left by past work that ignores translations between humans and machines in its focus on

the translation between different experts, and the interpretations of machine output by users (Agre 1995; Collins 1990; Forsythe 1993; Ribes and Bowker 2009; Suchman 1987). Rather than seeing knowledge as something easily extracted or converted to machine understanding, the Schema.org ontologists recognize the difficulty their task. While the knowledge engineers in past work see the problem as simply one of managing human description errors and extracting the "correct" knowledge, these ontologists see the problem as a matter of pragmatics. They understand that they are choosing the terms and relations, and understand that a trade-off between precision and practicality is usually inevitable. Thus their approach to development reflects their interpretations of the state of users, consumers, the state of the Web, and the current state of the ontology.

In line with past work in assemblage theory (Delanda 2006; Deleuze and Guattari 1987; Galloway 2012; McCarthy 2017), I find that as attempts to capture meaning are increased and their complexities more accurately mapped, those attempts only ever create more complexity. Thus, development and use of Schema.org occurs at high levels of granularity, making those additional complexities largely invisible. I will argue that their pragmatic, generalized, and a/contextual approach to development can lead to problems of path dependence and indeterminacy. While, these approaches are not empirically distinct, their analytic distinction helps to show the way each contributes to the problems listed. Furthermore, I argue that the specific ways the ontologists and users navigate these problems can have unintended consequences that stem from issues of complexity and imprecision. Lastly, I argue, that this tension between precision and practical deployment is largely unresolvable, as it either over-determines a domain area or offers little in the way of coverage.

6.1 The Community

The Schema.org project is overwhelmingly dominated by a limited number of individuals, companies, and other organizations. However, Schema.org is openly available to anyone able and willing to contribute or that has cause to use its markup in creating semantic data. It is also worth noting that despite this central dominating cluster, modifications to the ontology have originated and continue to originate from marginal collaborators, and that Schema.org's explicit operating principle is to make contributions, changes, and extensions as simple and as free as possible. While the evidence is in line with other work that suggests perceived prestige and meritocracy serve as the differentiators of soft organizational power, contributions by competent and able community members are welcomed (Coleman 2004; Halupka and Star 2011). Steering group members often encourage users and members with proposals to establish working groups to expand the ontology's reach and scope. While, the project is shepherded by the steering group, which ultimately decides courses of action, their communications and directions are largely transparent. Additionally, it is the general practice of steering group members to establish rough community consensus on additions, changes, and other courses of action[20]. Furthermore, while debate and disagreement are common, particularly on complicated issues, the general timbre of discussion is professional and collegial. Moreover, despite the large differences in understanding and mastery of both the semantic web and Schema.org, those community members who are more knowledgeable and experienced are polite and helpful to those who are less well informed.

[20] Rough consensus here means that they informally poll the people involved or working on a specific issue. Involvement can range from two to fifteen participants and agreement is denoted by a +1 or a "thumbs up" emoji.

6.2 Pragmatism

The Schema.org ontologists take a pragmatic approach to development and they encourage a pragmatic approach to use. Developers and users alike opt for what works over what could work best. For the ontologists, this means grounding their development in existing use cases and empirical evaluation of their environment – the Web. For users, this means mimicking examples found on Schema.org's main website, Stack Overflow[21], and examples of other similar from the community message boards, despite perhaps not being perfectly accurate semantics. There are many moments in the development process where a community member, or a potential user initiates a proposal to cover a domain, specify an existing one, or clarify an area with vague or ambiguous semantics, only to face a question from steering group members, or other central contributors, regarding potential or already existing use cases. This is not in and of itself surprising since, as a commercial product, it makes sense to have work flow decisions be guided by such practical considerations. However, the value of practicality and the benefits that pragmatism in decision making have are not necessarily evenly distributed among the potential actors involved or affected by those decisions, as smaller proposals backed by the sponsor companies are adopted with little disagreement.

A somewhat recent example is with the *claimsReview* addition. Recently, Google, as well as a small collection of fact checking sites, wanted to add a way for their checked content to automatically appear in Google's news aggregator next to relevant articles. To do so, it had

[21] Stack Overflow is a question and answer forum for programmers to ask for and give coding advice.

Schema.org ontologists create the *claimsReview* class with which a fact checking site, or other content provider, can use to encode their content to say that they are reviewing a given a claim, by a given article or outlet, all with the attendant semantic markup for Google news to interpret. Google has its own undisclosed mechanisms for determining the improper use of the *claimsReview* markup, but its existence feasibly opens up complications that come along with automatically presenting anything claiming to be a fact check and the publicly uncertain means to verify and police improper use. This addition to the ontology was included for release despite having fewer than ten use domains, and without any real discussion. However, it was introduced to the community in the same way as all issues are entered into the project. With Google's support for the proposal It is not surprising that the there was little push back. I see that as relatively unproblematic, but it does mean that things of alternate interest and long tail domains don't receive the attention that they could, or perhaps should. There is a limited number of people involved with the actual work in Schema.org, and a smaller number still with the sophistication to complete the growing workload. When those people are overburdened as they currently attest to being, they are not likely to have the time, energy, or inclination to do extra work to cover domains that may have only small numbers of actual uses.

In most cases the concerns of covering use cases weighs heavily on decision making. The data are rife with examples and instances where core community members question requests for coverage and/or new proposals. One member asks, "are you aware of publishers who are putting relevant structured information into HTML sites already that would be candidates for adoption? Or that are doing other kinds of Web-based data sharing?" and in a separate case, "are there tools looking for religious events specifically to justify this addition?" In each excerpt,

this core member is inquiring about the potential demand for two different vocabulary addition issues. These particular cases were formal proposals made on GitHub. However, the most common place where use is questioned is on the W3C message boards, where proposals are most often less well developed than they tend to be on GitHub. Members have similar questions in this forum, such as, “I know of a lot of useful things we could do with the information. The question is whether there is someone who is going to do something, because until that point I am reluctant to support adding the terms” and “But before we get too far, what is the actual use case for this data? Who is going to process it and present it?” In many cases these fledgling proposals are just simple requests for coverage and are informally vetted before reaching the GitHub stage. This vetting process is not required to introduce a proposal to GitHub, but is instead a less visible means for participating. To paraphrase a conversation I had with a steering group member, the GitHub repository is akin to the main stage and many community members have expressed anxiety about participating at that level of visibility.

Use case driven pragmatism does not just concern new creation. Schema.org is a form of living document, as term definitions and expected values change overtime as novel applications of the existing vocabulary are introduced by the community of users or uncovered by the ontologists. Many frequent changes are of the following nature,

> A common change is for a property to be marked as applicable to some previously unrelated type, or to expect to take values of a new type. When this happens the textual definitions are often adjusted slightly too. This is either by listing the new types explicitly, or through the use of a more general term like "item", "entity" or "thing".

Schema.org’s ontologists reflect on usage and adjust the ontology in a way the seeks to avoid adding constraints. The steering group, as representative of the project itself, does not want to “define or dictate… and notion of ‘mandatory’ property”. The ontologists expect that users will

have different amounts of available information to encode and that those amounts or types cannot be determined a priori. They also know that the needs and abilities of consuming services vary. This means that they attempt to satisfy the broadest possible set of users and consumers, as their driving goal is creating coverage. The practical implications are that when they see or hear of a use driven result that they were initially unaware of, they implement it if it is easy, compatible, and does not invalidate markup. Also, it means that they operate under the assumption that some markup is better than no markup. Regarding this last point, the ontologists reflexively adapt guidance on the vocabulary to be more inclusive in order to accommodate less detailed or less sophisticated usage. This was particularly evident in the ability of the *listItem* and *recipe* types to both include any things that have steps and any things that have both ingredients (broadly construed) and steps, respectively. While the property values of these types would coerce a "how to", or instructional, guide's semantics, particularly in a non-culinary context, simply omitting property markup would allow the steps to be displayed, only absent the rich descriptive detail enabled by properties that allow for formally correct semantics and cross instance mapping.

This use driven approach often even supersedes precision in content markup. In one example, discussed in greater detail below, the community debated the most appropriate way to code recurrence into the vocabulary. One of the steering group members, who is often reluctant to add precision for precision's sake, makes the following argument,

> I agree that we might need to improve the wording for schema:Event, but I hesitate to add a new type for this. I would not fight against EventSeries, but I do not see any actual benefit for neither publishers nor consumers of data. In the end, what matters is not whether you can disambiguate a concept by a philosophical analysis but whether the distinction brings any benefits for automated data processing.

The basic argument, which is often repeated in cases where the drives for complexity, precision, and practicality conflict, is that if an instance can be modeled by drawing on other parts of the existing vocabulary, in this case date specification, then a computational agent does not need to understand that the instance is of a certain different type. This is especially the case, where the agent is only required to represent the instance according to generalized context. Here, that means showing an event series as an event that occurs at different times. However, as I discuss below, that over-determines a Web developer or consumer's intent, and the underlying state of the instance.

The data indicate that the drive for semantic precision most often prevails when simplicity and practicality agree with it. This agreement in most commonly found where proposals are small, clear, less detailed, and introduce novel and/or superficial coverage to the vocabulary. A noted example was when the community added *Exhibition* to the vocabulary and then further marked the distinction between *exhibition* as *creativeWork* and *exhibition* as *event.* In this example there was consensus about ease of use, ease of implementation, and semantic accuracy. Thus, there was alignment between the forces that push increased precision, coverage, and ease of use (McCarthy 2017). Though this is not exclusively the case, as major semantic gaps prompt majority agreement for the need to modify or complicate the ontology, especially where modification is use driven. However, as with the recurring events case above, those issues that are subject to significant levels of debate are often longer term problems and remain complicated and unresolved as the community weighs the impact across other sections of the vocabulary.

Breakdown in semantic accuracy occurs when specifications are made to existing domains or when new domain proposals involve a high degree of semantic complexity. As one member remarks about a proposal for integrating legal decisions and terminology,

> The problem is that internationally there are SO many different vocabularies—I'm not just talking about language differences, I'm talking about the way legal concepts are referred to and thought of in different jurisdictions. Not to mention at all levels from international… to national to regional (state, province) to municipal and other local levels. So trying to include all those existing vocabularies is not only monumental but probably unworkable. In a situation like this it is actually more helpful to… come up with high level generic terms to which individual vocabularies can be mapped.

In this example, semantic complexity is simply too great to manage, let alone integrate in a way that developers could easily use. The trade-off in this case was a proposal to use generic legal terms, but even those terms ran into semantic difficulties because they were based on the U.S. legal system and were not sensitive to the differences in other local regional, and global legal systems. This follows the pragmatic approach described in one member's explanation as to why Schema.org does not code detailed data provenance information into its vocabulary, "the world is too rich, complex and interesting for a single schema to describe fully on its own. With schema.org we aim to find a balance, by providing a core schema that covers lots of situations, alongside extension mechanisms for extra detail."

Thus, my data suggests the opposite of the work by Agre (1995), Collins (1990), and Forsythe (1993). Rather than seeing knowledge as something easily extracted or converted to machine understanding, the Schema.org ontologists recognize the difficulty of such a task. While social scientists see the problem as intractable, and knowledge engineers see the problem as simply one of managing human description errors, these ontologists see the problem as a matter of pragmatism. They understand that they are choosing the terms and

relations, and understand that there is a measure of “semantic violence” that occurs (Agre 1995). However, they are tasked with implementing the semantic web, and so they prefer to approach modeling in a way that is based on empirical use and compatibility, unlike Forsythe’s (1993) knowledge engineers. If a declared term or relationship causes little disturbance and resonates with the existing set of terms and relationships, they tend to accept that it is a good declaration. They do not, however, see themselves as building worlds, but rather representing them. They resolutely believe that they can approach an adequate representation of the world though discussion and consensus. As one steering group member put it,

> The problem that any knowledge representation system has is fully describing its system so that all users across all cultures and languages will understand it and have the capacity to expand upon that system while maintaining agreement across all cultures and languages. 'Negotiating the limits' often times just means that folks have to simply "talk it out", in email, GitHub issues, or meetups, until a consensus is formed.

6.3 Generalizability

Related to their pragmatic approach the Schema.org ontologists prefer to adopt a generalized application of the RDF model. This is in some ways inseparable from the use case driven approach and is arguably an improvement over the patchwork open approach of the pre Schema.org semantic web, at least in terms of operating at Web scale. Simply put, if the ontology is too complex, either no one will use it or they will use it incorrectly. A survey from one of the pioneering members of Schema.org, and the semantic web more generally, found that nearly 60% of all semantic markup on the Web was incorrect (Guha 2014). This leads Schema.org’s ontologists to advocate for restraint when it comes to adding to the existing vocabulary. Among many other places, this approach is evident in the long standing debates surrounding recurrence. In these debates, the fundamental issue for data modeling is whether

to materialize facts individually or encode decision rules into the vocabulary. This dilemma reflects two opposing positions within the Schema.org community. The debate revolves around forcing users to mark their content for each possible occurrence in each possible place, or allow for a recurrence rule that defines a repeating pattern of events, times, or definitions.

This debate exposed two different but related debates about recurrence. The first has to do with repeating dates, while the second has to do with recurring events. Adding markup for repeating dates would be broadly applicable across Schema.org. One steering group member notes,

> Recurring dates will be relevant for many branches of schema.org, namely for opening hours, but also offer validity and prices specifications, so I prefer a generic solution. Once we have this, there is IMO no need for a specific mechanism for recurring events, because we can simply use the recurring dates mechanism to specify the dates of the event

The member is arguing for a generalized approach to recurrence so that the underlying grammar of the ontology can model across contexts. At issue is that Schema.org has no ability to model change. While mutable, the world that Schema.org creates is not dynamic. That is, pursuant to the materialization approach, a user would need to encode all parts of a process in which change or differing states occurred individually as discrete artifacts. Working from the preceding excerpt, hours would need to be specified for each individual day with each exception marked at every instance. For offer and price validity, all fluctuations and price differences would likewise need to be changed manually at each instance depending on the markup syntax in use. Since prices are often generated algorithmically on many of the major e-commerce sites that use Schema.org markup, this materialization approach can create a mismatch between how price is modeled and how pricing occurs in actuality.

In a rule based approach, recurrence is modeled and the exceptions to the recurrence rules are materialized ad hoc. Adopting a decision rule approach is more complicated, but allows for the easier coverage of change so long as that change is predictable. Such an approach would require that the rule system be simple to adopt, specific enough to be of practical benefit to users, and also be general enough for use across contexts. However, as one member notes,

> The problem with recurrence rules is that no-one ever thinks to put an end-date on them (even if only a tentative one that can be extended later), so you end up with meetings/events with rules like "10-11am every Tuesday" and then look at your calendar a few years into the future and see it full of events that will almost certainly never happen.

Here, a member highlights one pragmatic problem associated with the specific proposal to integrate a recurrence rule into the ontology. The dilemma is that ontologists can either model change, thereby increasing the applicability of the ontology and reducing the potential error of users having to add markup for each and every possible point of change in an ad hoc basis, or run the risk of implementing a complex rule that produces user error and/or negligence.

There is a similar tension evident in the proposal to add compositional data modeling to the ontology. Compositional data modeling is an approach that algorithmically combines pre-existing parts of the ontology to model more complex concepts. One steering group member puts it thusly,

> We often have the same underlying concepts appear in several places in the hierarchy, as part of more complex concepts. So for example we might have GolfCourse (which combines the idea of a place where you do a sport i.e. a SportsActivityLocation with some specific sport, Golf). But we might also have a (something like...) GolfSportsMatch, which combines that same sport - Golf - with a different concept - the idea of being a kind of SportsEvent devoted to a particular sport.

> As schema.org grows, we get more and more of these situations. It becomes a burden to keep track of the different ways in which we've built terms that combine simpler underlying ideas (like specific sports). We don't want to clutter schema.org with every possible sensible combination of these ideas e.g. a GolfingTutorial or a CricketSportsEvent or a CyclingCompetition or a CricketTutorial or a GolfingSportsEvent or a GolfingCompetition or a CyclingTutorial. The combinations soon become unmanageable.

In keeping with their approach to not add highly granular subtypes into the ontology, a steering group member made a proposal to replace the input of a name-value pair linked to a specific Schema.org IRI with a function set that mimics an IRI with a series of arguments that reference type, property, and class variables in the ontology. In JSON-LD this turns an *@id* string from something like, "@id: http://schema.org/about" to the much more complex string, "@id: http://schema.org/<Function>arg1name=arg1value&arg2name=arg2value". Ostensibly, this will make the ontology much more expressive with little added complication to the ontology itself. It adds an intra-ontological reference system whereby the vocabulary can be broadened and types can be added as acceptable values where they may not be predefined or otherwise allowed. Problematically, this proposal adds substantial complexity to an already complex set of technologies, unsurprisingly leading to a number of the core community members to question the applicability of the compositional model. They state,

> Yes, the approach is structured but the implementation pattern would be so unlike the rest of Schema.org the result would be as unnatural as a shotgun marriage between SOAP and REST. Both pushing the same direction but a union incomprehensible to the outside world
>
> trying to understand it makes my brain hurt because of the - for me - too abstract description, thus making me skeptical about the practicality of the proposal for everyday webmasters… About that, first off, I've literally lost count of the amount of times I've seen theme developers utterly destroy graphs because they don't really understand how Drupal (and/or markup) works (and/or aren't aware that structured data is part of its core). Often leading to developers either letting broken markup get published or disabling those parts of Drupal that add structured data to a page. So unless [name] gets

> involved in the gazillion custom templates out there I wouldn't get your hopes up about seeing a lot of proper implementations.

These critical voices are not only experts in semantic web technologies, but are also experts in, and founding community members of, Schema.org. Furthermore, this proposal brings up important questions for the markup standards as there is some uncertainty about if and how they can encode such functional strings. Ultimately, the proposal is on hold and not attached to a major version update target as concerns about the proposal's usability and empirical value remain.

That tension between usability, generalizability, and specificity is a frequent issue for ontologists, who often argue that, "we need to find a sweet spot that covers 80 - 90 % of the cases. A formally proven perfect solution for the underlying problems is IMO beyond what Schema.org can achieve." Thus, their approach to data modeling is one where simplicity and generalizability prevail over complexity and specificity. Ontologists often shy away from adding types and properties to existing parts of the ontology, preferring instead to allow markup to be used in the broadest possible ways working with markup standards to strategically leverage the flexibilities that they offer. This often occurs despite potential problems with semantic accuracy and the unintended consequences of ontic occlusion and over-determination that can result.

The potential for over-determination is evident in this generalized approach. Surrounding these attempts to model change, is a large and diverse series of issues attempting to model things in series. In one example the community was trying to add conceptual clarification to events that have separate and distinct, or specific, events within them.

By way of example, they were discussing both the Salzburg Festival and the Olympics for use cases where there is a super event with many sub events that occur within it. The semantic problem that occurs, particularly if a simple recurring dates model is applied to the event designation, is captured in the following two descriptions of the problem by two steering group members,

> An EventSeries is a collection of events that share some unifying characteristic. For example, "The Olympic Games" is a series, which is repeated regularly. The "2012 London Olympics" can be presented both as an Event in the series "Olympic Games", and as an EventSeries that included a number of sporting competitions as Events (member 1).
>
> "The Ashes" (cricket contest between Australia and England) is repeated frequently, but not according to any long-term stable schedule. The course "Beowulf: the manuscript and the man" can be an EventSeries consisting of individual seminars with the same participants, and an EventSeries of the course being held repeatedly over a decade (member 1).
>
> For example consider if there is a 2016 seminar series (e.g. each of 6 x 1h sessions on successive weeks), and then later the course is re-run as a 2017 series with the same structure and topic, but different people. This also relates to whether we want to say that EventSeries is a subtype of Event (or at least carries Event-like properties). [member 1] outlines two views. One is to see the 6 sessions in 2016 as an EventSeries i.e. '2016 Beowulf course', the other is to see the sequence of annual runs of the course - i.e. 'Beowulf: the manuscript and the man' - as an EventSeries. Either way, superEvent would point up the hierarchy, from e.g. '2017 session 3' (an Event) to '2017 Beowulf course' (an Event or an EventSeries) and from there to the top level 'Beowulf: the manuscript and the man' which is presumably an EventSeries but not an Event since it represents just the high level notion of the series rather than something that you can say is happening at a particular place and time.
>
> The mid-level is the most awkward. Intuition pulls it towards being both Event ("An event happening at a certain time and location" e.g. summer 2016, Madrid) and also an EventSeries (approximately "a collection of events that share some unifying characteristic e.g. relationship") (member 2).

Here, the problem is as follows. A given instance is ontologically indeterminate, meaning that it can legitimately and empirically occupy two or more ontological states simultaneously. As

member 1 notes, the 2012 London Olympics can be both an event in a series and an event series. Likewise, the Beowulf course can be two different types of event series, depending on perspective, or an event in a series of events relating to an organization's course offerings. So, depending on the particular context under consideration, *event* and *eventSeries* could both be considered super types and subtypes. This means that marking the order of relationships and inheritances will necessarily over-determine the semantics of both *event* and *eventSeries* while occluding the alternate meaning.

By way of example, omitting such a distinction would mean the difference between marking the something like the World Trade Organization meetings (WTO) as an event with the relevant description and associated semantic properties that surround it, and modeling the WTO as an event series, with the 1999 Seattle WTO as a specific event in that series. One completely occludes the possibility of representing specific and important information relating to an event, the other enables a specific case to be represented and allows for the possibility of supplementary information to be included and linked to both the overarching super-event and the subject of that event, in this example the famed protests of the 1999 WTO meeting. Marking such distinctions is beneficial and relevant to consumers of information whose interpretations and intentions cannot be specified a priori. This issue also bears upon issues of reversibility and the constraints of standards discussed in the previous chapter, as well as, potential cross compatibility problems with other ontologies. This problem of precision, complexity, and usability is not fully resolvable.

This general approach to organization can also lead to problems of path dependency within the vocabulary. Consider the following example,

> In the previous medical/health additions we made the mistake of including terms whose name was implicitly contextualized to medical/health scenarios. So we "used up" the word "action" on a property that we now call… muscleAction and so on… every change needs to make sense [a]cross domain, since schema.org is a cross domain vocabulary.

In this situation, prior to the development of the hosted medical extension, ontologists created a general set of terms to cover a nascent medial domain, the problem was that this general term, along with its very large set of inheritances, was then used in a large number of use cases. Here a member points out that a previous vocabulary specification came into conflict with the development of an ontology extension because when the ontologists added specificity to a type of action they precluded its use for the extension to which it was eventually most relevant. This forced them to encode additional complexities into the extension that have caused misuse and conceptual confusion in modeling. Thus, we see their general affinity for broad or generic types and properties. This affinity however created problems as the ontology grows. Another member notes, "As Schema.org grows, generic names for types and properties leads to unfortunate collisions." Here, one of the steering group members brings up a long list of general terms and properties that at the time constrained the possible ranges of use and interpretation. For example, *season* was artificially limited to references to seasons of TV shows, precluding mention of any other interpretations. Likewise, *code* referenced software code omitting other potentially serious code types, like medical or legal codes which are both immediately relevant to Schema.org's internally hosted extensions.

Adding new subtypes and filling in additional context specific details would seem like the easy solution, but doing so forces the inheritance of an entire range of additional subtypes and properties, many of which might be inappropriate to the newly specified types. Alternatively, depending on how the developers code the new specification, it could mean the

displacement of properties and subtypes already deployed in a given publisher's markup. There are no one size fits all approaches to deal with the problem of path dependence. The ontologists, especially the steering group members, carefully weigh the consequences of updates. In some rare cases they deprecate the original term or relocate property inheritances under the new subtypes, but both solution occur at the expense of existing use cases. In most other cases the developers broaden the expected set of inputs and make use of flexibilities within the markup standards, though at the cost of further dulling semantic precision and reducing domain specific context.

6.4 a/Contextuality

Context in Schema.org, and in the semantic web more generally, is complicated. As noted in an earlier chapter, semantic web ontologies are constructed acontexually (Waller 2016). Purportedly, this allows them to be applied across contexts, but as Ginsberg's (2008) research shows, this can lead to serious complications including ontological over-determination and occlusion, as the context where markup is applied defines the socio-cultural range of interpretations (Agre 1995; Collins 1990; Knobel 2010). However, past work on meaning and context shows that the actual deployment of, and meanings surrounding, technologies never occurs outside of a specified context. Much of this past work shows that the definitions of concepts and their attachments to properties always necessarily rely on social context for their provision. That is, all claims are inextricably embedded in specific contexts (Bakhtin 1981; Heidegger 1973; Lakoff 1987; Waller 2016). For the semantic web, context is stripped from factors of its creation and then imputed to the individual instance. This resonates with Seaver's

(2015) claim that context is everywhere and at all times a matter of making decisions and statements about the state of the world. However, this re-imputed context is determined entirely by the boundaries set in the ontology and it unavailable to consumers in the way that past work stipulates (Agre 1995, 2003; Collins 1990; Ford and Graham 2016; Waller 2016).

While my investigation does not reveal that ontologists assume a single correct definition or context for every concept, as does other work (Forsythe 1993; Ginsberg 2008; Ribes and Bowker 2009; Waller 2016), ontologists have no method to address intentionality regarding use context and so are constrained in their ability to model context. I attribute this as one of significant reasons for adopting a pragmatic and generalized approach to modeling. Since by allowing term reuse, you obviate the need to account for context. Further, this superficially appears to eliminate the problem of imputing contexts to instances. However, I argue that while imputing context seems to be avoided at the local level of the instance, context is entirely determined at the level of the ontology. So far as semantic data is concerned, the ontology entirely determines an instance's ontological state, as it sets the boundaries of what can and cannot be.

The approach to modeling data inherent to ontologies necessarily over-determines relationships and their inheritance paths as they are defined within a specific domain that cannot be context agnostic. The alleged agnosticism is evident in the approach detailed earlier. Recall the previous discussion on how Schema.org over-determines service provision or agency by organizations in commercial contexts. The decisions to model types and properties relating to services as a commercial offer of a product was a pragmatic decision to cover an assumed significant majority of their use cases, but it has the problem of imputing commercial contexts

to things like volunteering, aid, pro-bono work, and other similar actions by organizations and individuals. Waller (2016:8) notes similar patterns with the actions afforded to the vocabulary type *Person*, where whether "'alive, dead, undead or fictional', are focused on involvement in movies, videogames, TV series, radio and sport and buying and selling.". This problem of context is not necessarily a fault of ontologists, but the fault of ontologies in particular, and the modeling of the world in simple terms of objects and their relations, in general. In this system of thinking, context is not locally determined by the characteristics and peculiarities of situations, but rather determined globally by the declaration of a canonical reference point, in this case, a context file or ontology.

In terms of context, Schema.org's enactment of data is captured well by one steering group member,

> We very often have the following workflow:
> someone suggests a new property + textual definition
> someone else says, "hey what about re-using property such-and-so?"
> they agree to re-use an existing property, perhaps tweaking its definition to mention or permit this new usecase.
> result: the original more contextualized definition gets lost.

and a Web developer who notes that, "Schema.org does seem to encourage reuse of terms that have the same semantics in different contexts." Ontologists call this boundary traversal "scope creep". However, these understandings of context display an understanding of context that fits within the paradigm noted in earlier work on AI, namely a context that ignores some level of constructivism implied in ontology development. While these community members are correct in that Schema.org encourages the pragmatic and generalized application of terms across contexts, as discussed above, this implies that context is somehow erased, rather than displaced and enacted in every instance. Much as Seaver (2015) argues that rather than the

analysis of big data stripping context from data points, it instead recreates it, the cross contextual application of Schema.org enacts contexts in the process of selecting the markup syntax, type and property usage, and underlying content to be created as data. Here, context is precisely the thing in question.

Moreover, this application obscures the way that context informs a vocabulary's creation. In every instance, vocabulary proposals are made with a specific single or set of use cases in mind. The proposer may be a Web developer, community member, steering group member, industry group, or sponsor company. While each different group comes to the community with different levels of sophistication and knowledge of the domain, the ensuring process is relatively similar. First, a proposal is made. As discussed previously, this includes the range of potential use domains, clarification of why it is needed, and what additions or modifications might be required, among other details that fit the particular proposal. Second, the community evaluates the need for the proposal and the potential for use in related contexts. This measurement of scope creep is the first stage in the abstraction of context from the initial proposal. Third, the best approach to model that use context is debated and measured for its fit into the existing ontology and its ability to be modeled appropriately with the three markup syntaxes. This is a next step in deconstructing context, as the community evaluates the amount and extent to which the proposed extension or addition can be satisfied with existing terminology. This has the effect of removing context specificity from a domain. Interestingly, this new domain proposal only happens because a user determines that the existing ontology does not fit the domain specific context required. The last stages are where the ontologists, often with the domain expertise of the users, set implementation details and

timelines to integrate the proposal into the core vocabulary or as an extension of it. The cumulative process of moving from the proposal stage to the implementation stage is one where small and compatible additions and modifications are stripped of their particular context to fit into the core vocabulary and hosted extensions in way that minimizes the impact to the existing ontology. As noted, this means reusing or adapting existing terminology wherever possible, abstracting meaning from its origin and set up as a standing reserve from which new contexts and new meanings can be imputed. The new terms are then put to use in markup recreating and imputing a new set of contexts, that may or may not be aligned with the ones contributing to their genesis.

To sum up, context informs creation, context is then removed as the new creation is integrated with the older set of decontextualized relationships, and then context is newly introduced at both local and global levels as the ontology is deployed by users. While this may result in little actual change to the initial context, that matter depends on the potential impact to the existing ontology, as well as the level of overall and the perceived value of the addition as negotiated by the proposer and the steering group members. In some instances, context bears some resemblance to the one initially supplied, but its use occurs in a diluted form, as with proposed mapping of the pre-existing European Legislation Identifier Ontology (ELI) ontology. In this case, integrating this vocabulary with Schema.org would alter the semantic context of the original vocabulary, as Schema.org's *validFrom* does not equal the ELI's *dateInForce* and *applicableDate,* and Schema.org's *basedOn* is not the same as the ELI's *madePossibleBy* in legal contexts. These examples among others in this proposal highlight the inadequacy of existing acontextualized Schema.org terms to encode nuanced understanding of

how the legislative process works. This is something further complicated by property directionality and its availability to be specifically modeled.

This case is additionally complicated by the fact that there is no single legislative process to be modeled, but rather many different processes. Returning to the excerpt from page 137,

> The problem is that internationally there are SO many different vocabularies—I'm not just talking about language differences, I'm talking about the way legal concepts are referred to and thought of in different jurisdictions. Not to mention at all levels from international… to national to regional (state, province) to municipal and other local levels. So trying to include all those existing vocabularies is not only monumental but probably unworkable. In a situation like this it is actually more helpful to… come up with high level generic terms.

We see that this user is responding to a proposal to add markup coverage for legal decisions discussed earlier in the chapter. He correctly highlights the complete inadequacy of trying to model variation in legal terminology across a wide feature set, as there is significant variation in legal structure, terminology, publication, and process across all levels of jurisdiction. This particular part of the problem is well discussed in the data, and can be seen as a unifying point connecting frequent proposals to add legal and legislative proposals that cover a wide variety of case types and jurisdictions. Currently, efforts to integrate the two largest and most detailed legal proposals (E.U. and U.S. based) into a hosted legislation extension are ongoing. Users and community members in this and related discussions, debated the differences in the European and U.S. legal systems establishing the intractability of the problem they faced. While establishing “high level generic terms” may be the best of a number of poor solutions, it is actually only the best where those high level terms draw out the correct legal contexts. As one user makes note,

> My only concern is that some legal issues aren’t about the government-to-citizen relationship (like civics would imply), but about family matters (divorce, child custody,

child support, domestic violence — these are some of the most common-searched legal help topics), corporate matters, contracts, property, etc.

Presently such concerns are not included in the pending extension. While these concerns have been noted by users and emphasized by steering group members, making additions likely, they will require careful consideration and preservation of existing real world context to have meaning in this domain. Additionally, as this proposal remains underdevelopment, new interest groups are advocating for coverage of other legal domains, structures, and jurisdictions, further complicating the addition. In meaning spaces where precision is paramount, system-level over-determination of legal concepts, options, and advice may have severe consequences for information seekers.

6.5 Conclusion

Schema.org ontologists adopt a pragmatic, generalized, and a/contextual approach to development. This empirically driven approach contrasts with the approaches detailed in earlier work on knowledge engineering (Agre 1995; Forsythe 1993a, 1993b; Star 1995). Furthermore, the ontologists and users recognize the semantic problems of their respective tasks and do not believe that they are simply representing a stable worldview, but rather, take the pragmatic view that if consensus based decision making can come to a developed representation that aligns with the existing set of them, they can construct a more or less true vision of the world. That is, they are acutely aware of the fact that they might be embedding certain cultural understandings in their work. Again, this runs counter to existing work (Forsythe 1993a, 1993b). Their pragmatic approach allows them to navigate between an entirely constructivist or realist path to development. By basing their development decisions on actual use on the Web, the

ontologists and Web developers can enact an ontology that while constructed, resonates with the existing and agreed upon interpretations of their information environment. The assemblage based approach allows us to uncover the ways that those states are exposed by attending to the situated interactions of components. In this chapter, those components are the rationales, strategies, modeling techniques, and communities of practice.

Problematically for this pragmatic approach is the tension between precision and applicability. Following from their use-driven approach, ontologists and Web developers alike, opt for applicability over precision, as that tension is not often able to be resolved. Relatedly, ontologists and users try to make the ontology as generalizable as possible. This has the effect of making it easier to use and more adaptable across domains. However, in addition to suffering the similar consequences of pragmatism, generalizability tends to lead to both over determining meaning, as it applies classes and types broadly across content domains, and to creating a path dependency problem. Path dependency creates a problem where the ontology faces two undesirable outcomes, invalidating existing markup or adding new coverage into the ontology with the potential to create semantic problems relating to class and property inheritance.

Finally, the ontology, far from being acontextual, determines contexts in two different ways. First, it serves to impute a specific context to a particular application at the local level. This removes markup from its original contextualized development and embeds it into a new domain. As the ontology is developed to generalize, this acontextual mapping creates a degree of indeterminacy, as both domains can make equal claims to legitimacy with the markup. This has the effect of losing the ability to specify semantics in any precise manner at that micro, or

local, level. Second, the ontology determines context in ontological terms by constructing a delimited state of semantic meaning at the that larger global level. Here, existence is determined entirely through the specific construction of the ontology, in this case, Schema.org.

7. Conclusion

In this concluding chapter it will be helpful to revisit my motiving questions and to briefly return to the chapters where they were investigated both in order to explain how I came to answer them, and to show how those answers address the literature that informed them. Motivated by calls in the emerging field of data studies, I initially sought to understand how data assemblages came to be. These calls stem from a lack of large scale studies about how data assemblages come together to produce data. This question can only ever be answered in part, as the semantic web is composed of many different assemblages, which themselves include various components. It would be impossible to completely cover the scope and scale that this question implies. Nonetheless, my investigation does explain how one major semantic web assemblage is enacted. Answering this question required engaging with literatures surround the various components of a given data assemblage. Major components addressed here included standards, developers, users, design philosophies, and habits of logic (Kitchin 2014b).

Answering this question required that I first answer two other major questions. First among them examined the way that ontological states are declared in the processes of developing and using semantic web ontologies. At the onset of this study I wrote that this question was primarily theoretical in that it attended to the particular delimiting of meaning space. While I maintain that the question is primarily a theoretical one that deals with the processes of ontological over-determination and occlusion that occur any time data assemblages take shape, it is also an empirical question. Empirically, the question shifts to my second subsidiary question which demanded that I understand how humans represent an

indeterminate world to a deterministic machine. It was important in this particular case to move past the long standing and intractable problem of accuracy in representation, to acknowledge that representations are happening and to understand how. Approaching these questions of ontology in terms of assemblage theory meant that I needed to understand how components are implicated in the assemblage and how they interact to express it. This required drawing on recent work in STS on empirical ontology and practice to remedy gaps in literatures surrounding standards, expert systems and AI, as well as emerging work on the semantic web.

Chapter five detailed the ways that markup standards shape the semantic web in general and Schema.org in particular. These markup standards translate the work of ontologists and the work of Web developers in ways that alter the both sets of practices. Not only do the standards force ontologists to cater to the dictates of the standards, but also the resulting situated markup may differ in both form and function for different markup standards as they offer different affordances and constraints on how semantics can be expressed and what forms of markup are allowed. This section showed how Microdata forced ontologists to modify the ontology to enable reverse properties and to allow for more permissive, though semantically less accurate, markup rules to account for Microdata's lack of reverse properties and multiple types across contexts. These modifications result in over-determining semantic states and ambiguous semantics. However, it also showed the ways that ontologists creatively work with standards to enact compromise and/or modifications to the standards.

This chapter also showed the ways the markup standards create and exacerbate problems with drawing on multiple ontologies. Using markup standards to encoded semantic content is a complicated activity. This is made all the more complicated as Web developers

need to have a deep understanding of the definitional and structural relationships between ontologies if they are to interlink them. Absent those deeper understandings, the problems of term override, namely ontological over-determination and occlusion, lead to problems where the human intent does not translate to machine interpretation. Furthermore, where Web developers lack an ability or motivation to draw on multiple ontologies, the space for semantics is constrained. Ultimately, the ways that ontologies are adapted to fit the constraints and affordances of markup standards can result in the differential enactment of the semantic web as the ontologies are developed both through seeking to accurately model an environment and through negotiations with the limits set forth by markup standards. The potential for differential enactments of semantic data is especially significant as ontologies grow in scope and size where they themselves take on the role of a standard.

My investigation into the interactions between communities of practice and markup standards leads to the following conclusions. In line with past work on standards, the various actors involved with Schema.org have a tenuous relationship with the markup standards used to enact semantic data (Bowker and Star 1999; Epstein 2007, 2010; Heimer 2001; Whooley 2010, 2014; Whooley and Horwitz 2013). In part, I agree with this research, finding that while standards do enable interoperability in many ways, they require users to work around their limitations. Here this takes the form of deploying a different standard to circumvent an imposed constraint. While workarounds exist, they were not the most common negotiation in my data, particularly as development progresses. Ontologists do not like to advocate for using multiple standards, and they prefer not to modify the ontology unless absolutely necessary. Instead, I side with Halpin (2016) finding that rather than working around the limitations

imposed by standards, actors usually act creatively to work with and through them. This happens in a few main ways. I show that standards force ontologists to compromise by modifying types and properties to fit into syntax rules that might otherwise not have been appropriate to use in such circumstances, mitigating the standards' limiting influence. I also show that ontologists have some measure of control over the standards themselves, modifying them to fit the ontology. Additionally, I show that users deploy the ontology in ways that while semantically over-determine their content, provides some level of semantics. These practitioners negotiate the limitations of different standards differently depending on the standard implicated, the possibility of affecting large amounts of markup, the complexity of the solution for the ontology, and the added complication of a fix for users. This affirms much of the extant literature, attesting to standards' jussive power. Additionally, it contributes to the recent debate on how practitioners negotiate the limitations imposed by standards.

Furthermore, there is evidence to suggest that the various markup standards work to create multiple ontological states, as the unique affordances and constraints of each contribute to the differential enactment of semantics. In the limited example in this study, we can see such an effect with the differential abilities of Microdata, RDFa, and JSON-LD, particularly as they pertain to allowing for multiple context files and property types for a single instance. Certain markup standards are simply more expressive than others for a given ontology. While I have not done additional empirical research on this matter, I would expect new work to support that suspicion since there are many other markup standards that are not supported by Schema.org and are used for other semantic web ontologies. Each with their own set of affordances and constraints.

Chapter six discusses the underlying processes of development. It does so in two main ways. First, it outlines the underlying design philosophies and systems of thought that ontologists use to craft Schema.org. Second, it investigates the development and deployment of the ontology in terms of the interactions between ontologists and users. I find that the ontologists and users alike, adopt a pragmatic view of ontologies and their work. By basing many of their development decisions on actual usage on the Web, they opt for an empirically grounded approach to model world relations in a way that they believe best resonates with their understanding of it. Furthermore, they do not assume that they have a necessarily correct world image in the way that past work has portrayed knowledge engineers and ontologists (Forsythe 1993, 2001; Ribes and Bowker 2009; Waller 2016). Instead, Schema.org's ontologists debate both the existing state of the world and the best way to represent it. They attempt to arrive at consensus for both the world image and the way to best model that image. Central to this view is that they recognize the problem inherent to their task. As a means to navigate the tension between precision and applicability ontologists and users try to make the ontology as generalizable as possible. This helps to make the ontology easier to use and more adaptable across domains, as ontologists develop in a way that permits broad interpretation and users deploy markup in creative ways, something particularly useful for navigating the limitations discussed in chapter five.

By adopting this approach ontologists and users create a complication of context. While ontologists use such design models to allow their ontologies to be deployed independent of particular domain contexts, they actually determine context in both local and global ways. Locally, contexts are removed from their initial development and the reasserted as terms and

properties are used across domains. It serves to impute a specific context to a particular application at that local level, despite the sometimes problematic imputation. Simultaneously, the ontology determines global context by constructing a bounded meaning space. This confirms the concern raised in recent work that an ontology constructs a world taken as a whole (Ford and Graham 2016; Halford et al. 2013; McCarthy 2017; Waller 2016). Ontological states are circumscribed at both levels of granularity. At each level, what can and cannot be said or interpreted is dictated by the ontology. At the local level, what can be said in a domain is limited to what domain specific knowledge is encoded in the ontology, or what is generalizable from others. At the global level, what can be said in total, as well as where the localities sit in relation to others, is defined and exposed by its existence in the ontology, or occluded by its omission (Knobel 2010).

This contributes to, and extends recent work on practice in STS. Following this "turn to ontology" (Lynch 2013; Woolgar and Lezuan 2013, 2015; Sismondo 2015). I do not make the assumption that there is a picture of the world that ontologists are getting right or wrong. Instead I add to this turn, by exploring how ontologists and users are engaged in negotiating with the various situated components of the data assemblage. These engagements enact a particular understanding of the world for machines, but are dependent on the ontology used, the standards deployed, the way markup is performed, and the content that is encoded. Thus, ontology in this respect can be said to be multiple, particularly as contexts are continually re-determined (Mol 2002; Sismondo 2015). However, while STS work on multiple ontologies eventually converge to enact the thing in question, for the semantic web, ontologies remain multiple. That is, their states do not converge to create a unified thing. Additionally, I add to

this work by showing precisely how "being" is premediated by the technological systems (Aneesh 2017).

This also contributes to gaps that exist in our understanding of AI. In this study, I attempt to reinvigorate long dormant work on AI and expert systems from the slumbers of the AI winter[22]. Present day ontologists, and the knowledge engineers of early AI systems have different views of what they are doing. My data suggests the opposite of the work by Agre (1995), Collins (1987, 1990), Forsythe (1993a, 1993b), and Star (1995). Rather than seeing knowledge as something easily extracted or converted for machine understanding, the Schema.org ontologists recognize the difficulty of making these translations. While the knowledge engineers in that work see the problem as simply one of managing human description errors, these ontologists see the problem as a matter of pragmatic consensus building. Where consensus is agreement amongst user, experts, ontologists, and the ontological orderings of Schema.org. They understand that they are choosing the terms and relations, and understand that a trade-off between precision and practicality is usually inevitable. Thus their approach to development reflects their interpretations of the state of users, consumers, the state of the Web, and the current state of the ontology. Furthermore, I fill the gap left by past work that ignores translations between humans and machines (Agre 1995; Collins 1990; Forsythe 1993a, 1993b; Ribes and Bowker 2009) through their specific focus on the translation between different experts and cultural communities.

[22] This term reflects the temporary abandonment of AI across governments, academics, popular press, tech companies, and venture capital as computing power and actual applications could not live up to the promises of artificial intelligence. See Wikipedia (2017) for further details.

Where the space for expressing semantics is constrained by ontological consolidation and frictions between human and machine translations, Web consumers are provided with limited understandings of the world. Of course while humans have the capacity to exercise their socio-cultural knowledge (Collins 1990) to interrogate and develop meaning from information, they are limited where the provenance, methodology, and contexts of that information is hidden. Overall, these questions contribute broadly to sociological concerns of information asymmetries and access. As the AI winter thaws, and Web environments continually shift in ways that make that contextual information less visible and answers more ready to hand, understanding how systems like AI and semantic search are developed becomes more important. While the semantic web and AI mediated search offer many promises for efficiency and knowledge acquisition, a critical analytic eye must remain on their implementations.

This study is not without its limitations. I can identity at least two. First, this study only examines one semantic web ontology. Ideally, a study that attempts to understand the multiple ways that ontological states are determined for AI and other computational agents would be able to examine the multiple ways that they are developed across ontologies. As I argue throughout this work, the given state is a function of the complex interactions between assemblage components. Thus a comparative study would have been illustrative on this front. Problematically, I know of no such comparative site for research. There are relatively few ontologies that are actively developed, and fewer still that are developed as openly and with as much supporting evidence as Schema.org. Wikidata does offer some window into its construction, but it is not as rich or comprehensive as Schema.org. Additionally, it does not have the same type of community involvement, so there is not a similar level of exposure to the

ontology's user base. Despite this limitation, my research convincingly shows the ways that the interactions between components produce a specific ontological state. I think that my work, and the limited work on other ontologies, allows for the logical assumption that the specific interactions in other ontological assemblages might produce other states.

A second limitation, noted in the methods chapter, is that I was not able to conduct in-depth interviews with community members. While the data used contain all of the public discussion on how to best alter and apply the ontology, it does not contain any backchannel communications. My research indicated that backchannel communications happened regularly between steering group members. Most of the time those discussions are disclosed to the group with a summary of the results, but the actual content of some of those discussions is lost. This may be fine for continued work on the project from the perspective of a developer, but from an analytic perspective, there could be information important to my findings, for or against. Additionally, since the data are entirely textual and interviews were not performed for this study, there is no possibility for follow up or clarification questions that are not covered or alluded to in the two main research sites. However, the vast majority of decisions and rationales, as well as all modifications and development were public and so these limitations seem negligible.

One final aspect of this research that is missing, but perhaps not a limitation per se, is that there was no practicable way to investigate the actual displayed and final form of user's markup. Markup is available in source code, but the sheer volume of markup applied is beyond the scope of even a team of researchers. However, a systematic survey of semantic markup in use could allow for a discursive analysis similar to the type used by Ford and Graham (2016).

While one could certainly interrogate sematic markup, problematically, its provenance is lost. This is not a limitation of this study, as it is outside of the scope and scale here. Additionally, because these systems are deterministic, one can fully comprehend the range of possibility by understanding the standards and ontology coverage at a given time. However, it poses interesting empirical questions about in situ deployment of Schema.org markup.

This study closes some questions, but doing so, opens others. Recent work in data and code studies has been largely oriented towards critique. While my research certainly took a critical eye towards the limitations of Schema.org and the semantic web more generally, critique in a more critical theoretic tradition was not my aim. However, following from that tradition an analysis of the power relationship that exist within and through the semantic web's development would be productive. Work in this area would do well to take critical feminist and postcolonial approaches to the ways that the semantic web builds worlds to foreclose on certain knowledges while expanding on others, such as the decidedly Western and commercial orientation Schema.org has. Such analyses might adopt critical discourse analyses of the coverage of ontologies by mapping out what different domains are covered by major ontologies, what sort of equivalencies they establish, and how they reflect and reinforce dominant power relationships that exist at large.

Additional work might look at other forms of bias that enter into technical systems. Such critical approaches might also investigate the ways in which peer production creates its own set of biases on technical systems. Conventionally, peer production would seem like a way to mitigate the existence of bias in code based systems, but as other work has shown, biases that exist at larger levels of social organization often manifest themselves in technical systems

(Barocas and Selbst 2016; Friedman and Nissenbaum 1996; Noble 2013; Sweeny 2013). This is particularly the case with AI systems that rely on deep neural networks which are unavailable for even machine learning experts to scrutinize.

Currently, semantic technologies are developed manually by individual or teams of ontologists. They debate the ways to define terms, relationships, and their domain contexts all for users to apply to their content. The processes outlined by various scholars all involve the situated practical activity of individuals to translate expertise and organize knowledge in a way that is amenable to computer scientists (Collins 1990; Forsythe 1993; Khazaree and Khoo 2011; McCarthy 2017; Millerand & Bowker 2009, Ribes and Bowker 2009, Ribes and Finholt 2009). In this way, the process of constructing ontologies is still available for scrutiny. However, recent advances in natural language processing and state of the art machine learning algorithms like Word2vec promise to automate these processes. At the moment these famously black boxed neural network algorithms are too computationally expensive to run at Web scale, but as processing units become more powerful, smaller, and more efficient, we may soon see Pandora's box close for good.

References

Abbate, Janet. 1999. *Inventing the Internet*. Cambridge, MA: MIT Press.

Agre, Philip E. 1995. "The Soul Gained and Lost: Artificial Intelligence as a Philosophical Project." *Stanford Humanities Review* 4(2): 1-19.

Agre, Philip E. 2003. "Writing and Representation." Pp. 281-303 in *Narrative Intelligence,* edited by M. Mateas and P. Sengers. Amsterdam: John Benjamins Publishing Company.

Alač, Morena. 2008. "Working with Brain Scans: Digital Images and Gestural Interaction in an fMRI Laboratory." *Social Studies of Science* 38(4): 483–508.

Alvesson, Mats. and Dan Karreman. 2000. "Varieties of discourse: On the study of organizations through discourse analysis." *Human relations* 53(9): 1125-1149.

American Sociological Association. 2008. "Code of Ethics." Retrieved January 6, 2017 (http://www.asanet.org/sites/default/files/code_of_ethics.pdf).

Anderson, Neil. 2005. "Building Digital Capacities in Remote Communities Within Developing Countries: Practical Applications and Ethical Issues." *Information Technology, Education and Society* 6(3): 43-54.

Andrejevic, Mark. 2014. Big data, Big Questions: The Big Data Divide. *International Journal of Communication* 8: 1673-1689.

Aneesh, Aneesh. 2017. "Epilogue: The Global Horizon for Technoscience." *Science, Technology, and Society* 22(1): 124-127

Arora, Payal. 2016. Bottom of the Data Pyramid: Big Data and the Global South. *International Journal of Communication* 10: 1681-1699.

Bakhtin, Mikhail M. 1981. *The Dialogic Imagination: Four Essays*. Translated by Michael Holquist and Caryl Emerson. Austin, TX: University of Texas Press.

Barad, Karen. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglements of Matter and Meaning*. Durham, NC: Duke University Press.

Barocas, Solon. and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104: 1-62.

Becker, Howard. 1982. *Art Worlds*. Berkeley, CA: University of California Press.

Berg, Bruce L. 2004. *Qualitative Research Methods for the Social Sciences*. 5th ed. Boston, MA: Allyn and Bacon.

Berg, Marc. 1997. "Of Forms, Containers, and the Electronic Medical Record: Some Tools for a Sociology of the Formal." *Science, Technology, & Human Values* 22(4): 403-433.

Berners-Lee, Tim, Hendler, Jim and Lassila, Ora. 2001. "The Semantic Web." *Scientific American* 284(5): 34–43.

Bing. 2017. Retrieved January 17, 2017 (https://www.bing.com/ webmaster/help/webmaster-guidelines-30fba23a).

Boellstorff, Tom. 2013. "Making Big Data, in Theory." *First Monday* 18(10).

Bourdieu, Pierre. 1977. *Outline of a Theory of Practice*. Cambridge, UK: Cambridge University Press.

Bowker, Geoffrey C. 2005. *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Bowker, Geoffrey C. 2014. "The Theory/Data Thing." *International Journal of Communication* 8: 1795–1799.

Bowker, Geoffrey C., and Susan Leigh Star. 1996. "How Things (Actor-Net) Work: Classification, Magic and the Ubiquity of Standards." *Philosophia* 25(3–4):195–220.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press.

Bowker, Geoffrey C., Karen Baker, Florence Millerand and David Ribes.2010. "Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment." Pp. 97-117 in *International Handbook of Internet Research*, edited by Jeremy Hunsinger, Lisbeth Klastrup and Matthew M. Allen. New York: Springer.

boyd, danah and Kate Crawford. 2012. Critical Questions for Big Data. *Information, Communication and Society* 15(5): 662–679.

Brekhus, Wayne. 2007. "The Rutgers School: A Zerubavelian Culturalist Cognitive Sociology." *European Journal of Social Theory* 10(3):448–464.

Brickley, Dan. 2016. "How we work" Retrieved January 17, 2017 (https://www.w3.org/community/schemaorg/how-we-work/).

Brine, Kevin R. and Mary Poovey. 2013. "From Measuring Desire to Quantifying Expectations: A Late Nineteenth-Century Effort to Marry Economic Theory and Data." Pp. 61-76 in *'Raw*

Data' is an Oxymoron, edited by Lisa Gitelman. Cambridge, MA: MIT Press.

Busch, Lawrence .2011. *Standards: Recipes for reality*. Cambridge, MA: MIT Press.

Busch, Lawrence. 2014. "A Dozen Ways to Get Lost in Translation: Inherent Challenges in Large Scale Data Sets." *International Journal of Communication* 8: 1727-1744.

Chaffey, Dave. 2017. Retrieved March 3, 2017 (http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/mobile-marketing-statistics/).

Cheney-Lippold, John. 2011. "A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control." *Theory, Culture & Society* 28(6): 164–181.

Coleman, Gabriella. 2004. "The Political Agnosticism of Free and Open Source Software and the Inadvertent Politics of Contrast." *Anthropological Quarterly* 77(3): 507–519.

Collins, Harry M. 1990. Artificial Experts: Social Knowledge and Intelligent Machines. Cambridge, MA: MIT Press.

Collins, Harry M. 2012. Expert Systems and the Science of Knowledge." Pp. 321-340 in *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*, edited by Wiebe Bijker, Thomas Hughes, Trevor Pinch, and Deborah Douglas. Cambridge, MA: MIT Press.

Coopmans, Catelijne, Janet Vertesi, Michael E. Lynch and Steve Woolgar. 2014. "Introduction." Pp, 1-14 in *Representation in Scientific Practice Revisited*, edited by Catelijne Coopmans, Janet Vertesi, Michael E. Lynch and Steve Woolgar. Cambridge, MA: MIT Press.

Coyle, Karen. 2008. "Meaning, Technology, and the Semantic Web." *The Journal of Academic Librarianship 34*(3): 263-264.

DeLanda, Manuel. 2006. *A New Philosophy of Society: Assemblage Theory and Social Complexity*. London: Bloomsbury.

DeLanda, Manuel. 2011. *Philosophy and Simulation: The Emergence of Synthetic Reason*. London New York: Continuum.

Deleuze, Gilles. 1964 [1995]. *Difference and Repetition*. Translated by Paul Patton. New York: Columbia University Press.

Deleuze, Gilles and Felix Guattari. 1987. *A Thousand Plateaus: Capitalism and Schizophrenia.* Translated by Brian Massumi. Minneapolis, MN and London: University of Minnesota Press.

DeNardis, Laura. 2009. *Protocol Politics: The Globalization of Internet Governance*. MIT Press.

Derrida, Jacques. 1967[2016]. Of Grammatology. Translated by Gayatri Chakravorty Spivak. Baltimore, MD: John Hopkins University Press.

DiMaggio, Paul J. 1987. "Classification in Art." American Sociological Review 52: 440–455.

DiMaggio, Paul J and Walter W. Powell. 1983. "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields." *American Sociological Review* 48: 147–160.

DiMaggio, Paul J. and Walter W. Powell. 1991. "Introduction." Pp.1–38 in The New Institutionalism in Organizational Analysis, edited by Walter. W. Powell and Paul. J. DiMaggio. Chicago: University of Chicago Press

Douglas, Mary. 1986. *How Institutions Think*. Syracuse, NY: Syracuse University Press.

Edwards, Paul N. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Edwards, Paul N., Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L. Borgman. 2011. Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science* 41(5), 667–690.

Epstein, Steven. 2000. *Inclusion: The Politics of Difference in Medical Research*. Chicago: University of Chicago Press.

Epstein, Steven. 2009. "Beyond the Standard Human?" Pp. 35-54 in *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*, edited by Martha Lampland and Susan Leigh Star. Ithaca, NY: Cornell University Press.

Fetto, John. 2016. *Mobile Search: Topics and Themes*. Retrieved January 17, 2017 (http://hitwise.connexity.com/070116_MobileSearchReport_CD_US.html).

Fiormonte, Domenico, Desmond Schmidt, Paolo Monella, and Paolo Sordi. 2015. "The Politics of Code. How Digital Representations and Languages Shape Culture." *ISIS Summit Vienna 2015:* 66–68.

Floridi, Luciano. 2012. Big Data and Their Epistemological Challenge. *Philosophy & Technology*, 25(4): 435–437.

Ford, Heather, and Mark Graham. 2016. "Semantic Cities: Coded Geopolitics and the Rise of the Semantic Web." Pp. 200-214 In *Code and the City*, edited by Rob Kitchin and Sung-Yeh Perng. London: Routledge.

Forsythe, Diana E. 1993b. The Construction of Work in Artificial Intelligence. *Science, Technology & Human Values* 18(4): 460-479.

Forsythe, Diana E. 1993a. Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. *Social Studies of Science* 23(3): 445-477.

Foucault, Michel. 1973 [1994]. The Order of Things An Archeology of the Human Sciences. New York: Vintage.

Foucault, Michel. 1977 [1995]. Discipline and Punish: The Birth of the Prison. New York. Vintage.

French, Martin. 2014. "Gaps in the Gaze: Informatic Practice and the Work of Public Health Surveillance." Surveillance & Society, 12(2): 226-243.

Friedman, Batya and Helen Nissenbaum. 1996. "Bias in Computer Systems. *ACM Transactions on Information Systems* 14(3): 330–347.

Friese, Carrie. 2010. Classification Conundrums: Categorizing Chimeras and Enacting Species Preservation. *Theory and Society* 39(2): 145–172.

Galloway, Alexander R. 2004. Protocol: How Control Exists After Decentralization. Cambridge, MA: MIT Press.

Galloway, Alexander R. 2006. Protocol. *Theory, Culture & Society*, *23*(2-3): 317-320.

Galloway, Alexander R. 2012. *The Interface Effect*. Malden, MA: Polity.

Gandy Jr., Oscar. 1993. The Panoptic Sort: A Political Economy of Personal Information. Boulder, CO. Westview Press.

Genschel Philipp. 1997. "How Fragmentation Can Improve Co-ordination: Setting Standards in International Telecommunications." *Organization Studies* 18: 603–622.

Gillespie, Tarelton. 2014. "The Relevance of Algorithms". Pp. 167-194 in *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton. Gillespie, Pablo J. Boczkowski and Kirsten A. Foot. Cambridge, MA: MIT Press.

Ginsberg, Allen. 2008. "Ontological Indeterminacy and the Semantic Web." *International Journal on Semantic Web and Information Systems (IJSWIS)* 4(2): 19-48.

Gitelman, Lisa and Virginia Jackson. 2013. "Introduction". Pp. 1-14 in 'Raw Data' is an Oxymoron, edited by Lisa Gitelman. Cambridge, MA: MIT Press.

Giri, Kaushal. 2011. “Role of Ontology in Semantic Web.” *DESIDOC Journal of Library & Information Technology*, 31(2): 116-120.

Goffman, Erving. 1959. *The Presentation of Self in Everyday Life*. New York: Anchor Books.

Google. 2017. Retrieved January 17, 2017 (https://support.google.com/webmasters/answer/3069489?hl=en).

Guha, Ramanathan V. 2014. “Search, Structure, and Knowledge on the Web.” Presented at *SemTech*, San Jose, CA.

Guha, Ramanathan V., Dan Brickley, and Steve MacBeth. 2015. “Schema.org: Evolution of Structured Data on the Web. *ACM Queue 13*(9): 1-28

Hacking, Ian. 1986. “Making Up People.” Pp. 222-236 in *Reconstructing Individualism*, edited by Thomas Heller. Palo Alto, CA: Stanford University Press

Hacking, Ian. 1999. *The Social Construction of What?*. Cambridge, MA: Harvard University Press.

Halavais, Alexander. 2009. *Search Engine Society*. Cambridge, UK: Polity Press.

Halford, Susan, Catherine Pope and Mark Weal. 2013. “Digital Futures? Sociological Challenges and Opportunities in the Emergent Semantic Web.” *Sociology* 47(1): 173–189.

Halpin, Harry, Patrick Hayes, James P. McCusker, Deborah L. McGuiness and Henry Thompson. 2010. “When owl:sameAs Isn’t the Same: An Analysis of Identity in Linked Data.” Pp. 305-320 In *The Semantic Web - ISWC 2010: 9th International Semantic Web Conference pt. 1,* edited by Peter F. Patel-Schneider, Yue Pan, Pascal Litzer, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks and Birte Glimm. Heidelberg: Springer.

Halpin, Michael. 2016. “The DSM and Professional Practice: Research, Clinical, and Institutional Perspectives.” *Journal of Health and Social Behavior* 57(2): 153–167.

Halupka, Max and Cassandra Star. 2011. The Utilisation of Direct Democracy and Meritocracy in the Decision Making Process of the Decentralized Virtual Community Anonymous. In *Australian Political Studies Association Conference*. Canberra, AT.

Hanseth, Ole, Eric Monteiro and Morten Hatling. 1996. Developing Information Infrastructure: the Tension Between Standardisation and Flexibility. *Science, Technology and Human Values* 21(4): 407–426.

Haraway, Donna. 1991. Simians, Cyborgs, and Women: The Reinvention of Nature. Abingdon, UK: Routledge.

Heidegger, Martin. 1973. *Being and Time*. Translated by John Macquarrie and Edward Robinson. Oxford: Basil Blackwell.

Heimer, Carol A. 2001.Cases and Biographies: An Essay on Routinization and the Nature of Comparison." *Annual Review of Sociology* 27(1): 47-76.

Herman, Ivan, Ben Adida, Manu Sporny and Mark Birbeck. 2015. Retrieved September 13, 2015 https://www.w3.org/TR/xhtml-rdfa-primer/.

Hickson, Ian. 2013. "HTML Microdata" Retrieved September 13, 2015 (https://www.w3.org/TR/microdata/).

Hine, Christine. 2000. *Virtual Ethnography*. London: Sage,

Hoffman, Steve G. 2015. "Thinking Science with Thinking Machines: The Multiple Realities of Basic and Applied Knowledge in a Research Border Zone." *Social Studies of Science* 45(2): 242–269.

Hsu, Wendy. 2014. "Digital Ethnography Towards Augmented Empiricism: A New Methodological Framework." *Journal of Digital Humanities* (3)1. Retrieved 1/22/2017 (http://journalofdigitalhumanities.org/3-1/digital-ethnography-toward-augmented-empiricism-by-wendy-hsu/).

Introna, Lucas D. 2005. "Disclosive Ethics and Information Technology: Disclosing Facial Recognition Systems." *Ethics and Information Technology* 7(2): 75–86.

Introna, Lucas D. 2014. "Towards a Post-Human Intra-Actional Account of Sociomaterial Agency (and Morality)." *The Moral Status of Technical Artefacts.* 31–53.

Introna, Lucas D. and Helen Nissenbaum. 2000. "Shaping the Web: Why the Politics of Search Engines Matters" *The Information Society* 16(3): 169–185.

Introna, Lucas D. and Niall Hayes. 2011. "On Sociomaterial Imbrications: What Plagiarism Detection Systems Reveal and Why It Matters." *Information and Organization* 21(2): 107–122.

Jansen, Bernard J., Amanda Spink and Tefko Saracevic. 2000. "Real Life, Real Users and Real Needs: A Study and Analysis of Users' Queries on the Web." *Information Processing and Management* 36(2): 207-227.

Jansen, Bernard J. and Amanda Spink. 2003. "An Analysis of Web Information Seeking and Use: Documents Retrieved Versus Documents Viewed." Pp. 65-69 In *Proceedings of the 4th International Conference on Internet Computing*, Las Vegas, Nevada. June 23-26, 2003.

Kallinikos, Jannis, Aleksi Aaltonen and Attila Marton. 2013. “The Ambivalent Ontology of Digital Artifacts.” *MIS Quarterly* 37(2): 357–370.

Khazaree, Emad and Michael Khoo. 2011. “Practice-Based Ontologies: A New Approach to Address the Challenges of Ontology and Knowledge Representation in History and Archaeology.” Pp. 375-386 in *Metadata and Semantic Research*, edited by Elena Garcia-Barriocanal, Zeynel Cebeci, Aydin Ozturk and Mehmet C. Okur. Berlin: Springer-Verlag Berlin Heidelberg.

Kitchin, Rob. 2014a. “Thinking Critically About and Researching Algorithms.” *Programmable City Working Paper #5*. Retrieved July 13, 2015 (http://ssrn.com/abstract=2515786).

Kitchin, Rob. 2014b. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. New York: Sage Publications.

Kitchin, Rob. 2014c. “Big Data, New Epistemologies and Paradigm Shifts.” *Big Data & Society,* 1(1): 1-12.

Kitchin, Rob. and Dodge, Martin. 2011. *Code/Space: Software and Everyday Life*. Cambridge, MA: MIT Press.

Kitchin, Rob, & Tracey P. Lauriault. 2014. “Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work.” *Geoweb and Big Data*, 1–19. Retrieved July 13, 2015 (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2474112).

Kitchin, Rob, & Tracey P. Lauriault and Gavin McArdle. 2015a. “Urban indicators and dashboards: epistemology, contradictions and power/knowledge.” *Regional Studies, Regional Science* 2(1): 43–45.

Kitchin, Rob, & Tracey P. Lauriault and Gavin McArdle. 2015b. “Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards.” *Regional Studies, Regional Science* 2(1): 6–28.

Klose, Alexander. 2009. *The Container Principle*. Cambridge, MA: MIT Press

Knobel, Cory P. 2010. “Ontic Occlusion and Exposure in Sociotechnical Systems.” Ph.D. dissertation, Department of Information Studies, The University of Michigan, Ann Arbor.

Knorr-Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge.* Cambridge, MA: Harvard University Press.

Knorr-Cetina, Karin. 2001. “Objectual Practice.” Pp. 175-188 In *The Practice Turn in Contemporary Theory,* edited by Theodore Schatzki, Karin Knorr-Cetina & Eike von

Savigny. London: Routledge.

Koehne, Kristy, Bridget E. Hamilton, Natasha Sands and Cathy Humphreys. 2013. "Working Around a Contested Diagnosis: Borderline Personality Disorder in Adolescence." *Health* 17(1): 37–56.

Kozinets, Robert V. 2010. *Netnography: Doing Ethnographic Research Online*. London: SAGE Publications.

Kress, Gunther and Theo van Leeuwen. 2006. *Reading Images: The Grammar of Visual Design* (2nd Ed). London: Routledge.

Krippner, Greta. 2011. *Capitalizing on Crisis: The Political Origins of the Rise of Finance*. Cambridge, MA: Harvard University Press.

Lakoff, George. 1987. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*. Chicago, IL: Chicago University Press

Latour, Bruno .1987. *Science in Action: How to Follow Scientists and Engineers Through Society*. Princeton, NJ: Princeton University Press.

Latour, Bruno. 1988. *The pasteurization of France.* Translated by Alan Sheridan and John Law. Cambridge, MA: Harvard University Press.

Latour, Bruno. 1991. *We Have Never Been Modern*. Translated by Catherine Porter. Cambridge, MA: Harvard University Press.

Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor Network Theory.* Oxford, UK: Oxford University Press.

Latour, Bruno and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts.* Princeton, NJ: Princeton University Press.

Lauriault, Tracey P. 2012. "Data, Infrastructures and Geographical Imaginations." Ph.D. dissertation, Department of Geography, Carleton University, Ottawa.

Law, John. 2010. "The Materials of STS" Retrieved July 13, 2015 (http://www.heterogeneities.net/publications/Law2008MaterialsofSTS.pdf).

Law, John. and Marianne E. Lien. 2012. "Slippery: Field Notes in Empirical Ontology." *Social Studies of Science* 43(3): 363–378.

Legg, Catherine. 2013. "Peirce, Meaning, and the Semantic Web." *Semiotica* 193: 119-143.

Lengwiler, Martin. 2009. “Double Standards: The History of Standardizing Humans in Modern Life Insurance.” Pp. 95-114 in *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*, edited by Martha Lampland and Susan Leigh Star. Ithaca, NY: Cornell University Press.

Leonelli, Sabina. 2014. “What Difference Does Quantity Make? On the Epistemology of Big Data in Biology.” *Big Data & Society* 1(1): 1-11. DOI: 10.1177/2053951714534395

Lupton, Deborah. 2014. “Apps as Artefacts: Towards a Critical Perspective on Mobile Health and Medical Apps. *Societies* 4(4): 606–622.

Lustig, Caitlin and Bonnie Nardi. 2015. “Algorithmic Authority: The Case of Bitcoin.” In *Proceedings of the Annual Hawaii International Conference on System Sciences,* 2015–March: 743–752. http://doi.org/10.1109/HICSS.2015.95

Lynch. Michael. 2006. "The production of scientific images: Vision and re-vision in the history, philosophy, and sociology of science." Pp. 26-40 In *Visual cultures of science: Rethinking representational practices in knowledge building and science communication*, edited by Luc Pauwels. Lebanon, NH: Dartmouth College Press.

Lyon, David. 2007. *Surveillance Studies: An Overview*. Cambridge, UK: Polity Press

MacCormick, John. 2013. *Nine Algorithms That Changed the Future: The Ingenious Ideas That Drive Today's Computers*. Princeton, NJ: Princeton University Press.

Mackenzie, Adrian. 2005. “The Performativity of Code: Software and Cultures of Circulation.” *Theory, Culture and Society* 22(1): 71-92.

Manovich, Lev. 2013. *Software takes control*. New York: Bloomsbury.

Markham, Annette N. 1998. *Life Online: Researching Real Experiences in Virtual Space*. Lanham, MD: AltaMira Press.

May, Todd. 2005. *Gilles Deleuze: An Introduction*. Cambridge, UK: Cambridge University Press.

McCarthy, Matthew T. 2016. “The Big Data Divide and its Consequences.” *Sociology Compass*, *10*(12): 1131-1140.

McCarthy, Matthew T. 2017. “The Semantic Web and Its Entanglements.” *Science, Technology and Society 22*(1): 21-37.

Meller, Hilla and Pascal Cohen. 2015. “The State of the Mobile Web.” SimilarWeb retrieved January 15, 2017 (https://www.similarweb.com/corp/the-state-of-mobile-web-in-the-us-2015/).

Meyer, John and Brian Rowan. 1977. “Institutionalized Organization: Formal Structure as Myth and Ceremony.” *American Journal of Sociology* 83: 340–63.

Millerand, Florence and Geoffrey C. Bowker. 2009. “Metadata Standards: Trajectories and Enactment in the Life of an Ontology.” Pp. 149-166 in *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*, edited by Martha Lampland and Susan Leigh Star. Ithaca, NY: Cornell University Press.

Mittelstadt, Brent D. and Luciano Floridi. 2015. “The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics* 22(2): 303–341.

Mol, Annemarie. 2002. *The Body Multiple: Ontology in Medical Practice*. Durham, NC: Duke University Press.

Moser, Ingunn. 2008. “Making Alzheimer’s disease matter. Enacting, interfering and doing politics of nature.” *Geoforum* 39(1): 98-110.

Nevile, Charles. 2014. “*Schema.org – What, how, why?”* Presented at Open Web Camp. San Jose, CA. Retrieved July 13, 2015 (https://www.youtube.com/watch?v=hcahQfN5u9Y).

Neyland, Daniel. 2014. “On Organizing Algorithms.” *Theory, Culture & Society* 32(1): 119–132.

Nissenbaum, Helen. 2004. “Privacy as Contextual Integrity.” *Washington Law Review* 79(119): 1-31.

Noble David F. 1984. *Forces of Production: A Social History of Industrial Automation*. Oxford/New York: Oxford University Press.

Noble, Safyia U. 2013. “Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible.” *InVisible Culture* (19). (http://ivc.lib.rochester.edu/google-search-hyper-visibility-as-a-means-of-rendering-black-women-and-girls-invisible/).

O’Halloran, Kay L. 2011. Multimodal discourse analysis. Pp. 120-136 in *Continuum companion to discourse analysis*, edited by Ken Hyland. London: Bloomsbury.

O'Halloran, Kay L. 2005. *Mathematical Discourse: Language, Symbolism and Visual Images*. London and New York: Continuum.

Oremus, Will. 2014. Here Are All the Different Genders You Can be on Facebook. Retrieved February 20, 2017. (http://www.slate.com/blogs/future_tense/ 2014/02/13/ facebook_custom_gender_options_here_are_all_56_custom_options.html.)

O'Toole, M. 1994. *The Language of Displayed Art*. London: Leicester University Press.

Parchoma, Gale. 2012. “The Contested Ontology of Affordances: Implications for Researching Technological Affordances for Collaborative Knowledge Production.” *Computers in Human Behavior* 37:360–368.

Poirier, Lindsay. 2015. “The Stickiness of Difference in the Semantic Web.” *Cyborgology* Retrieved September 5, 2015 (http://thesocietypages.org/cyborgology/2015/07/20/the-stickiness-of-difference-in-the-semantic-web).

Raley, Rita. 2013. “Dataveillance and Counterveillance.” Pp. 121-146 in *Raw data is an Oxymoron*, edited by Lisa Gitelman. Cambridge, MA: MIT Press.

Ribes, David and Geoffrey C. Bowker. 2009. Between meaning and machine: Learning to Represent the Knowledge of Communities. *Information and Organization* 19(4): 199–217.

Ribes, David and Steven J. Jackson. 2013. “Data Bite Man: The Work of Sustaining a Long-Term Study.” Pp. 147-166 in Raw data is an Oxymoron, edited by Lisa Gitelman and Virginia Jackson. Cambridge, MA: MIT Press.

Ribes, David and Thomas A. Finholt. 2009. “The Long Now of Technology Infrastructure: Articulating Tensions in Development.” *Journal of the Association for Information Systems* 10(5): 375-398.

Rogers, Richard. 2013. *Digital Methods*. Cambridge, MA: MIT Press

Sandholtz, Kurt W. 2012. Making Standards Stick: A Theory of Coupled vs. Decoupled Compliance.” *Organization Studies* 33(5–6): 655–679.

Suchman, Lucy A. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge: Cambridge University Press.

Shankar, Kalpana, Kristin Eschenfelder and Greg Downey. 2016. “Studying the History of Social Science Data Archives as Knowledge Infrastructure.” *Science and Technology Studies* 29(2): 62–73.

Schatzki, Theodore. 2001. “Introduction: Practice theory.” Pp. 10-23 In *The Practice Turn in Contemporary Theory,* edited by Theodore Schatzki, Karin Knorr-Cetina and Eike von Savigny. London: Routledge.

Schema.org. 2017a. “Welcome to Schema.org.” Retrieved January 1, 2017 (http://schema.org/).

Schema.org. 2017b. “About Schema.org.” Retrieved January 1, 2017 (http://schema.org/docs/about.html).

Schema.org. 2017c. “FAQ.” Retrieved January 1, 2017 (http://schema.org/docs/faq.html).

Schema.org. 2017d. “GovernmentPermit.” Retrieved May 21, 2017 (http://schema.org/GovernmentPermit)

Schema.org. 2017e. “CollegeOrUniversity.” Retrieved May 21, 2017 (http://schema.org/CollegeOrUniversity).

Schutz, Alfred and Thomas Luckmann. 1973. *The Structures of the Life-World Volume 1*. Evanston, IL: Northwestern University Press.

Schutz, Alfred and Thomas Luckmann. 1989. *The Structures of the Life-World Volume 2*. Evanston, IL: Northwestern University Press.

Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven, CT: Yale University Press.

Seaver, Nick. 2013. “Knowing Algorithms.” Presented at *Media in Transition 8*, Cambridge, MA. Retrieved January 17, 2017 (http://nickseaver.net/papers/seaverMiT8.pdf).

Sismondo, Simon. 2015. “Ontological Turns, Turnoffs and Roundabouts.” *Social Studies of Science* 45(3): 441–448.

Solove, David J. 2004. *The Digital Person: Technology and Privacy in the Information Age.* New York: New York University Press.

Spink, Amanda and Bernard J. Jansen. 2004. “A Study of Web Search Trends.” *Webology* 1(2): (http://www.webology.org/2004/v1n2/a4.html).

Spivak, Gayatri C. 1999. *A Critique of Postcolonial Reason*. Cambridge, MA: Harvard University Press.

Sporny, Manu. 2015. “RDFa Lite 1.1.” Retrieved May 21, 2017 (https://www.w3.org/TR/rdfa-lite/).

Sporny, Manu, Dave Longley, Gregg Kellogg, Markus Lanthaler, Niklas Lindstrom. 2014. “JSON-LD 1.0: A JSON-based Serialization for Linked Data.” Retrieved September 13, 2015 (https://www.w3.org/TR/json-ld/).

Springer, Noah J. 2015. “Publics and Counter-publics on the Front Page of the Internet: The Cultural Practices, Technological Affordances, Hybrid Economics and Politics of Reddit’s Public Sphere.” Ph.D. Dissertation, Department of Journalism and Mass Communication, Boulder, CO: University of Colorado.

Stanley, Matthew. 2013. “Where is That Moon, Anyway? The Problem of Interpreting Historical Solar Eclipse Observations.” Pp. 77-88 in *Raw data is an Oxymoron*, edited by Lisa Gitelman and Virginia Jackson. Cambridge, MA: MIT Press.

Star, Susan Leigh. 1995. “Introduction.” Pp. 1-38 in *Ecologies of Knowledge: Work and Politics in Science and Technology*, edited by Susan Leigh Star. Albany, NY: State University of New York Press.

Star, Susan Leigh. 1999. “The Ethnography of Infrastructure.” *American Behavioral Scientist* 43(3): 377-391.

Star, Susan Leigh and Martha Lampland. 2009. “Reckoning With Standards.” Pp. 3-34 in *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*, edited by Martha Lampland and Susan Leigh Star. Ithaca, NY: Cornell University Press.

Strauss, Anslem and Juliet Corbin. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques.* London: Sage Publications.

Sweeney, Latanya. 2013. “Discrimination in online ad delivery.” *ACM Queue* 11(3) (http://queue.acm.org/detail.cfm?id=2460278).

Thatcher, Jim. 2014. “Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data.” *International Journal of Communication* 8: 1765-1783.

Thévenot, Laurent. 2009. “Governing Life by Standards: A View From Engagements.” *Social Studies of Science* 39: 793–813.

Timmermans, Stefan and Steven Epstein. 2010. “A World of Standards but not a Standard World: Toward a Sociology of Standards and Standardization”. *Annual Review of Sociology* 36(1), 69–89.

Traweek, Sharon. 1988. *Beamtimes and Lifetimes: The World of High Energy Physicists*. Harvard University Press: Cambridge, Mass.

Turkle, Sherry. 1995. *Life on Screen: Identity in the Age of the Internet*. New York: Simon & Schuster.

Vaidhayanathan, Siva. 2011. *The Googlization of Everything and Why We Should Worry*. Berkeley, CA: University of California Press.

Vertesi, Janet. 2015. Seeing Like a Rover: How Robots, Teams, and Images Craft Knowledge of Mars. Chicago, IL: University of Chicago Press.

Vis, Farida. 2012. “A Critical Reflection on Big Data.” *First Monday* 1–14. (http://ojs-prod-lib.cc.uic.edu/ojs/index.php/fm/article/view/4878).

Warf, Barney and John Grimes, “Counterhegemonic Discourses and the Internet” *Geographical Review* 87(2): 259-274.

Westbrook, Laurel, & Saperstein, Aliya. 2015. “New Categories Are Not Enough: Rethinking the Measurement of Sex and Gender in Social Surveys.” *Gender & Society* 29(4): 534–560.

Whooley, Owen. 2010. “Diagnostic Ambivalence: Psychiatric Workarounds and the Diagnostic and Statistical Manual of Mental Disorders.” *Sociology of Health & Illness* 32(3): 452–469.

Whooley, Owen. 2014. “Nosological Reflections: The Failure of DSM-5, the Emergence of RDoC, and the Decontextualization of Mental Distress.” *Society and Mental Health* 4(2): 92–110.

Whooley, Owen and Alan Horwitz. 2013. “The Paradox of Professional Success: Grand Ambition, Furious Resistance, and the Derailment of the DSM-5 Revision Process.” Pp. 75–92 in *Making the DSM-5*, edited by Joel. Paris and James. Phillips. New York: Springer.

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): 321–340.

Woolgar, Steve and Javier Lezaun. 2015. “Missing the (Question) Mark? What Is a Turn to Ontology?” *Social Studies of Science* 45(3):462–467.

W3C (World Wide Web Consortium). 2011. “Mixing HTML Data Formats.” Retrieved September, 16, 2015 (https://www.w3.org/wiki/Mixing_HTML_Data_Formats).

W3C (World Wide Web Consortium). 2017. “JSON Intro.” Retrieved February 5, 2017 (https://www.w3schools.com/js/js_json_intro.asp).

WHATWG (Web Hypertext Application Technology Working Group). 2016. “HTML Living Standard.” Retrieved February 5, 2017 (https://html.spec.whatwg.org/multipage/).

Westbrook, Laurel and Aliya Saperstein. 2015. “New Categories Are Not Enough: Rethinking the Measurement of Sex and Gender in Social Surveys.” *Gender & Society* 29(4):534–560.

Wikipedia .2017. “AI Winter.” Retrieved May 10, 2017 (https://en.wikipedia.org/wiki/AI_winter).

Wittgenstein, Ludwig. 2009. *Philosophical Investigations*. Edited by Peter Hacker and Joachim Schulte. Chichester, UK: Wiley-Blackwell.

Yasunori Baba and Ken-ichi Imai. 1993. “A Network View of Innovation and Entrepreneurship: the Case of the Evolution of the VCR Systems.” *International Social Science Journal* 45: 23–34.

Zbaracki, Mark. 1998. “The Rhetoric and Reality of Total Quality Management.” *Administrative Science Quarterly* 43: 602–636.

Zerubavel, Eviatar. 1996. “Lumping and Splitting: Notes on Social Classification.” *Sociological Forum* 11(3):421–433.

Zhao, Wei. 2008. “Social Categories, Classification Systems, and Determinants of Wine Price in the California and French Wine Industries.” *Sociological Perspectives* 51(1):163–199.

Zuckerman, Ezra. 1999. “The Categorical Imperative: Securities Analysts and the Illegitimacy Discount.” *American Journal of Sociology* 104: 1398–438.

Zuckerman, Ezra. 2000. “Focusing the Corporate Product: Securities Analysts and De-diversification.” *Administrative Science Quarterly* 45: 591–619.

Appendix:

Glossary of Terms

Commits: These are the specific changes to a repository.

Enumerations: A set of available categories and the set of values that rest within those categories

Fork: This is the process of marking a separate version of the master repository. When one wants to make a change to the master repository, one creates a fork from the master, makes a specific set of changes, and then uploads those changes for approval.

Issues: These are the main tasks that are created for specific work. On GitHub, they take a threaded form, each with a hyperlinked number. Issues are varied and can be for specific coding tasks, simple documentary changes, goal setting, etc. They are linked to pull requests and commits when action is taken on them. Issues can either be merged with other issues when substantially similar, closed when completed, deemed irrelevant, unimportant, or open when still actively being worked on.

Literals: Specific string values that indicate a value precisely without IRIs

Pull Requests: Pull requests are notifications sent to the working group indicating that a developer has made changes on a project that are awaiting review.

Repository: These are the stores of all versions, code, and metadata for a project.

Matthew T. McCarthy

Department of Sociology
University of Wisconsin – Milwaukee
P.O. Box 413. Bolton Hall 714
3210 N. Maryland Ave. Milwaukee, WI 53201

Education

M.A. University of Wisconsin – Milwaukee, Sociology (2012)
Thesis: Anonymous' and WikiLeaks' War Against Information Asymmetry

B.S. Fitchburg State College (University) – Fitchburg, MA: International Business and Economics (2006)

Dissertation Title

Enacting the Semantic Web: Ontological Orderings, Negotiated Standards, and Human-Machine Translations

Research and Teaching Interests

Science, Knowledge, and Technology
Surveillance Studies
Social Theory
Communication and Information Technologies
Critical Data/Code Studies
Standards and Infrastructure
Social Movements

Peer Reviewed Publications

McCarthy, Matthew T. 2017. "The Semantic Web and Its Entanglements." *Science, Technology, and Society* 22(1): 21-37.

McCarthy, Matthew T. 2016. "The Big Data Divide and Its Consequences." *Sociology Compass* 10(12): 1131-1140.

McCarthy, Matthew T. 2015. "Toward a Free Information Movement." *Sociological Forum,* 30(2): 295-332.

Fellowships and Awards

2015 Best Paper Award, Department of Sociology, University of Wisconsin – Milwaukee

2012 Chancellors Fellowship, University of Wisconsin – Milwaukee $4000

Teaching Experience

Instructor

SOC 376 Modern Sociological Theory (in person) – Fall 2016, 2015, 2014, 2013; Spring 2014

SOC 376 Modern Sociological Theory (online) – Spring 2016; Summer 2015

SOC 103 World Society (online) – Fall 2016; Summer 2016

Teaching Assistant

SOC 760 Advanced Statistical Methods in Sociology w/ Dr. Gordon Gauchat – Spring 2015; w/ Dr. Noelle Chesley – Spring 2013, 2012

SOC 261 Introduction to Statistical Thinking in Sociology w/ Dr. Heeju Shin - Fall 2012; w/ Dr. Pat Rubio Goldsmith – Fall 2011

SOC 241 Criminology w/ Dr. Donald Green – Spring 2011; Fall 2010

Research Assistant

2015 Quantitative data collection/entry – Dr. Gordon Gauchat, Summer

2013 Quantitative data analyst for the Task Force on Family and Medical Leave Report w/ Dr. Stacey Oliker, Dr. Kate Kramer, and Dr. Amanda Seligman, Summer

Conference Presentations

2015 Co-authored presentation: "Big Data and the Rise of the System Identity," *American Sociological Association annual meeting*, Chicago, IL, August

2015 Co-authored presentation: "The Uses of Personality: Personal, Bureaucratic, and System Identities" *University of Wisconsin – Milwaukee Sociology Department Colloquium Series*, Milwaukee, WI, April

2014 Single author presentation: "High Fidelity?: Trust, Attitudes, and ID Systems" *American Sociological Association annual meeting,* San Francisco, CA, August

2013 Single author presentation: "Towards a Free Information Movement: The Cases of WikiLeaks and Anonymous" *Midwest Sociological Society annual meeting,* Chicago, IL, March

2012 Single author presentation: "Anonymous' and WikiLeaks' War Against Information Asymmetry" *University of Wisconsin – Milwaukee Sociology Department Colloquium Series*, Milwaukee, WI, April

Service

2016 Vice President – Sociology Graduate Student Association

Research Software Expertise

R, Tableau, Dedoose, Pajek, SPSS, Stata