

**KNOWLEDGE REPRESENTATION IN SCRUFFY  
WORLDS  
AN ETHNOGRAPHY OF SEMIOTIC  
INFRASTRUCTURE DESIGN WORK**

By

Lindsay Poirier

A Dissertation Submitted to the Graduate  
Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the  
Requirements for the Degree of

**DOCTOR OF PHILOSOPHY**

Major Subject: **SCIENCE AND TECHNOLOGY STUDIES**

Examining Committee:

---

Dr. Kim Fortun, Dissertation Adviser

---

Dr. Michael Fortun, Member

---

Dr. Dean Nieuwma, Member

---

Dr. James Hendler, Member

Rensselaer Polytechnic Institute  
Troy, New York

April 2018  
(For Graduation May 2018)

© Copyright 2018  
by  
Lindsay Poirier  
All Rights Reserved

# CONTENTS

LIST OF TABLES . . . . .	vi
LIST OF FIGURES . . . . .	vii
ACKNOWLEDGMENT . . . . .	viii
ABSTRACT . . . . .	xii
1. INTRODUCTION . . . . .	1
1.1 Introduction . . . . .	1
1.2 Defining Semiotic Infrastructures and Semiotic Technologists . . . . .	6
1.3 Indexing Semiotic Infrastructure Design . . . . .	11
1.4 Encoding Digital Semiotic Expertise . . . . .	16
1.5 Double Binds of Expert Boundary Drawing . . . . .	22
1.6 Ontologically Representing (in) Real Worlds . . . . .	29
1.7 Mapping the Dissertation . . . . .	34
2. THINKING ABOUT SEMIOTIC INFRASTRUCTURE: A GENEALOGY	38
2.1 Information Architecture as Semiotic Infrastructure . . . . .	40
2.2 Infrastructurally Inverting the Platform for Experimental Collaborative Ethnography . . . . .	43
2.2.1 Ethnography . . . . .	45
2.2.2 Collaborative . . . . .	47
2.2.3 Experimental . . . . .	53
2.2.4 Platform . . . . .	57
2.3 Conclusion . . . . .	65
3. TROUBLING TRADEOFFS IN SEMIOTIC INFRASTRUCTURE DESIGN	67
3.1 Knowledge Representation Oppositions . . . . .	71
3.1.1 Encoding Knowledge: Depth vs. Breadth . . . . .	74
3.1.2 Encoding Syntax: Logic vs. Procedures, Neat vs. Scruffy . . . . .	77
3.1.3 Encoding Semantics: Expressivity vs. Tractability . . . . .	83
3.1.4 Battle Grounds . . . . .	87
3.2 Knowledge Representation on the Semantic Web . . . . .	91

3.2.1	A Syntax for Web Knowledge . . . . .	92
3.2.2	A Semantics for Web Knowledge . . . . .	94
3.3	Knowledge in the "Real World" . . . . .	99
4.	PURSUING THE LIMITS OF KNOWLEDGE REPRESENTATION ON THE SEMANTIC WEB . . . . .	103
4.1	Genealogies of Logical Restrictions . . . . .	105
4.2	Shape-Shifting Knowledge . . . . .	108
4.2.1	Encoding World Wide Knowledge . . . . .	110
4.2.2	Encoding Knowledge Worldwide . . . . .	113
4.3	Delaying Semantic Commitment . . . . .	115
4.4	Learning to Swim in Troubled Waters . . . . .	120
4.4.1	Scoping Schema.org . . . . .	123
4.4.2	Modeling Schema.org . . . . .	125
4.5	Expertise in Encoding Différance . . . . .	127
4.6	Semiotic Tricksters . . . . .	130
5.	CATACHRESTICALLY DESIGNING SEMIOTIC INFRASTRUCTURES FOR THE HUMAN SERVICES . . . . .	133
5.1	Information and Referral as a "Real World" Search Engine . . . . .	136
5.2	Semiotic Heterogeneity in the Human Services . . . . .	140
5.2.1	Semiotic Regulations, Bureaucracies, and Networks . . . . .	140
5.2.2	Semiotic Styles, Habits, and Commitments . . . . .	143
5.3	"Sorting" Semantic Differences with Semiotic Infrastructure . . . . .	145
5.4	Finding a Term's Logic Niche: The Design of the AIRS/211 LA County Taxonomy . . . . .	147
5.4.1	Bracketing, Ordering, Scoping, and Re-ordering . . . . .	148
5.4.2	Discerning Authoritative Language . . . . .	150
5.4.3	Translating For Diverse Communities . . . . .	153
5.4.4	A Standard Language for the Human Services? . . . . .	154
5.5	Creating an Interlingua for the Human Services: The Design of the Human Services Data Specification . . . . .	155
5.5.1	Modeling Semantics Relationally and Contextually . . . . .	157
5.5.2	Using the Term "Standard" in Air Quotes . . . . .	159
5.6	Strategically Encoding Semiotic Infrastructure . . . . .	161

6. CONCLUSION . . . . .	167
6.1 Studying Data and Data Infrastructures Ethnographically . . . . .	170
6.2 Shifts in Expert Sign Systems . . . . .	172
6.3 The Need for New Kinds of Data Expertise . . . . .	174
6.4 Learning to Critically Represent Data . . . . .	177
6.5 Conclusion . . . . .	181
REFERENCES . . . . .	182

## LIST OF TABLES

1.1	Ideologies of semiotic infrastructure design. . . . .	28
-----	---	----

## LIST OF FIGURES

1.1	Google Knowledge Graph Results for 'IC 434'   (Google and the Google logo are registered trademarks of Google Inc., used with permission.) . . .	2
1.2	David Woods, a computer scientist involved in the design of several Semantic Web frameworks, re-purposed this classic cartoon to demonstrate one semiotic technologist ideological spectrum. (Reprinted with permission). . . . .	24
1.3	Google Results for 'Ontology'   (Google and the Google logo are registered trademarks of Google Inc., used with permission.) . . . . .	30
2.1	Annotations on an artifact page in PECE represent diverse interpretations of the artifact's cultural import. Clicking on a link to a user's annotation will display how different users responded to annotation questions. . . . .	60
2.2	This was the annotation form in an earlier version of PECE. Each field of this form was a shared question to which researchers would respond for a particular artifact. As fields of a form, it was technically challenging to disaggregate responses to this form in order to compare how the question was answered by different researchers or for different artifacts.	61
2.3	On Analytic pages in the newer version of PECE, users can compare the responses to a single annotation question - across users and across artifacts. . . . .	62
2.4	By creating a PECE Essay, users can pull together data from across the platform and configure it into "kaleidoscopic" views. . . . .	63

## ACKNOWLEDGMENT

I am grateful to many friends, colleagues, and infrastructures that supported me through the writing of this dissertation. First, I would like to thank my committee members - Kim Fortun, Mike Fortun, Dean Nieusma, and Jim Hendler - for their thoughtful advice, encouragement, provocations, and critique over several years of research and writing. As an advisor, teacher, mentor, and friend, Kim shaped my thinking in far more profound ways than will be represented in this dissertation. She consistently pushed me to figure my skills and my thinking in different grounds, yet always encouraged me to stay grounded with the things that inspired me. Mike helped me learn to creatively endure (and represent) double bind; his provocations made me a better semiotic technologist. Dean, through his comments on numerous versions of papers, helped me develop focus and deliberateness with my writing - a skill that I will carry with me throughout my career. Jim inspired my fascination with the World Wide Web when I took his Web Science course as an undergraduate. His continued support and interest in my work helped me move in diverse scholarly worlds, and his knowledge of the history of AI was incredibly helpful to writing this dissertation research.

I would also like to thank all of those I've collaborated with over several years of work on the Platform for Experimental Collaborative Ethnography (PECE). In particular, I would like to thank Eric Bigras, Brian Callahan, Brandon Costelloe-Kuehn, Dominic DiFranzo, Brad Fidler, Kim Fortun, Mike, Fortun, Renato Gomes, Aalok Khandekar, Alli Morgan, Luis Felipe Murillo, and Sharon Traweek. Many folks listed here volunteered many hours to this project, helping to shape it into a system that now supports international research projects and several university courses. I'm incredibly proud of this work and the team that made it possible.

My fabulous and brilliant peers made graduate school both enjoyable and challenging. I will always treasure my time reading, writing, and reflecting with Laura Rabinow, and I cannot thank Alli Morgan enough for lending an ear whenever I needed to decompress. Thanks also to: Leo Matteo Bachinger, David Banks, James

Birmingham, Michael Bouchey, Ben Brucato, Brian Callahan, Mitch Cieminski, Pedro de la Torre III, Thomas De Pree, Mara Dicenta Vilker, Kevin Fodness, Ellen Foster, Colin Garvey, Wynne Hedlesky, Rebecca Jablonsky, Scott Kellogg, Michael Lachney, Dan Lyles, Khadija Mitu, Lee Nelson, Karin Patzke, Hined Rafeh, N. Bucky Stanton, Guy Schaffer, Kate Tyrol, and Kirk Winans. I'd also like to give special thanks to my friend and mentor, Alison Kenner, whose encouragement kept me chugging along through even the most frustrating of times.

Thank you also to RPI's STS faculty, who, each in their own way, revolutionized my understanding of knowledge: Atsushi Akeru, Steve Breyman, Nancy Cambell, Linnda Caporael, Faye Duchin, Ron Eglash, Kim Fortun, Michael Fortun, Tamar Gordon, John Gowdy, Abby Kinchy, Jim Malazita, Michael Mascarenhas, Dean Nieuwma, Raquel Velho, Langdon Winner, and Ned Woodhouse.

The Tetherless World Constellation (TWC) at Rensselaer Polytechnic Institute has been such a welcoming space for me to engage in interdisciplinary work. Thank you to Jim Hendler, Deborah McGuinness, and Peter Fox, who introduced me to many of my interlocutors, provided technical feedback on several sections of this dissertation, and more generally supported my socio-technical perspective. I'd also like to thank Dominic DiFranzo and Kristine Gloria, who were wonderful collaborators on several papers and presentations. I so enjoyed the time thinking about and writing about the Web with you.

The Research Data Alliance (RDA) has also been a wonderful space for me to explore the intersection of data infrastructure, research, and society. I would like to thank Fran Berman for creating a space for humanities voices at the RDA (while also demonstrating how to be a rockstar female computer scientist).

I have received financial support from several grants, scholarships, and institutions that made this research possible. An RPI Seed grant supported a year of my work developing the information architecture for the Platform for Experimental Collaborative Ethnography. A two-year fellowship from the School of Humanities, Arts, and Social Sciences enabled me to focus on my fieldwork and writing in my final two years of the program. The Sloan Foundation also supported me through an RDA Data Share Fellowship, which enabled me to integrate in the RDA community

and informed some of the research presented in this dissertation. Finally, both the Tetherless World Constellation and the Institute for Data Exploration and Applications supported my travel to several conferences and workshops, where I conducted much of the fieldwork and interviewing for this dissertation.

I am grateful to each of my interlocutors for taking the time to share their histories, thoughts, and insights with me. This dissertation would not have been possible without your years of dedication and contributions to your respective fields. I am also indebted to the 42nd St. Branch of the New York Public Library, where most of this dissertation was written. There were days where working in the Rose Reading Room, which is probably the most magnificent room I've ever encountered, was the only thing that made the loneliness of dissertation writing tolerable. I'm incredibly appreciative of New York City's investment in maintaining this beautiful public infrastructure.

My mom, Kim Woodward, has been scaring me out of quitting things since I was a little girl. She helped me develop stamina, fostered my creativity, and honed my ability to not stop writing until it's done and perfect; she's made me into a better scholar. My dad, Peter Poirier, has taught me patience and how to communicate with care; he's made me into a better ethnographer. My sister, Danielle Poirier, has been my emotional other half at every stage in my life; she's made me into a more caring teacher and activist.

I'm also so grateful for the love and support of Louise Sawyer, who continuously reminds me that I can be forever young, and Keith Noftle, whose conversations consistently encourage me to think deeper. My grandmother, Gertrude Bouvier, has been a wonderful cheerleader at every stage of my life. To my dear friends Jenna Bissonnette, Ashley Devine, Elena Krupin, Stephen Nock, thank you for encouraging me through the challenging times and celebrating with me through the great ones. To Sharon O'Brien and Trisha Miklic, thank you for providing with a soul-fulfilling distraction from academic life. Thank you to all of the lovely new family members that have come into my life since the start of graduate school - Paul Larson, Megan Marshall, Attie Poirier, Robert Poirier, Mary Wright, and Stu Wright. Thank you to my dog, Madison, whose spunk kept me sane through even

the most laborious days. And to Zach Wright, my partner and my love, thank you for supporting every exciting, treacherous stage of this process - from the worst of it to the best of it. I feel so incredibly fortunate to have you beside me.

## ABSTRACT

This dissertation examines the history, culture, and expertise of data infrastructure design work. More specifically it narrates how a community of researchers and practitioners that I refer to as *semiotic technologists* design data infrastructures to encode and represent the meaning of data. I demonstrate how semiotic technologists' ideas about language (about how meaning takes shape and evolves) animate how they approach the design of these data infrastructures, impacting how the infrastructures eventually order and represent data. Therefore, I refer to the infrastructures I study in the dissertation as *semiotic infrastructures*.

The research presented in this dissertation is based on four years of ethnographic fieldwork with diverse data communities – including the Semantic Web community and the human services informatics community. It is also based, in part, on reflections of my own involvement in a project to design a digital humanities platform - the Platform for Experimental Collaborative Ethnography (PECE). Through interviewing, archival work, and experimental design projects, I aimed to unpack how semiotic technologists in these communities learned to endure the limits of knowledge representation.

I argue that, while semiotic technologists bring particular language ideologies to their design work, they are often faced with competing injunctions – for instance, to make infrastructures more flexible for characterizing different interpretations of data's meaning, or to make them more structured so that diverse communities sharing data can use them to align their language. I follow semiotic technologists as they approach these tradeoffs, examining how their ideas about language and meaning shift as they learn to work in domains where the meaning of data is messy or “scruffy.” I show how they learn to assess the challenges of representing meaning in diverse data communities and how they design strategically and experimentally to address these challenges. In doing so, I argue that what it means to be an expert in encoding meaning is constantly evolving – stabilizing in particular times and contexts.

Focused at the scale of meaning-making, the dissertation contributes to litera-

ture (situated at the intersection of Science and Technology Studies and Information Studies) theorizing the history and politics of information infrastructures. It also furthers understanding of social practice and politics in big data contexts, demonstrating how data infrastructure designers learn to bring different assumptions, logics, and politics to their work. Finally, the dissertation recommends pedagogy for training the next generation of data and information scientists to recognize, communicate, and creatively endure the limits of representing knowledge in diverse data domains.

# 1. INTRODUCTION

Polysemy: "The fact of having several meanings; the possession of multiple meanings, senses, or connotations." - Oxford English Dictionary

Différance: "In the philosophy of Jacques Derrida: the impossibility of any sign within a system of signs having a fixed meaning; the process by which meaning is endlessly deferred from one sign to another within such a system." - Oxford English Dictionary

Polysemy always puts out its multiplicities and variation with the horizon, at least, of some integral reading which contains no absolute right, no senseless deviation – the *horizon* of the final parousia of a meaning at last deciphered, revealed, made present in the rich collection of its determinations. Whatever interest one might find in them, whatever dignity one might grant them, plurivocity, the interpretation it calls for, and the history that is precipitated out around it remain *lived* as the enriching, temporary detours of some passion, some signifying martyrdom that testifies to a truth past or a truth to come, to a meaning whose presence is announced by enigma. All the moments of polysemy are, as the word implies, moments of meaning. (Derrida, 1983, 350).

## 1.1 Introduction

If you navigated to Google in July 2017 and queried the engine with the phrase IC 434, you'd have noticed a box appear on the right hand side of the search results displaying images of an emission nebula in the constellation Orion. You'd also have noticed a Wikipedia snippet, describing IC 434, and beneath that a list of characteristics of the nebula, including its magnitude and distance.

These results display as part of Google's knowledge graph. The knowledge graph, introduced by Google in 2012, uses what is called *semantic search* to produce the results. In semantic search, search engines gather data, from sites like Wikipedia, the CIA World Factbook, and the Mayo Clinic that has been structured in such a way that a machine can interpret its meaning. Take, for instance, IC 434's distance



**Figure 1.1: Google Knowledge Graph Results for 'IC 434' | (Google and the Google logo are registered trademarks of Google Inc., used with permission.)**

from Earth, which is listed in the knowledge graph as "1500 ly" (or light years). This data was likely gathered from DBpedia – a project that extracts information from Wikipedia pages and then structures the information in ways that computers will eventually be able to "read."

The property "distance in light years" is defined in DBpedia as "dist ly," a property that can be used to describe that the integer "1500" on IC 434's Wikipedia page refers to its distance in light years (see above image). This same property can be used to describe the distance in light years of stars and nebulas that are the subject of other Wikipedia pages. For instance, the "dist ly" property also describes that the

integer 1600 on the Wikipedia page for "Messier 43" refers to the nebula's distance in light years, and that the integer 643 on the Wikipedia page for "Betelgeuse" refers to the star's distance in light years.

With the property "dist ly" defined in DBpedia and all the Wikipedia pages containing an integer that refers to the topic of that page's "distance in light years" structured with the dist ly property in DBpedia, I can run powerful search queries. I can request a list of all of the stars and nebulas that are less than 1500 light years away for instance. Or I can build a mobile app that allows me to compare the distance between stars that a person can see on a given night. And that's just one property. As of July 2017, DBpedia describes 4.58 million things in 125 languages. It contains 3 billion pieces of information, extracted from Wikipedia pages spanning topics relating to persons, places, creative works, organizations, species, diseases, and more (DBpedia, 2017). Prior to DBpedia, I could only search for a topic in Wikipedia and get linked to the topic's page. Search engines knew nothing about the information on that page or how it compared with information on other Wikipedia pages. With DBpedia, there is now a massive knowledge base of information that has been structured with properties that any user or app can query to integrate, compare, or pull apart information *within pages* from all over Wikipedia.

Google's knowledge graph extracts information from knowledge bases like DBpedia and stores them in a *graph database*. In graph databases, each data point is stored as a *node* that has properties, and is linked together with *edges* that describe relationships between data points. So, in a graph database IC 434 would be a node, with the property "dist ly." An edge would link the node IC 434 with the node for the constellation Orion. The edge would describe that IC 434 "is a part of" Orion. Thus, using graph databases, Google's knowledge graph links data collected from all over the Web and ascribes to their properties and relationships, so that it can later interpret that meaning to improve search.

In a Google blog post introducing the knowledge graph, Amit Singhal (2012) described the value of the graph as offering the capacity to search for "things, not strings." He noted that, "For four decades, search has essentially been about matching keywords to queries." However, with the knowledge graph, Google's search

engines could use semantic search to look up the properties and edges for a certain search string. So when I search with the string "IC 434," the search engine understands that I could be searching for a nebula, or alternatively I could be searching for the musical group IC 434. With polysemy rampant across the Web, Google displays a box that allows you to disambiguate between these alternative meanings for the string; clicking the relevant link filters the search results. Singhal noted, "This is one way the Knowledge Graph makes Google Search more intelligent—your results are more relevant because we understand these entities, and the nuances in their meaning, the way you do."

In July 2017, I can search Google for Chlorpyrifos - the chemical name for a pesticide used to kill insects and worms - and get information from the knowledge graph about its chemical make-up, mass, and melting point. I can search Carignan - a French/Spanish grape variety used to make dry red wine - and get information from the knowledge graph about the regions in which it grows, the sweetness of the wine it produces, and its scientific name.

But then I type in the search string "homelessness" - a term that has a strict scope in the human services sector to determine whether a person is eligible for supportive housing, special Medicaid services, and education services. Homelessness is also a term defined differently in the U.S. federal acts that regulate the distribution of these services and thus is in need of disambiguation when determining if a person is eligible for each service.<sup>1</sup> Google does not return any knowledge graph results. I type in abortion - a term that certainly needs more authoritative knowledge sources framing it. Again, Google does not return any knowledge graph results. Nor does it produce any results for "Medicaid," "WIC," or the "Supplemental Nutrition Assistance Program."<sup>2</sup>

One concern that motivates this dissertation is how knowledge bases end up with such rich structured information about astronomy, chemistry, and the study of

---

<sup>1</sup>The U.S. Department of Health and Human Services defines homelessness differently than the U.S. Department of Housing and Urban Development (HUD). HUD's definition is notably more restrictive.

<sup>2</sup>Fortun (2014) performs a similar exercise in her article "From Latour to Late Industrialism," querying Latour's knowledge base *An Inquiry into the Modes of Existence* and examining how the language ideologies interlaced in its design shape what knowledge comes to the fore. I will return to this exercise later in the chapter.

wine varietals, and such little structured information about social services - information that could add a great deal of civic value to Google's search. I am certainly not the first to raise this concern. Scholarship in STS has aimed to formalize a field of information infrastructure studies - assessing the socialities that digital infrastructures enable, the values that cohere within them, and the way knowledge work changes as a result (Bowker et al., 2009). In this vein, sociologists of information infrastructure have shown how what becomes knowable in an information system is in part guided by the way that a select group of experts decides how information will be named, structured, and processed (Halford et al., 2013). Sociologists have also shown how assumptions about the nature of knowledge - for instance, that it can be objectively credible - guides the design of information systems, in turn limiting what information can be extracted from them (Waller, 2016). Studies and commentary on contemporary information infrastructure - on classifications (Bowker and Star, 1999), ontologies (Ribes and Bowker, 2009; Leonelli, 2010), databases (Bowker, 2000), algorithms (Gillespie, 2014), and platforms (Bogost and Montfort, 2009; Gillespie, 2010) - have shown how political biases, corporate interests, and designer values and assumptions implicate how information systems order information and produce meaning.

In this dissertation, I approach these concerns with a different framework - by characterizing the role of *language* in shaping information infrastructures and the knowledge they structure and disseminate. I do so not only by examining information infrastructures that explicitly aim to encode language - what I call *semiotic infrastructures* - but also by teasing apart the diverse language ideologies that the designers of semiotic infrastructures - experts that I refer to as *semiotic technologists* - bring to their work.<sup>3</sup> I argue that *there are limits to encoding difference with semiotic infrastructures* (or in other words, that there are limits to representing all of the things that a certain thing is not). I argue that these limits are even greater

---

<sup>3</sup>Notably, researchers in critical code studies and software studies are similarly interested in examining the semiotics of information infrastructures but tend to do so by looking at "code as a text, as a sign system with its own rhetoric" (Marino, 2006). Methodologically, this often involves a "close reading" of digital code (Montfort et al., 2012). As an ethnographer, I instead approach the study of digital infrastructures and the ways in which they make meaning through ethnographic analysis of the language ideologies that infrastructure designers bring to their work.

in knowledge domains where language tends to be more heterogeneous, messy, and politicized - knowledge domains like cultural anthropology (Chapter 2), the World Wide Web (Chapters 3 and 4), and the social services (Chapter 5). Language iterates and drifts; words at once hold multiple meanings and have quasi-infinite possibilities for taking on new meaning; messages that operate at different scales can produce paradox, and differences can always be further multiplied. To represent knowledge to a computer - to "enact" representation - semiotic technologists have had to learn to employ new critical modalities - building out infrastructure to add *meaning* to data in domains where meaning itself is impossible to pin down. The dissertation is thus about expertise - a type of expertise that must learn to think, work, and represent knowledge in the face of limits.

Notably, the arguments I make in the dissertation extend to ethnographers, who too act as semiotic technologists, learning to represent knowledge (to "write culture" (Clifford and Marcus, 1986)) in domains where meaning is impossible to pin down. In my own ethnographic writing, I will try to make explicit the moments where I face limits to encoding difference - where polysemy emerges yet a summation of meanings are never enough to precisely define a term. This, of course, poses limits to precisely defining a semiotic technologist and encoding their expertise, but these are conditions under which semiotic technologists must learn to work and communicate. This dissertation is, then, in many ways an ethnography of that learning process - an ethnography of changing language practices and language ideologies that I hope may further the work of those learning communities.

## 1.2 Defining Semiotic Infrastructures and Semiotic Technologists

I refer to the information infrastructure that I study in this dissertation as *semiotic infrastructures* - tools, practices, and relations that are designed in order to communicate, filter, and translate meaning amongst and through signs. I examine ethnographically a community of practitioners that I refer to as *semiotic technologists* - a community of practitioners that support endeavors to make data interpretable, discoverable, and translatable by designing semiotic infrastructure

that models interpretations of how language works. In other words, semiotic technologists aim to enact the language of a particular domain through the design of semiotic infrastructure; they are experts in *knowledge representation*.<sup>4</sup> Their work and expertise involves delineating which "differences make a difference" for a given domain.<sup>5</sup>

Very careful definitional work has been employed in the study of infrastructure since Star (1999) issued a call for an "ethnography of infrastructure." For Star, this was a call to study "boring things" - the wires, protocols, and technical specifications that serve as foundations for systems that hold and transport products, people, and ideas. Yet, her call requested more than an analysis of the components of large-scale technical systems. Star and Ruhleder (1996) defined infrastructures as fundamentally relational; rather than "ready-to-hand" substrates, they argued that infrastructures represent a series of relationships that emerge as a result of organized practice. The hypertext protocol only becomes a component of an information infrastructure when distributed users leverage it to link information; pipes only become a component of water transport infrastructures when distributed users send water through them. In defining infrastructures, Star and Ruhleder suggest that information infrastructures only become visible upon breaking down.

Citing Star and Ruhleder's definition of information infrastructure has become almost a rite of passage for anyone contributing to an ethnography of infrastructure. By their own definition, this citation for information infrastructure has become its own information infrastructure. The definition serves as a specification - a foundation for communication in STS, anthropology, and information studies, transporting ideas from one scholar to the next. It is a shared conceptualization of meaning that emerges as a result of semantics, publications, and organized practice. It is also a glaringly visible semiotic infrastructure; the citation is acknowledged in thousands of publications, which reference it to signal that they are building from this founda-

---

<sup>4</sup>In later chapters, I will examine the field and history of "knowledge representation" in artificial intelligence communities in detail. However, here, I intend for the term to more broadly signify communities that designing the infrastructures for representing meaning.

<sup>5</sup>Semiotic technologists and semiotic infrastructures may be an object of study in what Collins (1992, 3) calls "knowledge science," which examines "how knowledge is made, maintained, transformed, and transferred" by studying "what communities know and the ways in which the know it."

tional work. This also makes its own (inevitable) breakdown glaringly visible, which Brian Larkin (2013, 336) describes eloquently in a review article on the poetics and politics of infrastructure:

Thus many studies that begin by stating how infrastructures are invisible until they break down are fundamentally inaccurate. Infrastructures are metapragmatic objects, signs of themselves deployed in particular circulatory regimes to establish sets of effects. It is commonplace, seemingly obligatory, for almost any study of infrastructure to repeat Star's (1999) assertion that infrastructures are "by definition invisible," taken for granted, and that they only "become visible on breakdown" (p. 380; see also Collier 2011, Elyachar 2010, Graham and Marvin 2001, Larkin 2008). But this assertion is a partial truth and, as a way of describing infrastructure as a whole, flatly untenable. Invisibility is certainly one aspect of infrastructure, but it is only one and at the extreme edge of a range of visibilities that move from unseen to grand spectacles and everything in between.

Larkin (2013, 330) goes on to offer an alternative way for thinking about the act of defining infrastructure:

Infrastructures are not, in any positivist sense, simply "out there." The act of defining an infrastructure is a categorizing moment. Taken thoughtfully, it comprises a cultural analytic that highlights the epistemological and political commitments involved in selecting what one sees as infrastructural (and thus causal) and what one leaves out.

In describing the defining of infrastructure as a "categorizing moment," Larkin points out how ethnographers of infrastructure themselves act as (what I'm categorizing as) "semiotic technologists," defining which differences make a difference in their own domain and for their own purposes. He also argues that "infrastructure" is an unstable concept - one that is open to semantic drift as it is employed in new contexts and for new purposes.

I define the infrastructures that I refer to throughout the dissertation as "semiotic" in part to highlight the semiotic aspects of building systems that transport and translate ideas. For instance, consider RDF - the Resource Description Framework. In technical terms, RDF would be considered a standard data format, specifying how data should be structured in order for the data to be interchangeable between different systems and capable of being merged with other data in other systems (even if those systems are structured differently). RDF was designed in order to make it possible to build expressions from data on the World Wide Web and to have these expressions translatable in diverse systems. To do so, RDF specifies a particular "syntax" for organizing data into "triples" - with a subject, predicate, and object. One piece of data on the Web would be the subject; the object would be another piece of data on the Web or a reference to something that exists off the Web, and the predicate would rely on Web schemas and ontologies (other components of the semiotic infrastructure) to describe the relationship between the two data points. As I will show throughout the dissertation, the design of infrastructure like RDF is semiotic not only because it digitally molds language, but also because it is described and ordered by ideas like these about how language works, and, as part of that work, produces and structures meaning.

In defining infrastructures as semiotic, I also intend to mark that the "semiotic infrastructure" that has supported my own thinking and work consists of ideas and practices developed in the field of "critical semiotics" by thinkers like Jacques Derrida and Gayatri Spivak - thinkers that position meaning and signification as perpetually deferred and thus only achieving stability in certain, fleeting times and contexts. Because meaning can never be fully exhaustive or enclosed, for critical semioticians all naming, defining, and categorizing can be categorized as "catachre-

sis" - or the (often forced) misuse of a signifier to represent a particular context.<sup>6</sup> Importantly, in order to advance politics - in order to make a case for something - naming, defining, and categorizing is both absolutely necessary and absolutely violent - a double bind in Gregory Bateson (1972)'s sense.<sup>7</sup> To be ethical and attentive to the politics of representing, cultural analysts must be both deliberate and reflexive in selecting which differences make a difference in our domains.

Thus, recognizing naming to be catachrestic becomes particularly important in my defining of semiotic technologists - a group of practitioners marked just as much by their differences as their similarities. Notably, the group of people I characterize as semiotic technologists do not necessarily share a disciplinary background, produce a common technology, or employ a shared set of practices. Semiotic technologists are trained in diverse fields of scholarship and practice - computer science, analytical philosophy, linguistics, database engineering, as well as more specific domains where informatics get applied, and they produce diverse semiotic infrastructures - taxonomies, schemas, data exchange languages, data models, ontologies, and recommendations for "best practice," for instance. They also often constitute diverse "thought collectives," bringing to their work different perceptual or cognitive dispositions of how the world is organized and how language works (Fleck, 1981). In categorizing "semiotic technologists," I am aiming to highlight the semiotic aspects of infrastructural design and work. I define semiotic technologists to mark a group of experts that share a commitment to grappling with the binds that emerge in designing infrastructure to make signs (that are always already incomplete and unstable) interpretable, discoverable, and translatable in diverse settings. Digital data infrastructure designers and ethnographers of infrastructure alike contend with such

---

<sup>6</sup>Derrida (1982b, 255-256) writes that "catachresis":

concerns first the violent and forced, abusive inscription of a sign, the imposition of a sign upon a meaning which did not yet have its own proper sign in language. So much so that there is no substitution here, no transport of proper signs, but rather the irruptive extension of a sign proper to an idea, a meaning, deprived of their signifier.

I put catachresis in quotes here to mark that defining catachresis itself is catachresis, and that this is the double bind of a deconstructive view of language (Spivak, 1993).

<sup>7</sup>I will elaborate on double binds later in the chapter, but here I will note that a double bind refers to the communicative dissonance that emerges when an individual is exposed to two conflicting demands.

binds; the differences (that make a difference) between them are primarily ones of media and use-context.

### 1.3 Indexing Semiotic Infrastructure Design

In semiotics, indexicals refer to words that take on new meaning as they move from context to context. For instance, words like "here," "now," and "you" refer to different places, times, and people respectively when used in different contexts. In linguistic anthropology, the study of indexicality considers how words take on particular meaning in certain social contexts. In this section I describe the social contexts in which *semiotic infrastructures* and *semiotic technologists* become meaningful reference points. As I will show, in the social context where semiotic infrastructure design has become a mode of expertise, the indexicality of digital data (or, in other words, the way that data takes on new meaning in new contexts) is perhaps the primary challenge to which semiotic technologists respond.

In the 1980s, CERN was not only the most prominent particle physics laboratory in the world; it was also at the cutting edge of coordinating collaborative international scientific research. Herwig Schopper (2014), director-general for CERN, 1981-1988, describes the time as provoking the birth of "a new 'sociology' for international scientific collaboration;" with over 30 countries participating in experiments, the challenges for keeping track of researchers, workflows, and scientific data were enormous.

Tim Berners-Lee was hired to CERN as a contract programmer in 1980. To help him keep track of the people and projects he encountered, in his spare time, he toyed with designing a program he called Enquire - an early model for what would eventually become the World Wide Web (WWW). Enquire was designed in line with emerging hypertext models (although Berners-Lee says that he was unaware of this research in the earliest stages of his thinking). Rather than organizing data in tree-like structures (such as folder structures), users could organize their data by creating links between documents stored in separate locations; this made it possible to jump from one document to another by simply clicking on a link. In this sense, all documents could be stored according to their relationship with other documents

rather than in a fixed hierarchical structure.

After leaving CERN for a short stint, Berners-Lee returned as a fellow in the data acquisition and control division in 1983 at a time when CERN was upgrading its computing infrastructure in order to better network globally distributed researchers. However, implementing infrastructure for data sharing was a particularly difficult obstacle. Researchers in different laboratories each followed their own research methods, used their own operating systems, and often spoke different languages. In describing the systems that were proposed for addressing these challenges, Berners-Lee (1999, 15) writes:

I had seen numerous developers arrive at CERN to tout systems that "helped" people organize information. They'd say, "To use this system all you have to do is divide all your documents into four categories" or "You just have to save your data as a WordWonderful document" or whatever. I saw one protagonist after the next shot down in flames by indignant researchers because the developers were forcing them to reorganize their work to fit the system. I would have to create a system with common rules that would be acceptable to everyone. This meant as close as possible to no rules at all.

The idea for the World Wide Web, documented in a proposal Berners-Lee drafted in 1989, responded to two challenges in coordinating international research. First, leveraging the Internet (which existed decades before Berners-Lee began planning the concept of the Web) to network computers over long distances, researchers could link to and access data stored in another lab, another country, and perhaps even on a different operating system.<sup>8</sup> Second, leveraging hypertext, dispersed and diverse researchers no longer needed to all agree upon a shared and static information model. With hypertext, a data point need only be stored in one place and

---

<sup>8</sup>In the early 1960s, Internet infrastructure was notably designed with a distributed network architecture, a design approach largely oriented by Cold War politics. With a distributed network architecture, each computer in the network is connected to multiple other computers in the network. Particularly concerned that a successful enemy attack could take down an entire communication system (as would be the case with a centralized or de-centralized network architecture), designing the Internet with a distributed network architecture ensured that if one line of communication came down, the system could still operate (Abbate, 2000).

then researchers could link to it however they deemed appropriate. They could thus establish their own organization styles rather than conforming to a standardized model.<sup>9</sup> Both base infrastructures were selected in recognition that the challenge of getting everyone to conform to the same networks, systems, and schemas was insurmountable.

The invention and spread of the World Wide Web has enabled networking both information and people in profound ways. For Manuel Castells (2009) this marks a new social era - the "network society" - where global societies, markets, politics, and science are organized around digitally networked information. Yet, the challenges of networking global information are enormous. Different cultures, institutions, and disciplines often follow different methods for producing, structuring, and describing information. In the network society, it is not only the volume of data that is unfathomably vast; data is also incredibly heterogeneous and often contradictory, and thus when exchanged, data produced in one place can be difficult to interpret and merge with data produced in another place. This issue has been discussed in technical literature as a problem of data integration - the "problem of combining data residing at different sources, and providing the user with a unified view of these data" (Lenzerini, 2002). While data integration is enabled by the creation and adoption of standards for identifying, formatting, describing, and linking disparate data sets, recent literature in Information Studies (IS) and Science and Technology Studies (STS) has shown that building standards for data integration is also complicated. As practitioners aim to make data shareable, different understandings of how data should be named, stored, and linked produce "friction" (Edwards et al., 2011).<sup>10</sup>

This is perhaps why the World Wide Web has been so successful. There are

---

<sup>9</sup>Hypertext was first conceptualized in Vannevar Bush (1945).

<sup>10</sup>The complications posed by data friction are described in a growing literature on data sharing practices in the sciences. Recounting the development of databases for cross-species comparison, Sabina Leonelli (2012) describes problems that database curators face as they attempt to integrate biological and clinical data. The conflicts that emerge as they work to determine what counts as reliable evidence and to select standards and metadata illustrate epistemic disagreement and controversies amongst disciplines. Christine Borgman (2012) argues that, while there are significant overarching rationales for sharing research data – such as enabling the verification of research results and disseminating academic work to the public – different research communities experience different motivations and incentives for making research data readily available, as well as different challenges. Thus, defining common sharing protocols is tricky.

few protocols that content authors need to follow to have their data woven into the Web: 1) format a document with code that a browser can render (HTML), 2) give the document a unique Web address, and 3) link to other unique addresses using the hypertext protocol (HTTP). However, there are no rules dictating how the content itself needs to be named, organized, displayed, or interpreted. Content authors can organize the data on their pages however they choose; they then create and evolve the Web's organization by linking to content stored elsewhere from within their own pages. They do not need to agree, for instance, that all of the stars and nebulas on the Web will be organized according to their constellations or according to their distance from Earth in light years. They can organize a Web document about IC 434 according to whichever schema they choose without concern that it will prevent them from linking to a Web page that organizes the information in a different way. This, in turn, means that they don't have to all agree that IC 434 is 1500 light-years away from Earth.

Teresa Pardo, Director of the Center for Technology in Government, described, during an October 2016 interview with me, that tools like this - tools that made it possible to leap over the standardizing phase - were some of the most transformative technologies for data sharing:

So there's the tools like NIEM [National Information Exchange Model] - the data standards and the structures ... - that increase the likelihood that the data in place A and place B and place C are structured in similar ways. ... [These tools] created the platform and the set of expectations about being able to pull data up and share it and use it in interesting ways. And then [there's] the old evolution of technology that says, "well, people don't want to do that." Now we have new kinds of technological tools that allow us to not have to invest all of our time and energy and effort creating common data infrastructures. So we can leapfrog to a phase of the development of the access and use of data that is not constrained by the lack of integrated similar core data infrastructure. ... So the tools that are at different generations and different phases of development (the tools now that are just creative javascripting) - [these

tools] are making it possible to be *data structure agnostic*.

Yet, while data structure agnosticism is saving time and effort and making it possible to share data without having to get everyone on the same page, there are considerable tradeoffs. When there is no common data structure or common data semantics, it becomes much more difficult to find information, and particularly to *find* the right information. For instance, if all of the information about stars and nebulas on the Web were organized according to their constellations, it would be much easier for Google's search engine to track down Web pages that contain information about IC 434; it would also be much easier for Google's search engine to know that the page's mention of IC 434 was referring to the nebula and not to the music group IC 434. This is the tradeoff that Google's knowledge graph is responding to - why Google wants to *understand* the information on a Web page. It is also the tradeoff to which community of practitioners working to enable a *Semantic Web* is responding. With so much heterogeneous information on the Web and no common data schema or data structure, how could semiotic technologists direct searchers to the information that they are searching for? How do semiotic technologists (and computers) make sense of the data? How do they structure it and describe it so that they - and we - can bring data together in innovative ways? How can they make data on the Web *meaningful* - so that it can be discovered, shared, and integrated? How do they *index* (in both the semiotic sense and the computing sense<sup>11</sup>) data so that it can be understood in diverse contexts?

The challenge of making heterogeneous data meaningful, findable, shareable, and integrable is partly a challenge of *translation* - translation of data's import from one social context to the next or one architecture to the next. When designing infrastructure to index data - to make it translatable into diverse contexts - semiotic technologists have to encode assumptions about how language works. Yet, the sta-

---

<sup>11</sup>In July 2017, the Oxford English dictionary lists 26 definitions for "index," used in a variety of contexts, including anatomy, computing, music, printing, algebra, optics, craniometry, crystallography, dynamics, and economics. Relevant (but not necessarily precise) definitions in this social context include: 1. "The fore-finger: so called because used in pointing." 4. a. "That which serves to direct or point to a particular fact or conclusion; a guiding principle." 5. b. "An alphabetical list, placed (usually) at the end of a book, of the names, subjects, etc. occurring in it, with indication of the places in which they occur." 5. d. "Computing. A set of items each of which specifies one of the records of a file and contains information about its address."

bility of meaning is always temporary and contextual; the meaning that gets applied to data inevitably iterates as it moves into new contexts - as data is exchanged and integrated.<sup>12</sup> Semiotic infrastructures and semiotic technologists act as translators, responding to the challenges that indexicality poses.

## 1.4 Encoding Digital Semiotic Expertise

In the academic field of semiotics, the creation of transportable meaning is referred to as "encoding" (Hall, 1980). In semiotics, "codes" refer to the social conventions that enable a community to interpret that a signifier (the word "tree") refers to a signified object (a "tree"). Encoding involves creating meanings that can be "decoded" in new contexts by referencing relevant codes.

During an April 2015 interview with my research group,<sup>13</sup> Deb McGuinness, a computer scientist, leader in the Semantic Web community, and wine aficionado, described what it means to her to represent knowledge to a computer (and in this particular case, knowledge about wine):

Well you've got to represent types of wine and food. So if you're eating salmon, then typically people will have a white wine with the dish. So I want to *encode* that kind of information, *encode* the information about the kind of wines there are and what colors of wines there are, and what goes with what. (emphasis mine)

For McGuinness, the "encoding" that she refers to involves bringing concepts together with a set of primitives defined in a knowledge representation language.

---

<sup>12</sup>Translation, as critical theorists Talal Asad (1986) and Gayatri Spivak (1993) argue, is never a neutral or balanced process. For Asad, because of an "inequality of languages," in translating "weaker" languages, translators tend to "push meaning ...in a single direction," and the power dynamics involved in translating tend to be masked (163). Further, for Spivak (1993), the rhetoric at play in the shadows of any statement is often eclipsed by the logic of translation.

<sup>13</sup>While my undergraduate degree in Information Technology and Web Science had trained me in articulating the aims and methodologies of the Semantic Web, I had not planned on making the Semantic Web a subject of my graduate research until well into my fourth semester of a PhD program in Science and Technology Studies. I had been working for several years on PECE, and in January 2015 we were awarded a grant to "semantify" the digital platform in collaboration with a knowledge representation expert, Deborah McGuinness, also at Rensselaer Polytechnic Institute. My entire research group interviewed Deb in order to better understand the assumptions and commitments that she brought to her work.

One primitive may be "is a;" another primitive may be "is a kind of." So, using the knowledge representation language, she can encode that salmon "is a kind of" food or that a fork "is a kind of" utensil. From this, the computer can build definitions of and represent relations between data. McGuinness described her expertise as being able to find the "edges of a term." Georgia Sales, who created an extensive taxonomy of social service terms, describes her expertise as finding a term's "logical niche." Derek Coursen, an information specialist in the human services and public health sector, described during an interview with me that:

... very experienced data modelers get to the point where they are able to not only reflect the semantics of a domain but they are actually able to work with the stakeholders of a domain that has very fuzzy semantics and *help them to untangle what's going on.*

For these semiotic technologists, encoding (in the semiotic sense) meaning is about disambiguation - about finding a set of boundaries between words and things, and then encoding (in the computing sense) those boundaries to a computer.

Yet other semiotic technologists that I have interviewed have suggested that to precisely define one word to a computer, you have to define the entire universe. For instance to define a chair, you have to define all of its properties and materials, and to define those properties and materials, you have to define their properties and materials, and so on. For these semiotic technologists, this is not "pragmatic." The term "pragmatic" also has multiple meanings in different contexts, and here, many semiotic technologists refer to it with double meaning - in one sense referring to technical feasibility, and in another sense referring to "pragmatics" in Peircean semiotics - or the theory's consideration of *indexicality (or context)* in the representation of meaning. For these semiotic technologists, encoding (in the semiotic sense) meaning by disambiguating terms is not only technically unfeasible; it also assumes that the context of an encoding does not matter - that the "edges of a term" are the same in every context.

In the communities that I have studied, it is widely recognized that there are divisions in thinking about how best to approach representing knowledge to a computer. Today my informants still cite classic distinctions between the *neats* and the

*scruffies* in the tradition of artificial intelligence referred to as "knowledge representation" - a tradition concerned with enabling a computer to model common sense facts about the world. Neats aimed to model knowledge cleanly, consistently, and completely, whereas scruffies were willing to sacrifice cleanliness, consistency, and completeness in order to get a system to work (subject of Chapter 3). Today, semiotic technologists use similar binary distinctions to compare and contrast approaches to, or styles of, knowledge representation; clean or messy, formal or loosey-goosey, for example.

Ethnographers studying knowledge representation communities and encoding their expertise have tended to characterize the neater spectrum of semiotic technologists, emphasizing how knowledge engineers "formalize" knowledge (Forsythe, 1993; Star, 1995; Adam, 1998; Ribes and Bowker, 2009). For example, in 1993, Diana Forsythe described the practice of knowledge engineering through an ethnographic study of an AI lab. Borrowing from Susan Leigh Star the idea that computer scientists tend to "delete the social" (cited as personal communication), Forsythe (1993, 53) argued that knowledge engineers tend to also "delete the cultural." In trying to delineate globally applicable rules for knowledge that could be codified into expert systems, knowledge engineers tended to employ "I am the world thinking" - taking their own perspective as representative of all reasoning. More recent ethnographic accounts have described an expertise in knowledge modeling that is divorced from domain expertise. For instance, Ribes and Bowker (2009) describe how knowledge engineers working in the geoscience community developed an "acquisition routine," or a "recipe" outlining the steps for articulating domain knowledge in a codified language. These experts moved from domain to domain following the same set of steps for establishing relations between concepts in traditionally silo-ed communities and applying formulas for resolving epistemic disagreements amongst domain experts.<sup>14</sup>

Examining communities where stabilizing the "edges of terms" is important for advancing global, interdisciplinary research agendas (like climate science, oceanog-

---

<sup>14</sup>Harry Collins (2004) has characterized the ability to communicate expert knowledge without actually practicing that knowledge as "interactional expertise." Ribes and Bowker (2009) describe ontologists as able to achieve an interactional expertise of the geoscience domain through interactions with the language of that domain.

raphy, and genetics), these ethnographic studies have shown how data sharing and knowledge production advance despite conflicting worldviews amongst the stakeholders involved. This literature has characterized how information infrastructure designers produce "neat" formalisms,<sup>15</sup> where meaning stabilizes (even if only temporarily) - or in other words, how such designers have produced *functional* semiotic infrastructures. It has shown that the processes involved in producing such formalisms are often invisible (Star, 1995), how incommensurate worldviews, habits, and methods impede consensus (producing "data friction") (Edwards, 2010), how communities work to find "common ground" (Edwards et al., 2011), and how certain ways of being in and knowing the world get excluded as meaning stabilizes (Bowker and Star, 1999).

In this literature, producing neat formalisms does not necessarily require consensus. Star (2010) describes many standards, classifications, and other information infrastructures as "boundary objects" - spaces, relations, and/or material objects that emerge in interdisciplinary communities when they cannot reach consensus. Boundary objects are designed to be flexible enough to be interpreted differently by diverse communities but also stable enough to maintain their integrity across such communities. Boundary objects enable science to function in the face of disagreement; they provide the stability of a boundary around which disciplinary communities can cooperate.<sup>16</sup>

The way that knowledge gets encoded (in both the semiotic and technical

---

<sup>15</sup>Susan Leigh Star (1995, 94) has argued that "formal representations" render the political decisions that go into their creation invisible. She writes:

The world does not come in neat layers visible to the naïve eye. Geological strata simply look like mush when a probe is inserted into the earth to pull them out (Bowker 1988); the neat formal representations of layers one sees in geology and paleontology texts are the results of many decades of abstraction and negotiation in the scientific community. ... And in the chip design team, what in the end is an extremely neat, fast bit of silicon and metal engraved in the logical, formal reasoning of a worldwide community of scientists and technicians, appears in the design process to be a battle for terrain and particular viewpoints (Star 1988).

<sup>16</sup>Several other concepts have emerged in STS to describe the way that science functions without consensus. 'Immutable mobiles' (Latour, 1986) refer to objects whose meaning (while flexible) temporarily stabilizes so that it can travel between communities. The concept of 'trading zones' describes a situation where scientists that may speak different "disciplinary languages" agree on a set of rules for communication exchange (Galison, 1997).

sense) has important implications for shaping what can be known in information systems. In their article "Between meaning and machine: learning to represent the knowledge of communities," Ribes and Bowker (2009, 215) argue that "Ontologies have their own epistemology: what and how the computer can 'know' is very particular, limited by the availability [of] description logics and the extant level of formalization." Ethnographers too are in a constant process of learning to represent the knowledge of the communities they study. Ethnographers' ontologies also have their own epistemologies; the way communities get represented is also very particular, limited by the logics brought to ethnographic work and the ways in which knowledge is formalized into writing. Notably, constructivist epistemologies like actor-network theory and grounded theory - traditions that privilege the study of how ideas and networks *stabilize*<sup>17</sup> - have oriented many existing ethnographies of digital semiotic expertise. Many of these ethnographies also originated in or work in the spirit of academic scholarship examining Computer Supported Cooperative Work (CSCW) and thus approach the research from a particular vantage point: considering how computers enable people to work cooperatively or alternatively, what disables them from doing so. Guided by such traditions, ethnographers encode a very particular image of knowledge representation work.

I came to the study of information infrastructures from a different set of genealogies and vantage points. First, I conducted my PhD coursework in a STS department that has historically critiqued constructivist approaches like Actor-Network Theory for disarming scholars from attending to both political power (Winner, 1993) and the politics of representation (Fortun, 2014).<sup>18</sup> Through participation in a collaborative research project designing digital infrastructure for experimental ethnography (see Chapter 2),<sup>19</sup> the set of research methods that I have engaged

---

<sup>17</sup>In the Social Construction of Technology (SCOT) tradition of technology studies, scholars similarly consider how different stakeholders hold different and sometimes conflicting interpretations of a technology's meaning, purpose, and design requirements. While this can give rise to conflict (or friction), the technology eventually reaches a point of temporary stability (Bijker et al., 1987).

<sup>18</sup>Susan Leigh Star (1990) too critiqued ANT in the early 1990s for uncritically celebrating scientific achievements and eclipsing that which is always already marginal to networks, but then endorsed a more critical version of ANT in later publications with Geoffrey Bowker (see for example (Bowker and Star, 1999)).

<sup>19</sup>Experimental ethnography refers to methodologies in cultural anthropology that emerged as the field began questioning the power dynamics in writing other cultures. Since 2013, I have served

with most closely are guided by critical anthropological texts such as *Writing Culture* (Clifford and Marcus, 1986) and *Anthropology as Cultural Critique* (Marcus and Fischer, 1986) - texts influenced by post-structuralist thinking that demand ethnographers to call attention to the power dynamics that emerge as they order and *encode* culture. Further, I began my dissertation research examining knowledge representation on the World Wide Web - an information infrastructure where conflict is widely considered to be just as foundational to the functioning of the infrastructure as cooperation.<sup>20</sup> Thus, the way that I've learned to represent the expertise and knowledge of communities has been oriented by a different set of commitments (epistemological, ontology and language ideological, to name a few) than many existing studies of knowledge representation. I will, as a result, tend to *encode* their thinking and work differently.

I make this point neither to disambiguate my approach to the study of knowledge representation *against* past studies nor to demarcate the "edges" of the arguments. I certainly do not aim to dismiss the studies for embodying a different epistemology, just as much as I do not aim to replicate the studies in order to reproduce their results. For experimental ethnographers and critical semioticians, demarcation, or the encoding of a difference that makes a difference, is a political act, and I very much ally with the politics that Star, Bowker, Forsythe, and other ethnographers of knowledge representation bring to their work.

But with every new reading of an empirical domain, there is an opportunity to thicken and expand the description of that domain - to leverage epistemological and other differences to layer new interpretations and draw out new insights. Privileging pluralism in ethnographic research helps to deconstruct some of the neat formalisms that have long encoded academic conventions - that, as an ethnographer and a scholar, I have to position myself with or against certain traditions or that I have to operate in a mode of either replication (reproducing studies to confirm their

---

as the Lead Platform Architect for the Platform for Experimental Collaborative Ethnography (PECE) (see [worldpece.org](http://worldpece.org)) - a digital humanities platform designed to embody the assumptions and commitments of experimental ethnography. My role in the project has involved translating many ideas from post-structural theory into information architecture.

<sup>20</sup>For instance, Web research has shown how Twitter has been used as a tool for dissent amongst counter-publics (Jackson and Welles, 2015) and how the content on Wikipedia pages often reflect conflict and disagreement (Viégas et al., 2004).

findings), extension (building off of studies to extend the scope of their findings), or critique (dismissing studies for their gaps). Instead, I can operate in a mode of *differential reproduction*, acknowledging that my reading of a domain will undoubtedly be structured by how past studies have encoded it, but also that my own perspectives can enable the characterization to iterate and take on new meaning, at times confirming past findings and at other times contradicting them.

From my own vantage point, and particularly in examining expertise in knowledge representation through the frames of semiotics, I have found that *stabilizing* meaning is not always an aim in the design of semiotic infrastructure. For instance, sometimes, semiotic technologists encode semiotic infrastructure in order to provoke a community to think differently about or to bring new meaning to data. Similarly, in examining infrastructure design semiotically, "data friction" appears less in my narrative than in studies that examine infrastructure design through its designers' cooperative work practices - studies that look at what knowledge engineers do and what resists and impedes them from doing it. Instead, Gregory Bateson's concept of "double bind" - or the communicative dissonance that occurs when an individual is exposed to two conflicting demands - will be a common theme throughout the dissertation - and the subject of the next section.<sup>21</sup>

## 1.5 Double Binds of Expert Boundary Drawing

In her anthropological review article, E. Summerson Carr (2010) argues that expertise is not something that a group of individuals has or holds, but instead that it is something that they do or enact.<sup>22</sup> For Carr, enacting expertise is a semiotic practice; experts verbalize and instantiate what they know according to a particular set of learned, authenticated, and institutionalized ideologies and logics.<sup>23</sup>

---

<sup>21</sup>I would characterize the distinction between friction and double bind to be a difference that makes a difference. Friction impedes forward motion, but there is still forward motion (albeit decelerating). A double bind, movement in any direction, by definition, marks movement away from a competing alternative.

<sup>22</sup>Carr's argument can be related to "performativity" - a concept that has been useful in gender studies (Butler, 1988) and queer studies (Sedgwick and Frank, 2003)) to mark that gender is not something that an individual has or is born with but instead something that they enact or perform according to certain social codes.

<sup>23</sup>Carr positions this argument against (Collins and Evans, 2002) - a controversial article that suggested that expertise is something that is acquired from experience or from interactions with

Semiotic technologists are experts at enacting semiotics - at verbalizing and instantiating what they know about language and meaning according to a particular set of ideologies.<sup>24</sup> They are expert encoders. An aim of this dissertation is to draw out, through ethnographic methods and analysis, the language ideologies according to which semiotic technologists distinguish which differences make a difference in empirical domains, or in other words, how they identify which differences deserve to be encoded, how they characterize the representativeness of the differences their infrastructures encode, and how they structure infrastructures to enlist other differences over time.

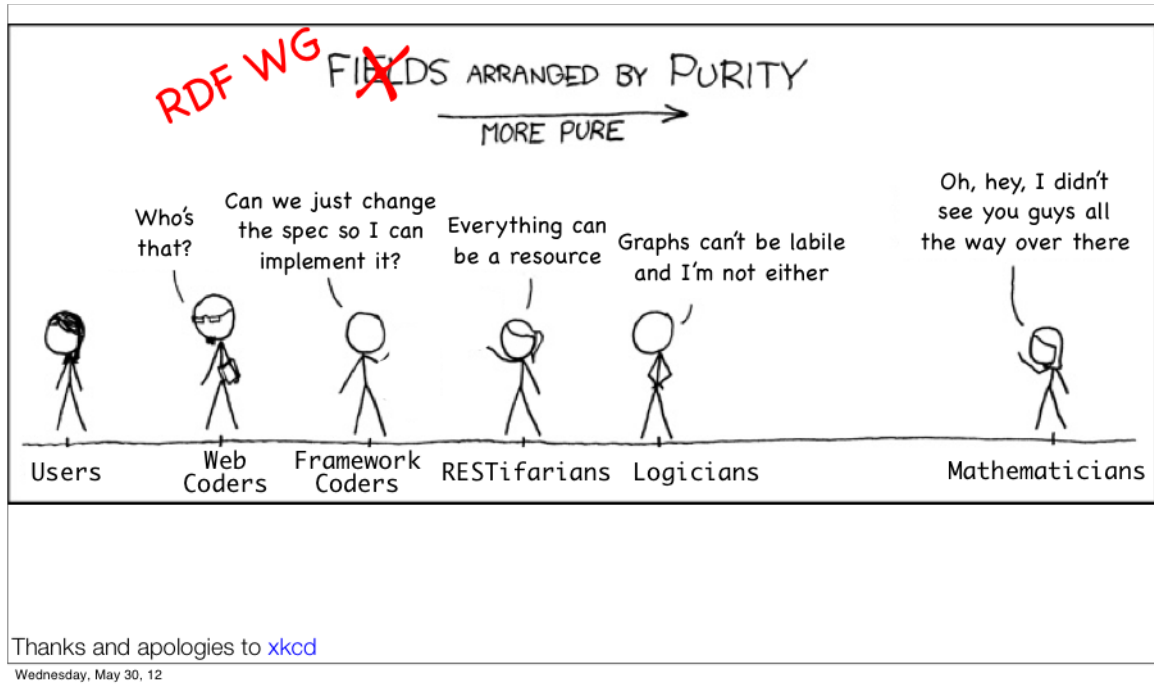
Notably, philosophical theories about logic and about language have already posited several ideological spectrums along which knowledge representation tends to fall. In the philosophy of logic, Jean Van Heijenoort (1967) described a distinction between "logic as language" and "logic as calculus." According to a "logic as language" ideology, logic is seen as universal, and the universe is understood to be fixed. According to this ideology, there is no meaning that exists outside of formal logic. A semiotic technologist enacting a "logic as language" ideology is often dogmatic in their view of "one true" method for formalizing language (which is often first order logic) and assigns meaning to data in top-down ways. On the other hand, a "logic as calculus" ideology assumes meaning to be relativist. According to this ideology, meaning emerges in contexts of differences within complex systems. Thus, rather than devising absolutist or universal models for data organization - models that dictate what a person can say with data - a semiotic technologist enacting a "logic as calculus" language ideology designs semiotic infrastructures that instead codify *how* you can say things with data; meaning then emerges from the bottom-up.

In the philosophy of language, another ideological spectrum along which knowledge representation falls concerns how semiotic technologists think about how meaning operates in particular language systems. According to a structural-functional ideology, meaning is assumed to eventually stabilize within a language system, even if there is initially polysemy or "interpretive flexibility." Thus a semiotic technologist

---

other experts.

<sup>24</sup>In using the word ideologies here, I am referring to the assumptions that communities hold about how sign systems operate.



**Figure 1.2:** David Woods, a computer scientist involved in the design of several Semantic Web frameworks, re-purposed this classic cartoon to demonstrate one semiotic technologist ideological spectrum. (Reprinted with permission).

enacting a structural-functional ideology often begins the semiotic infrastructure design process by working with communities to untangle the different meanings that they bring to concepts in order to get them to agree on some shared definitions. On the other hand, a post-structural ideology takes as given that there will always be perpetual slippage in language, or that meaning will never fully stabilize. Thus, a semiotic technologist enacting a post-structural ideology cobbles together semiotic infrastructures in ways that aim to tolerate and even embrace both difference and change, while sacrificing structure and stability.

No one semiotic technologist serves as the perfect archetype for any of the above categorizations, though there are certainly individuals that lean closer towards a certain ends of these spectrums. And the community often acknowledges the ways in which these spectrums orient design work. For instance, semiotic technologists often contrast "pure theorists" from "scruffy programmers." See 1.2.

However, there are technical tradeoffs to working at the farther ends of the

spectrums. Many semiotic technologists in the communities that I've studied acknowledge that "pure theorists" are really good at enabling certain kinds of representation and computation, while "scruffy programmers" are really good at dealing with the inconsistencies that emerge "in the real world." Often, in an effort to stretch the limits that any one language ideology poses, semiotic technologists describe attempts to find "middle ground." Reference to moving work towards the "middle" has arisen often in my fieldwork. Some semiotic technologists have described designing "middle" infrastructures - somewhere between classifying everything and classifying nothing. Others have suggested that, in semiotic infrastructure design, "a little semantics goes a long way." These efforts emphasize charting a space "in-between" by meeting somewhere in the middle, by trading off consistency for representing "real world" complexity or vice versa.

Yet, as I will show throughout the dissertation, there are often just as many tradeoffs to approaching knowledge representation from the middle as there are to approaching it from the farther ends of the spectrum. In other words, there are times when negotiating a middle approach results in more violent representations than representing knowledge with greater consistency or representing knowledge with greater complexity. Because of this, semiotic technologists are beginning to acknowledge their own language ideologies as assumptions about language and meaning that may not hold in every context, and they are beginning to acknowledge the merits of allowing alternative ideas about language to orient their work in certain times and contexts. In doing so, they are also coming to acknowledge the double binds of representing difference - that in order to represent difference, they have to encode it (or to represent it as homogenous), but as soon as they represent it as homogenous, they erase the differences within it. As they've found themselves in this paradoxical position, some semiotic technologists have begun to characterize their enactments of expertise as something other than negotiation or equivocation. They have described their work as performing "tricks," or what I call "devious design" strategies (Poirier, 2017).

Notably, as Gal and Irvine (1995, 974) argue, ethnographers too are experts, semiotically enacting expertise in drawing boundaries between particular languages

and language ideologies. In indexing the language ideologies of a particular culture, ethnographers run the risk of erasing linguistic differences within cultures:

Erasure is the process in which ideology, in simplifying the field of linguistic practices, renders some persons or activities or sociolinguistic phenomena invisible. Facts that are inconsistent with the ideological scheme may go unnoticed or get explained away. So, for example, a social group, or a language, may be imagined as homogeneous, its internal variation disregarded. Because a linguistic ideology is a totalizing vision, elements that do not fit its interpretive structure- that cannot be seen to fit- must either be ignored or be transformed.

Notably, in drawing boundaries around the language ideologies in the way that I have I pose the risk of "erasing" devious expertise - an expertise in creatively enduring double bind. Therefore, in order to trouble the spectrums that I have neatly described above, I introduce two additional ideologies - ideologies that are impossible to situate between two ends of the spectrum. Experts enacting these ideologies instead struggle to tolerate conflicting demands.

In enacting a critical ideology, semiotic technologists are often aiming to advance an agenda with a semiotic infrastructure - e.g. to make a case that a sexual assault took place or that an officer's interaction with a citizen represented police brutality. To do so, they must produce semiotic infrastructure that clearly defines terms like "sexual consent" or "police brutality." Yet, they also recognize that the more they nail down these definitions, the more likely they are to erase diverse and relevant cases that the definitions do not encompass.<sup>25</sup> Thus, (much like the ethnographer who catachrestically defines populations), in enacting a critical ideology, semiotic technologists are in the doubly bound position of at once needing to represent a meaning homogenously, while also needing to not completely homogenize its meaning.

In enacting a devious ideology, semiotic technologists aim explicitly to design semiotic infrastructures that undercut the force of dominant discourse. In this sense,

---

<sup>25</sup>This is akin to what Kim Fortun (2012) describes as "discursive risk."

their goal is not to reduce friction but instead to intensify it - to define "against the grain." They use essentializing definitions strategically, ironically, or deviously in order to expose how the historically-entrenched assumptions built into discursive systems can "erase" difference.<sup>26</sup>

My own drawing of boundaries between language ideologies that semiotic technologists bring to their work comes from a doubly bound position (one in which I acknowledge a need to characterize different language practices, and in the process of encoding these differences inevitably erase the variations within them). I began my fieldwork in the wake of the 2016 U.S. presidential election - a time where popular discourse began promoting the phrase 'post-truth' to describe a new form of governing. Post-truth discourse posed a risk that citizens, regulators, and scientists would lose "trust in numbers" (Porter, 1996) and data more generally, ignoring empirical evidence that global temperatures are rising, and racism is escalating. With this in mind, I approached my research with a genuine concern about the negative tenor that had recently emerged in much of the social science literature critiquing "big data" for assuming that numbers can "speak for themselves."<sup>27</sup> I wanted to not only characterize and also advocate for the critical modalities that data infrastruc-

---

<sup>26</sup>I consider feminist post-structuralist Luce Irigaray to be one such deviant semiotic technologist. Irigaray (1980) has been criticized for her "biological essentialism" as she offers an alternative model for dominant discourse - a model revolving around the lips (used polysemically to refer to both mouth and labia) in opposition to Jacques Lacan's privileging of the phallus. Yet, Maggie Berg (1991, 57) has noted that read as "ironic critique," Irigaray does not attempt to displace the phallus with the lips, but instead to render visible how the assumptions built into discursive systems have enabled the phallus to become a privileged signifier. Berg writes:

Irigaray mimics Lacan's phallus in order to expose it; elsewhere she explains that ironic imitation is a strategy for uncovering the repression of women: "To play with mimesis is thus, for a woman, to try to recover the place of her exploitation by discourse, without allowing herself to be simply reduced to it. It means to resubmit herself...to...ideas about herself, that are elaborated in/by a masculine logic, but so as to make 'visible,' by an effect of playful repetition, what was supposed to remain invisible." Irigaray's lips are a "playful repetition" of Lacan, exposing Lacan's failure to sustain his premise that gender, constructed entirely within discourse is unstable, arbitrary, and therefore open to choice.

<sup>27</sup>Since Chris Anderson (2008) provocatively suggested that big data marked an "end of theory," where swaths of research numbers could "speak for themselves," big data has garnered a great deal of criticism from social analysts. boyd and Crawford (2012) have written that the way that big data is understood culturally can be characterized as "mythology" - "the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy."

Table 1.1: Ideologies of semiotic infrastructure design.

	Aim	Foundation	Process	Privileged in their work
“Logic as language”	Completeness and Soundness	Logic	Mathematical Modeling	Universalism/Formalism
Structural-Functionalist	Consensus	Existing Definitions	Negotiation	Stability
“Logic as calculus”	Emergence	Systems	Standardizing Syntax	Relativism
Post-structuralist	Communication	Signs	Cobbling	Pragmatism in the Face of Drift
Critical	Advance a Political Agenda	Corner Cases <sup>a</sup>	Tricks	Politics of Representation
Devious	Undercut Dominant Discourse	Catachresis	“Ab-use” or Use from Below (Spivak, 2012)	Resistance

---

<sup>a</sup>Corner cases are use cases designed to demonstrate an exception.

ture designers were learning to bring to their work, acknowledging that not all data analysis follows one logic, one set of assumptions, or one set of politics. I divided these language ideologies to highlight this difference, recognizing that while drawing such boundaries must lead to erasure of difference - that such schemas will always be incomplete - it is also impossible to communicate difference without identifying it.

The figure of the semiotic technologist demonstrates how and when experts are provoked to deconstruct the binary oppositions that structure their work and their thinking about their work. In this dissertation, I call attention to the conditions under which experts learn to question their own semiotic enactments of expertise - re-evaluating their language ideologies and logics. Oftentimes such troubling is provoked as semiotic technologists attempt to design semiotic infrastructure "in the real world."

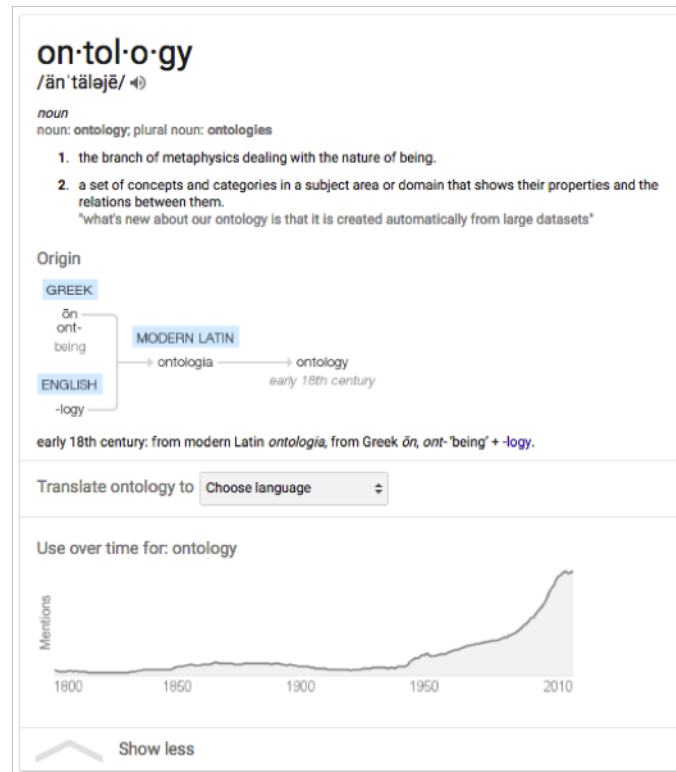
## 1.6 Ontologically Representing (in) Real Worlds

When I typed the word "ontology" into the Google search bar in July 2017,<sup>28</sup> no knowledge graph results displayed but two definitions appeared in a box above the search results. See 1.3.

I clicked down to look at the history of the term's use, recognizing a spike in the 1980s through today. The spike coincides with the timeline many of my ethnographic informants have given me for when the term "ontology" first was used in computer science. Tom Gruber (1993) published a paper defining an ontology as a "specification of a representational vocabulary for a shared domain of discourse – definitions of classes, relations, functions, and other objects." Since then, the term has been used to describe a formal specification for how the knowledge of a particular

---

<sup>28</sup>This, again, can be read as a (re)iteration of an exercise that Kim Fortun (2014) performs in her article "From Latour to Late Industrialism." When Fortun queries "ontology" in the AIME platform, several references to ontology in Latour's work appear as well as several comments from reviewers. In Latour's platform, the aim is to draw out the multiplicity of meanings that readers bring to inquiry - in an effort to "critique the moderns" (Latour, 2013). In this sense the platform does not operate according to a universalist language ideology. Yet, in the system, vocabularies are defined ahead of time in order to provide "common ground" through mediation, and the end goal is to provide "a more precise definition for the experiences gathered under the vague expression 'modernization'." In this sense, the platform can be said to operate according to a structural-functional language ideology.



**Figure 1.3: Google Results for 'Ontology' | (Google and the Google logo are registered trademarks of Google Inc., used with permission.)**

domain is modeled for computer consumption - how data is structured to become *meaningful* to a machine.

Notably, over the past decade, the term "ontology" has also taken on new meaning in the social sciences. In anthropology, "ontology" marks a philosophical "turn" that aims to take seriously that different cultures may not just perceive the world differently than Euro-American anthropologists, but also that being and existence is different in these cultures (Viveiros de Castro, 2002; Latour, 2013). In other words, the research conducted in this spirit is a "turn away from epistemology" towards the documentation of multiple "ontologies." The "turn to ontology" in STS is related but holds slightly different significance, marking a turn towards scholarship that grants an agential role to material objects in order to document how worlds and knowledges are enacted through scientific practice (Barad, 2007; Pickering, 2010).<sup>29</sup>

<sup>29</sup>Gad et al. (2015) argue that, while anthropology epistemologizes ontology, bringing new ideas and concepts to the study of multiple worlds, STS ontologizes epistemology, ethnographically

The word "ontology" is thus both polysemic and diachronic - two properties of language that pose incredible challenges to semiotic technologists. Many computer ontologies are designed to model *shared* definitions or conceptualizations; since terms are often defined in formal specifications by how they relate to other terms, computer ontologies do not cope well with semantic drift. Thus semiotic infrastructure design plays out differently in domains where language is more stable, and definitions are more agreed upon. In domains where language is messier, more contentious, and more open to change, semiotic technologists describe how the challenges of "real world" knowledge representation are provoking them to push the limits of semiotic infrastructure design. In other words, they describe how the "reality" of a domain pushes back on the way they perceive knowledge representation, provoking a retooling of their languages and language ideologies.

The distinctions between *epistemology* and *ontology* then become much scruffier in the design of semiotic infrastructures. In this domain, language and meaning are not just ideological. They are both epistemological and ontological - both real and representational, co-constituted through ideas about language and about reality, practical modes of representation, the materiality required to transport ideas (Hayles, 1993), and the material and semiotic dynamics that organize knowledge in an empirical domain.<sup>30</sup> Semiotic infrastructures both represent worlds and enact them.

In this sense, the figure of the semiotic technologist - an expert that is learning to critically represent polysemic, diachronic, and scruffy knowledge in a set of messy worlds - should intrigue cultural analysts - experts that are also learning to critically represent culture(s). Terminological disambiguations in social research have been rather "neat." A turn to ontology has been characterized as a turn away from epistemology (Kelly, 2014). Actors and actants, humans and non-humans, are to be treated *symmetrically* in these traditions (Callon, 1986; Ashmore et al.,  


---

studying how scientists enact and constitute worlds and knowledges.

<sup>30</sup>Yuk Hui (2016) grapples with philosophical questions concerning the "existence of digital objects" by examining the history and existential structure of several Semantic Web technologies in dialogue with Martin Heidegger and Gilbert Simondon - two theorists of Being. The question of whether there is a "digital real" was the subject of a set of essays published on the journal *Cultural Anthropology's* website in 2016 (Knox and Walford, 2016).

1994). Ontological worlds are often classified into categorical ontologies. Philippe Descola (2010), for instance, categorizes modes of being into a four-fold schema of ontologies - animism, totemism, analogism, and naturalism. And even as Bruno Latour (1991, 2013) seeks to deconstruct the nature-culture binaries that have oriented social research as he inquires into "modes of existence," he introduces new ones, distinguishing the moderns, who organize their worlds according to such binaries, from non-moderns, who don't. Kim Fortun (2014) thus points out that Latour's project is in fact quite ideological - operating according to a similar language ideology that is used to sustain industrial order.<sup>31</sup>

"In the real world" or in real worlds, such distinctions are not so neat - because cultural analysts only have access to alterity through language and representation that is always ordered by certain language ideologies; cultural analysts enact language through practical modes of representation, and language itself has multiple realities,<sup>32</sup> particularly in scruffy domains like that of linguistics and culture. Cultural analysts too may need to retool their languages and language ideologies to deal with the real world(s) - world(s) where *there are limits to encoding difference* and where moving "to the middle" is sometimes impractical, sometimes impossible, and sometimes politically ignorant. Tom Boellstorff (2016, 397) offers one suggestion:

Rather than a turn from epistemology to ontology (and inevitably, back to epistemology), I wonder about a kind of Heisenberg principle that could allow for the possessive co-constitution of ontology and epistemology as fact and perspective, like a photon can be a wave or a particle. Meaning is ontological-enacted in representational practice.

Karen Barad (2007, 176) too offers techniques for troubling the boundaries that have stabilized distinctions between epistemological and ontological approaches in science studies. Her concept of "agential realism" is based on a Bohr ontology -

---

<sup>31</sup>Fortun (2014, 321) describes Latour's project as aspiring "to provide a 'middle ground' for working through different ontologies, in the building of a common world." In trying to sort out and stabilize this "middle," Latour does not tackle the double binds of knowledge representation.

<sup>32</sup>Mangnus Course (2010) argues that the grammars of the languages ethnographers use to document other worlds (or other perspectival ontologies) have their own ontologies, and thus can distort the phenomena they seek to represent.

one that "does not entail some fixed notion of being that is prior to signification (as the classical realist assumes), but neither is being completely inaccessible to language (as in transcendental idealism) nor completely of language (as in linguistic monism)."<sup>33</sup>

Politically concerned with how semiotic infrastructures produce marginalities - how they render subalterns (those outside the network) speechless (Spivak, 1988) - from my own vantage point, I think that we need to acknowledge all of our ontologies, in many senses of the term, as what Derrida (1994) calls *hauntologies* - a term that acknowledges all ontologies to re(-)present always already absent presence - specters of meaning. It is this ghostly property of signification that provokes Derrida (1983, 351) to mark the differences that make a difference between polysemy and dissemination. Polysemy - an excess of meaning - emerges "within a semantic horizon;" we can only see meanings as multiple when we look at writing through a particular frame - through the lens of a singular context. Polysemy assumes that meanings can be untangled (that the "edges" of meanings can be disambiguated) because the concept of polysemy "forgets that its horizon is framed." Yet, for Derrida, writing does not move within a semantic horizon or in a singular context. Every mark, written or spoken, *disseminates*; "explodes the semantic horizon," taking on new meaning - dispersing, iterating - as it moves into new undeterminable contexts - into different ontological worlds (Derrida, 1982a, 45). A mark can only "explode the semantic horizon" because its meaning is never fully exhaustive or enclosed, and meaning that is never fully present can only be represented as hauntology.

As ethnographers index and encode (in) "real worlds," they work to recognize how their semiotic infrastructures - their own ontologies, again in many senses of the term - are never semantically or structurally complete, consistent, or stable. We need to enact a knowledge representation that acknowledges that the way we divide worlds and ideas, and the way that we catachrestically disambiguate the "edges" of our terms, will always erase "other" ways of knowing and being because it always happens within a particular semantic horizon, in a particular "moment

---

<sup>33</sup>Though, Barad (2007) too, at times, ends up "in the middle" of things - which can be seen in her directive to "meet the universe halfway." I don't want this footnote to be read as critique since sometimes escaping the middle is both necessary and impossible.

of meaning." We need to enact a knowledge representation that acknowledges that the "reality" of the domains that we write will inevitably play a part in deferring meaning, presenting new contexts and new modes of interpreting what is always already polysemous and more than polysemous.

This dissertation project studies those people trying to invent, deviously, those much-needed systems where meaning can be both overdetermined and underdetermined - systems that both multiply difference and recognize that plurivocity will never be exhaustive. I follow them as they figure out how their languages and their language ideologies (as well as my own) can limit their capacity to encode difference, and as they also figure out the "tricks" by which they (and I) figure out how to work in the face of those limits.

## 1.7 Mapping the Dissertation

In the chapters that follow I characterize the challenges of designing semiotic infrastructure for several messy "real world" contexts and the strategies that semiotic technologists in these domains have developed for enduring double binds. Notably, my own thinking about the interweaving of language and digital infrastructure began early in my graduate education through my involvement with the design of the Platform for Experimental Collaborative Ethnography (PECE) - a digital humanities platform that aims to support data sharing and collaborative analysis amongst anthropologists and historians. The team of researchers conceptualizing the design of PECE has been committed to advancing ethnographic research that is attentive to the politics of representation and assumes that the meaning an ethnographer ascribes to data is always temporary and partial. The challenge of building a digital platform to support this type of research is not just that it needs to archive, order, and present heterogeneous data, but also that the digital architecture of the system itself needs to embody our assumptions about language and our commitments towards representation. Many existing digital infrastructures do not. Thus, to start, Chapter 2 narrates how, as the Lead Platform Architect for PECE, I translated theories about language into information architecture - a task wrought with double bind. Chapter 2 argues that information infrastructures must be understood

through the lens of the genealogies of thinking that have informed their design, as well as the infrastructural and cultural limits that have constrained their design. I will reiterate these themes throughout the dissertation.

In working on PECE, I had an opportunity to interface with researchers in the Tetherless World Constellation (TWC) at Rensselaer Polytechnic Institute. TWC has been at the forefront of advancing the development of the Semantic Web. In collaborating with researchers at TWC on research projects and conference papers, I found that many of them brought to their work different ways of thinking about and talking about what it meant to "semantify" data. I also found that they would reference how their own ways of thinking about semantics were rooted in different research genealogies in artificial intelligence. I thus decided to launch a research project that ethnographically examined the Semantic Web community - seeking to understand how Semantic Web experts think about language and how their ideas about language get interwoven into digital infrastructure, impacting the shape and import of Web knowledge.

Chapters 3 and 4 are based on four years of historical and ethnographic fieldwork with the Semantic Web community. Chapter 3 narrates how diverse genealogies of thinking - about logic, language, and knowledge - in the field of knowledge representation have influenced how designers of Semantic Web infrastructure talk about and approach their design work. Drawing on ethnographic interviews, an extensive literature review of knowledge representation from the late-1960s to the early-1990s, and analysis of public archives where the design of Semantic Web infrastructures had been planned, the chapter argues that a series of oppositions (such as cleanness/messiness, breadth/depth, logic/procedures) have ordered Semantic Web infrastructure design. It demonstrates how these oppositions have been challenged as semiotic technologists grapple with making knowledge representation applicable to the "real world." Chapter 4 describes how advancing "real world" knowledge representation has forced Semantic Web designers to learn to cope with double bind and to experiment with trickier modes of knowledge representation. This, I argue, has prodded Semantic Web designers to rethink legacy assumptions about logic and about language, often provoking them to see knowledge representation more as a

pursuit than as an accomplishment. Chapter 4 draws on ethnographic interviews with key figures involved in conceptualizing the design of the Semantic Web, participant observation at Web conferences, and analysis of Semantic Web design forums.

Finally, Chapter 5 moves into a knowledge domain where the stakes for semantic stability and consensus are much higher. In March 2017, at the New York City School of Data workshop, I heard a presentation about an effort to design a data standard for information and referral - the branch of the human services domain that is concerned with helping people find public services for which they are eligible. For me, the effort embodied the political urgency of designing standard languages; developing common ways of talking about the human services is quite important for ensuring that people in need can quickly and accurately be directed towards services that can help them. Following this, I began interviewing designers of a variety of semiotic infrastructures for information and referral and analyzing design forums where the planning for these infrastructures took place. I recognized a lot of the same issues that had arisen in my study of Semantic Web infrastructure design; in the human services, due to bureaucracy, politics, and language habits, the meaning of words can be particularly challenging to pin down. In Chapter 5, I show how many semiotic technologists in the human services have had high hopes that semiotic infrastructures can help untangle muddled meanings in the human services, prodding the field towards adopting a shared language. However, these "standards" have never quite become standard. I argue that (mirroring trends that I describe throughout the dissertation) semiotic infrastructures in the human services have moved from very formal things to rather messy things. Semiotic technologists in this domain have come to understand standards as both necessary and impossible to fully achieve. They have experimented with opening these standards up, so that the meaning they represent can iterate and evolve. Newer, more experimental semiotic infrastructures in the human services are thus more about enabling collaboration and exchange than sorting out polysemy.

Together, the chapters advance understanding of data infrastructure design work, demonstrating how data practitioners learn to represent complex and messy knowledge. I examine how semiotic technologists, in many senses of the term, learn

to bring critical and deviant modalities to knowledge representation.

## 2. THINKING ABOUT SEMIOTIC INFRASTRUCTURE: A GENEALOGY

My thinking about infrastructure, about language, and about ethnography has been considerably shaped by my involvement with the design of the Platform for Experimental Collaborative Ethnography (PECE) - a digital humanities platform that aims to enable dispersed ethnographers to openly archive their data, collaboratively analyze it, and experiment with new forms of ethnographic publication. I have been collaborating with colleagues on the design of PECE<sup>1</sup> since 2012 - the final semester of my undergraduate education in Information Technology and Web Science (ITWS) at Rensselaer Polytechnic Institute (RPI). At the time, PECE was not a standalone platform. Instead, it was a series of websites, configured with the same data models, interfaces, and functionality,<sup>2</sup> supporting international research projects such as *The Asthma Files* and the *Disaster STS Network*.<sup>3</sup>

I was invited to participate in the design of PECE to help advance the development of new backend features. As it turned out, I did not have the coding skills to do this well. However, having a dual undergraduate major in ITWS and STS from RPI, I was particularly good at drawing up technical documents and communicating to developers what the design team wanted the platform to enable. I came to understand that my skill set was primarily in translation - more specifically, in translating between social scientists and computer scientists. I sometimes had to explain to the developers the theoretical reasons why the design team did not want

---

<sup>1</sup>The PECE design team refers to the group of researchers conceptualizing the design of the digital platform. While different folks have been involved on the team over the years, researchers that have made significant contributions to its design include: Erik Bigras, Brian Callahan, Brandon Costelloe-Kuehn, Dominic DiFranzo, Kim Fortun, Mike Fortun, Luis Felipe Murillo, and Lindsay Poirier. When I refer to "we" and "us" in this chapter, I'm referring to this team.

<sup>2</sup>Prior to my involvement with the project, PECE had gone through several iterations - starting out as series of PowerPoint slides, eventually moving onto a series of wikis, and finally ending up as a series website configured with the content management system Plone.

<sup>3</sup>*The Asthma Files* ([theasthmafiles.org](http://theasthmafiles.org)) is an international research project examining the culture of scientific research and governance responding to the global asthma epidemic. The *Disaster STS Network* ([disaster-sts-network.org](http://disaster-sts-network.org)) is an international network of scholars examining the cultural import of disasters across event, geography, time, and scale.

to follow technical "best practice" in the platform's design, and I sometimes had to explain to the PECE design team why it was technically difficult or impossible to configure the system in the ways that they wanted.

My graduate training in STS and ethnography helped me advance this translation skill. One particularly serendipitous semester, I took an independent study on Language Theories with Kim Fortun (one of the principal investigators on the PECE project) at the same time that I took a seminar studying design theories with Dean Nieusma. Doing the readings for these two courses in parallel helped me think through how the design team's assumptions about language influenced the ways in which we configured the design of the platform. It also helped me think through how the existing digital infrastructures that we leveraged to advance the platform's design were often themselves designed according to assumptions about language that clashed with our own. As we began to formalize PECE from a series of site configurations towards a standalone platform that any research team could download and install, my main role in the design of the platform was given a title - Lead Platform Architect. This role involved translating the theoretical commitments that guided the design team's work into digital architecture so that PECE would order ethnographic data in ways that embodied our language ideologies.

A key argument throughout this dissertation, and the main argument of this chapter, is that semiotic infrastructures should be interpreted according to the genealogies of thinking and research that produced them. I've placed this chapter before the chapters that constitute my primary research project, because it helps articulate the genealogies of thinking that helped shape my dissertation project into what it is. While my formal research questions have iterated often as my research has developed, since I began conceptualizing my dissertation project, two main themes have structured how I approach my field work - first, that information infrastructures have particular ideas and assumptions about language interwoven into their architectures, and second, that, in order to represent complex information, information infrastructure designers often have to figure out creative ways to endure double bind. In this chapter, I demonstrate how both of these theoretical commitments can be traced back to my involvement with the design of PECE. In

this sense, this chapter records provenance for the dissertation as a whole - providing historical metadata that should help the reader understand some of the reasons why I asked the research and interview questions that I did and why I privileged certain ethnographic observations over others.

I begin the chapter by characterizing how the design of an information system's architecture plays a role in how that information system produces meaning. This is important for understanding why the language theories we have brought to the design of PECE matter for how the platform represents ethnographic knowledge. I then go on to describe the structure and functionality of the platform through the lens of the diverse theoretical and methodological commitments that have informed its design. In the process, I characterize various challenges that we have had to learn to work through in order to advance PECE towards adequately representing complex knowledge about culture. Finally, I close the chapter by describing how the lessons I learned in working on PECE helped shaped the research that I will unpack in the remainder of the dissertation.

## **2.1 Information Architecture as Semiotic Infrastructure**

The information architecture of a digital system represents the shape of information in a digital system, how information is organized, and the possible ways that users can navigate, access, and interact with information. Compare a website to the blueprint for a building. The information architecture of the website details where things will be placed in different "rooms" (or pages) - or, in other words, how data on the page will be organized. It also details the pathways you need to follow to get from one room (or page) to the next. Sometimes these pathways are like doors; a page can display a series of links that represent where users can go next. Other times these pathways are like elevators; users are given options to input where they would like to go - by entering a term into a search box, for example. Notably, the information architecture of a digital system includes more than its data model, which specifies how data in the system is structured and relates to other data. Information architecture also includes the interfaces that visualize how data gets linked together and the tools that users can leverage to navigate, filter, and reconfigure

data. Information architectures structure how users experience a system's data.

Today, information architecture design has been formalized into a professional field. Many companies building digital systems hire information architects to plan and design robust user experiences. There is a professional organization - the Information Architecture Institute (IAI) - devoted to advancing research and education on information architecture design, and there is a journal - the Journal of Information Architecture - devoted to publishing peer-reviewed research on information architecture design. Yet, within this professional field, information architecture has been notoriously difficult to define - to the point where DTDT ("define the damn thing") has become a widely acknowledged acronym used to describe the field's ambiguity.<sup>4</sup> Morville and Rosenfeld (2006, 4)<sup>5</sup> have suggested that the reason the field has been so difficult to define is related to the reason why it is so difficult to design good information architectures:

We're talking about the challenges inherent in language and representation. No document fully and accurately represents the intended meaning of its author. No label or definition totally captures the meaning of a document. And no two readers experience or understand a particular document or definition or label in quite the same way. The relationship between words and meaning is tricky at best.

Information architects see their work as helping to shape how data becomes meaningful. By ordering information in particular ways and making it possible to navigate through information in particular ways, information architects enact an expertise in semiotic infrastructure design. They grapple with the challenges of representing knowledge in domains where the meaning of information can be difficult to pin down.

Information architectures represent meaning material-semiotically. In ordering digital interfaces and tools, information architectures exemplify what Lucy Suchman (2006) refers to as "ordering devices" - or the material-semiotic elements of technology design that calibrate a user towards using a technology in a particular way or

---

<sup>4</sup>Christina Wodtke (2001) has compiled a list of definitions for information architecture.

<sup>5</sup>Peter Morville is often considered the "founding father" of information architecture and Louis Rosenfeld is the co-founder of the IAI

deriving a particular meaning from a technology. Information architectures can be designed to make certain data inaccessible - materially silencing certain knowledge. They can be designed to enforce particular pathways through data - encouraging users to interpret the relationships between data accordingly. Information architectures are designed to connect users to information by formalizing and encoding how users move through a digital system. Thus, Morville (2011) describes information architecture work as "building bridges" that help a user conceptualize how parts of a system relate to the whole - how individual interactions with data relate to the grander narrative. Yet, STS literature reminds us that architecture, like bridges, are not apolitical. Citing Langdon Winner's (1986) famous example of Robert Moses's bridges on Long Island being built too low to allow public transportation to pass underneath, Star (1999, 389) writes, "There are millions of tiny bridges built into large-scale information infrastructures, and millions of (literal and metaphoric) public buses that cannot pass through them."

Information architecture may have the closest technical correlation to what Michel Foucault and Jacques Derrida refer to as an "archive." For both Foucault and Derrida, the term "archive" (like the term "information architecture") is ambiguous; as Derrida (1996, 90) writes: "nothing is less clear today than the word 'archive'." Today, we tend to think of archives as formal repositories for historical data. Yet, both Foucault and Derrida understand archives more comprehensively and critically as mechanisms that order knowledge. For Foucault (1982, 129), archives are understood to signify "systems of discursivity" - or the historical and epistemological conditions that govern "enunciative possibilities and impossibilities." Archives provoke these systems of discursivity by "ordering things"<sup>6</sup> in such a way that it becomes possible to say certain things and not others. I categorize this conception of "archive" in the same bucket as "information architecture" because they both emphasize how particular configurations of language and representation shape enunciative possibilities. As their common root "arch" - Greek for "rule" - sug-

---

<sup>6</sup>Suchman (2006) curated a suite of STS concepts under the heading "ordering devices" - including Grint and Woolgar's (1997) concept of "configuring a user" and Madeline Akrich's (1992) concept of "scripts."

gests, both govern<sup>7</sup> how discourse gets ordered. Information architecture is a form of semiotic infrastructure because it orients how people draw meaning from data.

While defining an "archive" more expansively than as a repository for historical knowledge, Foucault still understands archives to be quite historical. For Foucault, systems of discursivity are rooted in and shaped by historical, cultural, and infrastructural conditions. Categorizing information architectures as "archives" suggests that we need to understand them "genealogically," i.e. that we need to understand the legacies of thinking and the legacy infrastructures that have been interwoven into their design. In the next section, I demonstrate how legacies of thinking in cultural anthropology (as well as legacy infrastructures) have been designed into the information architecture of PECE, inevitably influencing how data becomes meaningful for the ethnographers that use it.

## 2.2 Infrastructurally Inverting the Platform for Experimental Collaborative Ethnography

In order to describe the design of PECE, I'm going to walk backwards through the acronym (moving from Ethnography to Collaborative to Experimental to Platform). I've ordered it this way because, with PECE, we have deliberately avoided beginning from the digital platform. Over the past decade, there's been a trend in the digital humanities - to take a new digital tool and to try and imagine how it can be used to advance a particular humanistic endeavor. However, often digital tools have been designed with cultural assumptions that do not align with those held by humanists. As Johanna Drucker (2011, 85-85) writes:

Positivistic, strictly quantitative, mechanistic, reductive, and literal, [digital] techniques preclude humanistic methods from their operations because of the very assumptions on which they are designed: that objects of knowledge can be understood as self-identical, self-evident, ahistorical, and autonomous.

In trying to fit their work into digital tools, digital humanists run the risk of undermining the assumptions and commitments they bring to their work, prodding

---

<sup>7</sup>For Foucault (1982, 44-49), language, and more specifically "objects of discourse," are governed according to a set of rules, or "discursive practices." Like ordering devices, these rules establish the relations between and "order of things" in discourse.

them to "converge" on overly simplistic renderings of culture.<sup>8</sup>

Citing Goffman, Susan Leigh Star (1991) has described how much of infrastructural design work occurs "backstage" - as what she calls "invisible work." Addressing this obstacle, Bowker (1994) recommends studying large-scale technical systems through a practice of "infrastructural inversion" - an approach that involves bringing to the fore the invisible work of infrastructural design. Accordingly, in designing the architecture of PECE, we started by examining the theoretical commitments that informed our work. We formalized these commitments into a series of "design logics" - which I've elsewhere (Poirier, 2017, 73) defined as "critical directives, informed by a design community's habits and assumptions about language, and by its philosophical commitments, which direct the architecture and arrangement of content in the systems they produce." Formalizing design logics helped us make visible some of the typically invisible assumptions (about language and knowledge) that tend to order the design of information architecture.

Notably, PECE is an open source platform - built from existing open source technologies and contributing new digital functionality to the open source community. In designing the architecture of the platform, we leveraged, configured, and manipulated existing open source technologies in order to align the digital system with our design logics. At times we needed to do this *deviously*, undercutting the logics interwoven into existing digital tools without simply cutting them out. As this dissertation progresses I return to and develop a set of related terms - devious, abusive, and tricky - that we (as well as other semiotic technologists we've engaged with) use to characterize the kind of practices and ways of thinking that exceed the

---

<sup>8</sup>Much of the political critique in the field of information infrastructure studies has centered around the inability of information infrastructures to mirror a complex and dynamic world. In narrating a shift in scientific work that has positioned the database as an end product, Geoffrey Bowker (2000) describes how, for the sake of making information modeling simple, classification schemas, databases, and ontologies are set up so as to represent the world in reductionist ways. His concept of "convergence" marks how digital representations and reality begin to appear to resemble each other. Bowker and Star (1999, 47) argue that, in large-scale information systems, the standards and categories that organize information are increasingly converging with the way we understand the world. In such cases, the political dimensions of cordoning off the world's information are rendered "invisible, erased by their naturalization into the routines of life. Conflict and multiplicity are often buried beneath layers of obscure representation." For Bowker and Star, convergence prods end users to fit their understanding of the world into new frames - frames designed according to a very particular set of values, assumptions, and commitments.

simple binary oppositions that allow for straightforward representation and information design. In this sense, rather than trying to fit our work into existing digital infrastructure, PECE emerged from the effort to install our theoretical commitments into a new kind of digital infrastructure that works "against the grain"—but still works. PECE - much like the other digital infrastructures I will examine in this dissertation - must be understood through the lens of these genealogies of thinking, research, and infrastructure that have informed it.

### 2.2.1 Ethnography

In the mid-1970s, inspired by poststructuralist writing in the philosophy of language by scholars like Jacques Derrida, anthropologists began calling attention to the language structures, tropes, and genres of key cultural texts. Historian and literary critic Hayden White (1975), in *Metahistory: The Historical Imagination of Nineteenth-Century Europe*, analyzed how the verbal structures employed in the writing of historical work shaped their broader narratives and arguments, structuring our knowledge of history today according to classical tropes of irony, tragedy, and so on. Around the same time, Edward Said's (1979) *Orientalism* demonstrated, through analysis of media depicting the non-Western world, how cultural representations shaped new forms of colonial power. And Gayatri Spivak's (1988) landmark article "Can the Subaltern Speak?" drew attention to how cultural analysts' tendencies to speak *for* marginalized populations rendered them speechless. The emergence of these new ideas and concerns, often (catachrestically)<sup>9</sup> labeled the "linguistic turn" in cultural anthropology, played an important role in the emergence of experimental ethnographic methods in the 1980s (Clifford and Marcus, 1986; Marcus and Fischer, 1986).

Working in the spirit of experimental ethnography, anthropologists began explicitly calling attention to the power relations provoked in their writing of other cultures, as well as the power relations that were always already present in the systems of language engaged in writing. Cultural anthropologists began considering

---

<sup>9</sup>Marcus and Fischer (1986, x) noted that labeling and defining the direction of the field ran counter to the very spirit underlying the transition – a spirit that encouraged ethnographers to not presume to know the direction the research would take them.

forms of writing that could challenge the authority of the anthropologist's voice. They also began to critique how depictions of culture in discrete moments in time tended to position culture as static and fixed, when in reality, culture is constantly iterating. In this regard, they began considering how experimental forms of writing could better represent "emergent forms of life" (Fischer, 2003). Finally, cultural anthropologists began calling attention to the diverse, multi-scaled systems, in which the communities they studied were embedded. Thus, while ethnography had traditionally examined culture in a geographically cordoned field site, in response to the linguistic turn, anthropologists began issuing new calls for "multi-sited ethnography" (Marcus, 1995).

Kim Fortun and Mike Fortun, who have led the development of PECE since its inception in 2006, have been at the forefront of conversations about innovating ethnographic methods towards these calls. Kim Fortun earned her Ph.D. in Anthropology at Rice University - where both George Marcus and Michael Fischer had been while working to advance experimental ethnography. Her (2001) book *Advocacy After Bhopal* is often cited as an exemplar of experimental ethnography. Mike Fortun has helped advance these methodologies in the field of science studies. His and Herbert Bernstein's book (1998) *Muddling Through* characterizes scientists as experimentalists; its commitment to experimentalism is embodied in its writing, which, much like the scientists it characterizes, meanders and moves by trial and error. Together, Fortun and Fortun served as journal editors for *Cultural Anthropology*, the academic journal perhaps best known for advancing experimental ethnographic writing. During their time as journal editors, they led efforts to conceptualize and build out the journal's online presence, introducing new online features that took advantage of the digital form to supplement the journal's essays with commentaries, author interviews, multimedia, and classroom discussion questions (Fortun and Fortun, 2015). These genealogies have informed the way they think about designing ethnographic research, and since PECE was conceptualized as a digital infrastructure for supporting ethnographic research, the genealogies have also informed the design of PECE.

By the time that I got involved in the design of PECE, the design team had

already formalized a series of "design logics" for the digital system, specifying theoretical assumptions and commitments (inspired by the feminist, post-colonial, and post-structural theories that prompted anthropology's linguistic turn) that they hoped the design of the digital infrastructure could embody. For instance, rather than cleaning up data or attempting to extract signal from noisy data, the system should enable researchers to collect and store more data than they think they need - in acknowledgement that signal may not emerge until later in the analytic process or in response to different queries. The PECE design team refers to this design logic as "valuing noise." Another design logic is "explanatory pluralism" - the idea that the system should invite multiple and diverse researchers to interpret data, in recognition that an individual researcher's perspective is always situated and partial (Keller, 2003, 303). Rather than aiming to derive "truths" from data analysis, the system should highlight where contradictions and paradoxes emerge in data analysis - not only because we presume that such disjunctures emerge inevitably in complex communications (Bateson, 1972), but also because highlighting the "faultlines" in data analysis can mark areas where deeply entrenched ideas about culture can begin to evolve (Traweek, 2000). We call this design logic "crossing scales, working double binds."

In my role as Lead Platform Architect for the system, I have worked to help ensure that the architecture of the digital infrastructure embodies these design logics. However, as the next two sections will show, we have all learned, in the process, that there are limits to designing an information architecture that fully realizes these design logics.

### **2.2.2 Collaborative**

In designing PECE, we have presumed that sharing ethnographic data - so that it can be interpreted and reinterpreted from multiple perspectives (and not just that of a solo anthropologist) - deepens and enriches the contextualization of data. Acknowledging that cultural analysts can only ever offer partial, situated perspectives on cultural phenomena - that their individual analyses will always marginalize certain communities and narratives - we have aimed to design PECE to invite col-

laborative analysis. PECE has been designed to advance "explanatory pluralism" - to elicit multiple interpretations of data - not in an effort to converge on one right interpretation of data, but instead in an effort to perpetually thicken and extend the description of data. As Evelyn Fox Keller (2003, 303) reminds us, in investigating complex phenomena, such explanatory pluralism "represent[s] our best chance of coming to terms with the world around us."<sup>10</sup>

This marks a considerable departure from predominant norms in ethnography. Ethnographers, for a long time, have tended to work in isolation. Typically, an ethnographer will go to a field site, collect their field notes, conduct interviews, analyze the data, and then formalize the analysis into a publication without ever sharing their data with other ethnographers. There are many reasons for this. First, because anthropologists tend to collect their data from human subjects, they remain greatly concerned about releasing sensitive information about vulnerable populations to the public. Further, anthropologists tend to get much more credit (towards tenure or other promotions) for solo publications than they do for co-authored publications.

These have become go-to responses for why building a discourse for data sharing in anthropology has been so slow-moving. However, I've come to understand there to be much deeper concerns about advancing data sharing in anthropology, and they have to do with the information architecture needed to make data sharing in anthropology possible. Ethnographers often do not see their work as producing data - at least not in the way that scientists produce data. Ethnographic analysis is typically developed through extended fieldwork and constant iterations of research questions. Ethnographic analysis *emerges* from interpreting observations in relation to each other; anthropologists continuously toggle between figure and ground - between individual field notes and bigger picture cultural landscapes - in order to eventually make cultural claims. Anthropologists also understand their observations to be inevitably biased according to the assumptions and commitments that they

---

<sup>10</sup>Further, taking seriously the efforts to dismantle colonial tendencies in anthropology, we believe that ethnographers should not have any special ownership rights to data they collect based on observations and analysis of other communities and that data produced as part of research projects funded with public grant money should be considered a public resource.

bring to their work; these biases shape the lens through which the anthropologist interacts with and observes a community. Thus, typically the first time that ethnographic observations are shared is after they've been formalized into a publication - once the ethnographer has had a chance to curate and contextualize observations collected over extended periods of time. In order to challenge an authorial voice that has historically positioned culture as simply "out there," ethnographers often experiment with forms of writing that weave themselves into the narrative. They do so in order to draw attention to how their own positioning within the research implicates how they go about observing, recording, and writing up culture.

On December 3, 2014, at the annual meeting for the American Anthropological Association in Washington DC, a roundtable was held examining the challenges of ethnographic data preservation. The roundtable was entitled "The Lifecycle of Ethnographic Information - Challenges in the Preservation and Accessibility of Ethnographic Data." Five presenters were invited to speak, and a discussion was held afterwards, led by the roundtable organizers, Lisa Cliggett and Elizabeth A Faier.

During the roundtable, Mark Turin presented on his project Digital Himalaya - a project to digitally archive ethnographic materials from the Himalayan region. He argued that new digital archiving tools provided a powerful opportunity to reassemble analog materials, potentially enabling new analytic collaborations. Deborah Winslow, the program director for cultural anthropology at NSF described how, through an NSF grant led by Lisa Cliggett and Oona Schmid, the AAA had recently set up an anthropological data registry, where anthropologists could link to digital repositories where their ethnographic data had been archived. The registry, she argued, helped meet the AAA's ethical guidelines - that 1) ethnographic data should be made available to those studied and 2) that ethnographic data should be preserved and archived.

In spite of these purported promises of digital archiving for anthropology, many folks in attendance acknowledged their reluctance to publicly share ethnographic data. Participants voiced concern that field notes - one of the most tangible forms of "data" for anthropologists - often include an anthropologist's

half-baked ideas. Field notes are taken from an anthropologist's unique perspective - a perspective that implicates what the anthropologist discerns as worthy of jotting down in their field notebook. The session's participants expressed concern that, in sharing their field notebooks, their half-baked ideas would be interpreted out of context - that others would attempt to derive meaning from the data without having the same deep knowledge of the domain and without sharing the perspective that led the anthropologist to jot the observation down.

The session, for me, captured the spirit of the double bind of data sharing in anthropology. To share ethnographic data ethically, anthropologists need semiotic infrastructure that can richly capture and transport the meaning of their data - so that others do not interpret disaggregated data out of context. Yet often the very goal of ethnographic data sharing is to enable diverse anthropologists to interpret cultural data in new ways, bringing new meaning to data. How can we design infrastructure to preserve the meaning of ethnographic data, while also enabling its meaning to iterate and evolve? How do we build infrastructure that respects the concern of anthropological data being interpreted out of context, while also acknowledging that interpreting data out of context may bring richer cultural analysis to data?

The information architecture supporting ethnography, conceived this way, is like a piece of knitting; observations are woven together to produce the final product, and if you were to remove one stitch, the whole thing could unravel. With this information architecture, no one observation can be interpreted apart from the entire narrative. However, "sharing data" presumes that ethnographers will disaggregate their extended observations into discrete components so that others can access them. Anthropologists are concerned that, in slicing up ethnographic data for inclusion in community databases, ethnographic research will become overly reductionist. As Alexander Galloway (2014) argues, the information architecture for this would look more like a crystalline structure - from which individual atoms can be extracted without jeopardizing the integrity of the structure.<sup>11</sup>

<sup>11</sup>Alexander Galloway (2014) has argued that new capitalist-oriented digital humanities work (championed by Google, Facebook, and others) tends to "atomize" cultural data - to cut it into distinct units - which can eclipse the complexity of cultural phenomena. This work assumes (in

Designing an information architecture to investigate complexity, then, is a double bind - because understanding complexity must be a collaborative endeavor (with diverse perspectives represented), but to enable collaboration, we need to structure a space that enables collaborators to respond to something common, and that very structuring reduces complexity. This is a semiotic challenge just as much as it is a practice-based challenge: we do not want to lose the richness of meaning of data collected in particular contexts, as part of particular collections, and according to particular assumptions and commitments. But we also want that meaning to have the possibility of taking on new meaning as it is examined collaboratively. This is one of the limits we have had to learn to pursue in designing the information architecture of PECE.

One of the ways we have approached this is by changing the architectural metaphor; PECE's information architecture is not like a piece of knitting, nor is it like a crystalline structure. Instead, it's like a kaleidoscope (Fortun, 2012). While data gets uploaded to the system in discrete units - as a single interview, image, field note, or document - it is almost always accessed from pages that situate it somewhere within a larger collection. We have designed PECE to support the "juxtaposition" and rearrangement of data; multiple features in the platform have been specifically designed to place ethnographic data side-by-side, encouraging analysis that considers not only data, but also the broader contexts in which it needs to be understood. Notably, this does not resolve ethnographers' concerns that data will be interpreted outside of the analytic lens from which it was produced, but it does enable bringing more diverse analytic lenses to the interpretation of data. At times, the way that data in the platform gets juxtaposed is random, aiming to subvert habitual modes of curation and to surprise researchers with connections they may have never before considered. At other times, in order to advance explanatory pluralism, we've enabled diverse researchers to curate each other's data into collage-like "PECE Essays," demonstrating how multiple, diverse narratives can emerge from the same data points.

We have also, over time, developed a considerable commitment to follow-  


---

 line with a computational logic) that data can be interpreted in discrete chunks and that culture can be understood as a sum of its parts.

ing best practices in research data management. We've learned from librarians, archivists, and scientists that researchers do not necessarily have to wait until data analysis and write-up to do the reflexive work of describing the history and context of data production.<sup>12</sup> Concerned about data being interpreted out of context, these communities helped develop and implement frameworks so that they could, to the extent possible, represent the meaning of their data to interdisciplinary communities interested in leveraging it for diverse types of analysis. For instance, they developed metadata frameworks that encourage researchers of all types to couple their data with descriptions of the data and the conditions that produced the data.

Notably, doing the work of describing the history and context of data production at the data level (rather than as part of a final write-up) is quite time-consuming. And it is difficult to encourage anthropologists to invest time in doing this work when the academic worlds in which they are embedded tend to give more credit to solo-authored research than to collaborative research. However, as Kim Fortun and colleagues (2014) note, the etymology of collaboration [Latin, *con-* ('with') + *laboro* ('work')] suggests that it is going to be laborious.

Notably, as anthropologists begin using PECE to organize and advance their ethnographic projects - as their ethnographic data gets ordered according to the information architectures the PECE design team has structured - we inevitably also become collaborators in their ethnographic projects. Our own assumptions and commitments about ethnography, formalized into the design of the system, also play a part in shaping the cultural claims made by ethnographers using the system. Thus, we've come to see how our design logics also function as metadata for PECE, describing the assumptions and commitments that inform it. To render visible the assumptions and commitments that have guided the design of PECE's architecture, we have created a page (that comes pre-packaged with every download of the platform) that details our design logics. We have made it so that new users cannot edit these design logics. We need users to understand the genealogies of thinking that are interwoven in the design of our system and that structure how

---

<sup>12</sup>And, of course, just because ethnographic analysis has been crafted into a publication does not necessarily ensure that others will interpret the narrative according to how the ethnographer intended.

their data will be ordered. Yet, in designing it in this way, we have undermined the very design logics that we seek to advance, hard-coding authoritative paradigms for how the system should be interpreted, rather than leaving it open to explanatory pluralism. This is a limit of collaborative experimentalism. We need collaborative insight to broaden partial perspectives, and to make collaborative analysis work, ethnographers need to acknowledge and render visible the genealogies of thinking that have shaped the data they share. But they can never fully do this from a partial perspective. As I will show in the next section, our commitment to experimentalism further complicates this bind.

### 2.2.3 Experimental

In *Anthropology as Cultural Critique*, one of the books that heralded the linguistic turn in cultural anthropology, Marcus and Fischer (1986, x) wrote:

A period of experimentation is characterized by eclecticism, the play of ideas free of authoritative paradigms, critical and reflexive views of subject matter, openness to diverse influences embracing whatever seems to work in practice, and tolerance of uncertainty about a field's direction and of incompleteness in some of its projects.

Their reference to "the play of ideas free of authoritative paradigms" can be read as an allusion to Derrida's (1970) article "Structure, Sign, and Play in the Human Sciences" - in which Derrida deconstructed the structural-functionalist tendency in anthropology to search for a centered meaning or a fixed structure to human culture. Derrida critiqued a tendency in the human sciences to immobilize the play (*jeu*, sometimes translated as "freeplay") of meaning and structure. In the article, often acknowledged as the inauguration of deconstruction in the U.S., Derrida marked a key assumption in post-structuralist theory - that meaning perpetually iterates and structures are only ever temporarily stable. Indeed, experimentalism in cultural anthropology is deeply informed by these post-structuralist conceptual commitments.

As Marcus and Fischer note, advancing experimentalism demands tolerating uncertainty and incompleteness. It demands that a researcher acknowledge that

she or he cannot fully know where the research will lead and that surprises that emerge along the way could dramatically shift habits of thinking. To do experimental research, anthropologists must allow for a full range of possibilities to be at "play" from the outset. Charting the direction of the field ahead of time, or aiming to produce a "complete" cultural narrative tends to immobilize these possibilities. Thus, while in the sciences the *reproducibility* of analysis tends to be an aim of data infrastructure design, in designing PECE we have instead aimed to design digital infrastructure that "differentially reproduces"<sup>13</sup> analysis. We did not want a platform designed to ensure every researcher approaching shared data interprets the data in the exact same way, but a platform that allowed and even encouraged an eclectic set of researchers to each bring a slightly different meaning to data, in play with other interpretations. In this sense, PECE aims to enable cultural analysis to iterate.

In order to enable this, we have had to tolerate the system being quite messy. Neat standards tend to formalize a fixed structure and often aim to reduce uncertainty, incompleteness, and iteration. PECE does not attempt to neatly organize data into logical collections; doing so would presume that users know what analysis they are trying to build toward before they actually do the analysis. Instead, PECE "values noise"; we assume that diverse researchers will draw different signal from noisy data - and that the signal drawn from noise will shift over time, as more data enters the system, as the data is interpreted in new contexts, and as new researchers find new moments of "play" in these data (signs) and their contexts (structure). We have come to learn that designing an information architecture that advances experimentation is very challenging. How do you build a structure for a mess? The very idea is paradoxical; as Derrida (1970, 247-248) confirmed, "one cannot in fact conceive an unorganized structure." And since we are building PECE primarily by leveraging and configuring existing digital architectures - architectures that have often been explicitly designed to neatly organize data and workflows - we have had to learn to identify and disrupt ordering tendencies.

Take, for instance, the content management systems (CMSs) on top of which

---

<sup>13</sup>Our ideas about "differential reproduction" have been influenced by Hans Jorg-Rheinberger's (1998) work on "experimental systems." I describe experimental systems in much greater detail in Chapter 4.

we have built PECE. A CMS is a Web application that serves as a foundation for data storage in many digital platforms. CMSs manage how users publish, update, and delete data in the digital systems they support. CMSs have their own information architecture; they specify the type of content that can be added to a system and how that content gets organized.

When I first became involved in the project in 2012, PECE had been designed on top of the CMS Plone. In Plone, users are prodded to create a series of folders into which content will be organized. Since users need to specify folders before the content gets added, built into the very architecture of Plone is an assumption that users already know how to classify their data. This ran counter to our commitment to experimentalism, disallowing the possibility that, through research, new categories could emerge reconfiguring how content would be organized. Further, since folders in Plone cannot overlap (or in other words, data cannot belong to more than one folder), the architecture also ran counter to our commitments to explanatory pluralism - the idea that, for different researchers using our system, data may belong in different folders. As David Golumbia (2009, 210) describes this as part of a "computationalist order":

Each object is defined hierarchically by where it fits into the larger scheme of objects; each object belongs in a class and does not belong in other classes; classes ([except] for some exceptional instances) do not overlap. ...but... the material world does not fit so neatly into the categories our scientific programs prefer. In the world of computers, though, objects do fit into hierarchies neatly, even if it is conceptually clear that the fit is poor.

In 2014, in response to many challenges working with Plone (including the difficulty of finding a developer that could work with us to undermine ordering tendencies), we began redesigning PECE on top of the CMS Drupal. With Drupal, most data is added to the system as a node. Nodes represent a row in a database table; they each have a unique identifier and a series of fields marked by the table's columns - e.g. titles, authors, dates created, file attachments, etc. There is no hierarchy to the nodes that get stored in Drupal; every node added to the system falls into one large pool of nodes. Instead of organizing that pool into discrete

structures like folders, Drupal provides tools to allow users to pull certain nodes out of the pool in particular circumstances.

One of the primary ways that nodes get organized in Drupal is through taxonomies. Every node in the system can be tagged with certain terms, and then later, users can create collections of all the nodes tagged with a certain term. It is possible to create taxonomies in Drupal that act like folders; a site developer can create a taxonomy that only includes a set number of pre-defined terms and then configure Drupal so that, when a user tags a node, they must select one term from this list. Doing so would mimic the act of categorizing the data into a folder. Some would go as far as to say that this is best practice for organizing data in Drupal (Drupal, 2017).

PECE does not take this approach; again, this would presume that we know what those terms are ahead of time. Instead, in PECE, users can tag their content with any term that they deem appropriate. In this way, instead of coming up with classification schemas ahead of time and then fitting data into them, in PECE, schemas can emerge and iterate as researchers move through their projects. The system does not lock the data into one organizational structure; it does not immobilize "play." Further, we've configured Drupal so that nodes can be tagged with more than one term, enabling different users to classify the same data in multiple and diverse ways. This helps advance explanatory pluralism. It also means that later, when a different user looks up a term in the system, they are presented with data from all over the system that their colleagues decided to classify with that term. This creates interesting juxtapositions.

Unfortunately, more often than not, this approach really does produce a mess - where data gets structured with diverse naming and formatting conventions. Different users have different habits of tagging. Some do not put spaces between multiple word phrases, and the system cannot tell the difference between "graymatter" and "gray matter." In this sense, the system cannot pull together data that has been classified according to these separate naming conventions. Similarly, slight differences in the spelling of a term can unnecessarily splinter how data gets classified; for instance, some users may tag content with the phrase "gray matter," while others

classify it with "grey matter." Sometimes users leverage tags to create hierarchies that help them find their own data later; for instance, they may tag a node "Data > Gray Matter > Government Document." But unless this convention has been shared with other researchers ahead of time, this hierarchical tag is not very useful for anyone but the user who created it.

Notably, collaboration can be quite challenging in a messy system. While we intend to privilege experimentalism, uncertainty, and incompleteness when thinking about the design of PECE's information architecture, we also acknowledge that we need users to be able to find data. And it is one thing for an ethnographer to sort through her or his own messy corpus of data, but it is another thing to sort through a messy corpus of data pulled together by several collaborators. About a year into running the project on Drupal, I had to establish some rules for tagging: "Be sure to check the existing tags to see if the tag you plan to use is already listed. Try and match the spelling." "Tag each phrase separately rather than conjoining multiple phrases with punctuation." "When tagging with a phrase, include a space between words."

We have learned in designing the information architecture for PECE that, while pre-imposed structures can counter our commitments to experimentalism, collaboration is impossible without some pre-imposed structures. Through designing the information architecture of PECE, we began thinking about how to advance the idea of "light structure" - structure that can frame spaces for collaboration with the assumption that the frames will both evolve and be interpreted in diverse ways (Poirier et al., 2014). In the following section, I describe some of the ways we have designed the idea of "light structure" into PECE.

#### **2.2.4 Platform**

PECE is a digital humanities platform, freely available on Github, which can be downloaded and installed in order to support any number of ethnographic projects. PECE is an underlying infrastructure for an ethnographic project; it comes pre-packaged with data models, workflows, and interfaces that, to a certain extent, structure the type of content users add to the platform, how that content gets orga-

nized, and how it gets visualized. In my role as Lead Platform Architect, I have led the PECE design team in thinking through how to technically configure these components of the system's information architecture. In doing so, I've thought about how we can take advantage of the unique affordances of the digital form in order to make the architecture both kaleidoscopic and lightly structured.

As literary theorist George Landow (2006) has argued, hypertext can be considered a manifestation of post-structuralist ideas; a hyperlinked text is de-centered and multi-vocal, representing not just the primary author's voice but also the voices of the distributed networks of authors linked from (and thus helping to constitute) the text. Focusing on the materiality of hypertext and what it means for how text is signified,<sup>14</sup> Katherine Hayles (2004) argued that hypertext is generated by fragmentation and recombination - by cutting up a text and then recombining it. In designing PECE, we have aimed to leverage these unique qualities of digital texts to our advantage, designing digital functionality that disaggregates and then recombines diverse analyses in order to de-center the authority of a single ethnographer's voice. This has enabled us to structure the system in ways that enable researchers to respond to something common (promoting collaboration), while also enabling the system to constantly pull apart and reorder analyses (promoting experimentation). Let me explain what this looks like technically.

The main way that users contribute data to an instance of the platform is by creating *artifacts*. The audio of an interview may get uploaded to the platform as an audio artifact, for instance; or a field note may get uploaded to the platform as a text artifact. When creating artifacts in the system, users assign them a title, upload a file attachment, and fill out fields specifying their ownership, permissions, tags, and several other metadata fields. Users can contribute artifacts to the platform as

---

<sup>14</sup>On the materiality of text, Hayles (2004, 72) writes:

The crucial move is to reconceptualize materiality as the interplay between a text's physical characteristics and its signifying strategies. This definition opens the possibility of considering texts as embodied entities while still maintaining a central focus on interpretation. In this view of materiality, it is not merely an inert collection of physical properties but a dynamic quality that emerges from the interplay between the text as a physical artifact, its conceptual content, and the interpretive activities of readers and writers.

a whole, or they can contribute them to a specific group with a more specialized focus.

PECE has been designed to support collaborative analysis of artifacts. Users can contribute structured sets of questions, or *structured analytics*, to the platform that will eventually be shared amongst researchers to annotate various artifacts. So, for instance, a structured analytic for "Annotating Peer Reviewed Research" may include questions such as "What is the main argument of this article?" or "How has this article impacted the way that policy makers think about x?" Once artifacts have been added to the platform, any user with the proper permissions to view the artifact can also annotate it with the questions listed in a structured analytic. Users need not respond to every question in the analytic, and they can also add their own questions to the analytic as they are responding, contributing to the shared pool of questions. Once a user annotates an artifact in PECE by answering several shared questions that have been contributed to the system, the annotation gets linked to the artifact. Another user can navigate to the artifact's page in the system and then follow a link to the author's annotation. Once several users annotate a single artifact in PECE, the artifact's page begins to render the artifact as multi-vocal, providing links to different annotations that analyze the artifact from different perspectives (see 2.1). As multiple users begin to respond to the same artifact with the same sets of questions, we begin to see how their diverse interpretations, or in other words their explanatory pluralism, deepen our understanding of that artifact's cultural import. This light structure - bringing researchers together around shared artifacts and shared analytics - does not aim to reproduce the same analysis, but instead to highlight the diverse and at times contradictory ways ethnographic data can be interpreted. It is produced through a common (structured) analytic frame, but it privileges difference.

Annotations can also be pulled apart and recombined to juxtapose diverse analyses. This was not always the case for PECE, but a product of iterative design. In our first attempts to design a data model, annotations were produced as a user filled out a form; each field of the form was a shared question (see 2.2). With this particular configuration of the data model, it was challenging to disaggregate

The screenshot shows a web page for 'THE ASTHMA FILES' with a logo of two lungs. Navigation links for 'LOG IN' and 'REGISTER' are in the top right. The main title is 'BREATHING CLEANER AIR: TEN SCALABLE SOLUTIONS FOR INDIAN CITIES'. Below the title, there are sections for 'PDF DOCUMENT', 'CONTRIBUTORS', 'CREATED DATE', 'CRITICAL COMMENTARY', 'LICENSE', and 'ANNOTATIONS'. The 'PDF DOCUMENT' section displays a thumbnail of the report cover, which features a circular diagram with ten segments representing different solutions: Energy, Transport, Agriculture, Industry, Buildings, Waste, Water, Air Quality, and others. The 'CONTRIBUTORS' section lists 'adkhandekar'. The 'CREATED DATE' is 'April 20, 2017'. The 'CRITICAL COMMENTARY' section contains a paragraph of text. The 'ANNOTATIONS' section lists several users: Kim Fortun, adkhandekar, Pankaj Sekhsaria, Vinay B, and prema\_srigyan.

**Figure 2.1:** Annotations on an artifact page in PECE represent diverse interpretations of the artifact’s cultural import. Clicking on a link to a user’s annotation will display how different users responded to annotation questions.

questions from specific annotations. For instance, it was rather simple to create a page that displayed a user’s annotation of an artifact; the page would simply list the responses to the form. However, it was much more challenging to create a page on the platform that displayed every user’s response to a single annotation question; it required extracting the question from the form and extracting each response to the question from each annotation contributed to the system. This was not impossible

**ANNOTATION**  
Use this to create a new annotation for this artifact from the list of structured questions.

[Show row weights](#)

TITLE	ANNOTATION TYPE	OPERATIONS
✚ Singapore as Asthmatic Space	Asthmatic spaces structured analytics	Edit Remove

**ADD NEW ANNOTATION**

Title \*

When did this space begin monitoring PM 2.5?

How prevalent is asthma, and who is tracking asthma prevalence?

What spatial units (cities, counties, provinces, regulatory regions, etc.) are important in this area?

How has asthma been covered by local media? Where have causes of and responsibility for asthma incidence been placed?

What kinds of civic organizations (environmental groups, caregivers groups) are involved in asthma surveillance and care?

What kind of research has been done on asthma-related issues in this area, and what are the findings?

Figure 2.2: This was the annotation form in an earlier version of PECE. Each field of this form was a shared question to which researchers would respond for a particular artifact. As fields of a form, it was technically challenging to disaggregate responses to this form in order to compare how the question was answered by different researchers or for different artifacts.





but required significant hacking.

In the newer data model, we aimed to take advantage of the fragmentary and recombinational features of hypertext so that we could juxtapose user analysis in more interesting ways. We deliberately made each question in the system its own entity (or its own *node*), and we deliberately made each response to each question in the system its own entity (or *node*). Each response was linked to the question to which it responded, and questions were curated into a structured analytic with tags. With this new model, it was easy to create pages in the platform that pulled apart annotations to compare responses across users and across artifacts. For instance, a user can click on any annotation question in the platform and see every response

**How would you characterize this city to someone unfamiliar with it? (geographic and population size, rate of growth, etc.)**

City Snapshot (/structured-analytics-questions-set/city-snapshot)

**ANNOTATIONS**

User	Artifact	APPLY
Enter a comma separated list of user names.		
Allisonstarr1994 (/Users/Allisonstarr1994) April 27, 2017	↳ In response to:  <a href="/content/history-community-displacement-mantua">The History of Community Displacement in Mantua (/content/history-community-displacement-mantua)</a>	
<p>Although Mantua was originally an "affluent white Philadelphia suburb," the city soon expanded and low-income African-Americans and immigrants moved into the area. Today, in fact, the majority of the neighborhood's residents are low-income African-Americans. The neighborhood has continuously been threatened by the expansion of Drexel University, with the first major development occurring in the 1950s. Mantua had been designated as an area for redevelopment in 1948 and 1950, and Drexel released a statement of its intent to do so in 1957. More than 1,700 buildings were taken over between 1910 and 1990. In 1970, however, a new development plan led to community outcry.</p>		
Allisonstarr1994 (/Users/Allisonstarr1994) April 27, 2017	↳ In response to:  <a href="/content/powelton-village-university-expansion-destroyed-community">Powelton Village: University Expansion Destroyed a Community (/content/powelton-village-university-expansion-destroyed-community)</a>	
<p>Expansion of the University of Pennsylvania reached its mid-1900s peak following a murder of an international student. Hoping to eradicate crime, the university formed a coalition with Drexel University, University of the Sciences in Philadelphia, and Presbyterian Hospital to obtain land from the area known as "Black bottom." Because this neighborhood was designated as blighted, they were able to seize it via eminent domain.</p>		
 Maliekms (/Users/Maliekms) September 3, 2016	↳ In response to:  <a href="/content/risky-gamble-manage-gentrification-sharswoodnorth-philadelphia">A Risky Gamble To Manage Gentrification-Sharswood/North Philadelphia (/content/risky-gamble-manage-gentrification-sharswoodnorth-philadelphia)</a>	
<p>The rate of growth in Sharswood is negative, riddled by crime, and school closing within the last 10 years it has become a place that has been forgotten by Philadelphians. However, with the Philadelphia Housing Authority, promising transformation, gentrification is coming to the neighborhood, and I anticipate a very different looking Sharswood in the next ten years.</p>		

**Figure 2.3: On Analytic pages in the newer version of PECE, users can compare the responses to a single annotation question - across users and across artifacts.**

to the question in the system - across users and artifacts (see 2.3). They can also filter this page to specific user responses or to responses to specific artifacts. In this sense, the digital form, and particularly the affordances of hypertext, have helped foster experimental collaboration on the system. Individual ethnographic analyses can be fragmented and recombined into collaborative analysis.

A second example of how we've leveraged the digital form to advance ethno-

The screenshot displays the PECE platform interface. At the top left is the PECE logo. On the top right are navigation links: MY ACCOUNT, DASHBOARD, and LOG OUT. The main content area is a collage of several artifacts:

- KALEIDOSCOPIK PERSPECTIVE:** A central image showing a complex, multi-faceted geometric pattern.
- (AB)USE AND PECE:** A text block discussing the concept of "ab-use" and its implications, with a "Read more" link.
- IN PURSUIT OF DIFFERENTIAL REPRODUCTION:** A video player showing a 3D visualization of complex, swirling patterns.
- ANNOTATION IN PECE:** A video player showing a screenshot of the PECE interface with annotations.
- LIGHT STRUCTURE IN THE PLATFORM FOR EXPERIMENTAL COLLABORATIVE ETHNOGRAPHY:** A text block with a purple icon.
- LIGHT STRUCTURE IN PECE:** A diagram showing several interconnected circular nodes.
- EXPERIMENTAL ETHNOGRAPHY ONLINE:** A text block with a purple icon.
- INTERVIEW WITH PECE DESIGN TEAM:** A video player showing an interview.
- RETURN TO PECE ESSAY METADATA:** A purple button at the bottom.

On the right side, there is a quote by Hans-Jorg Rheinberger: "[O]ne never knows exactly where it leads. As soon as one knows exactly what it produces, it is not longer a research system. An experimental system in which a scientific object gathers contours and becomes stabilized, at the same time must open windows for the emergence of unprecedented events." Below the quote is the author's name and affiliation: "Hans-Jorg Rheinberger, 'Experimental Systems, Graphematic Spaces,' in *Inscribing*".

**Figure 2.4:** By creating a PECE Essay, users can pull together data from across the platform and configure it into "kaleidoscopic" views.

graphic collaboration and experimentation is through a feature that we call the *PECE Essay*. When designing a PECE Essay, users curate a collage of artifacts in the system, created by diverse users. They can add headings to the collage and supporting text, but the bulk of the essay is composed of hyperlinked artifacts arranged to convey a particular ethnographic point. The tool enables ethnographers to "play" with different configurations of data. At any given moment, the PECE essay represents one frame of a kaleidoscope (see 2.4). However, the kaleidoscope can turn; ethnographers can add content to the essay and rearrange artifacts as their analysis develops. Two essays, created by different users, can arrange the same data in different ways to convey different points, each representing different frames of the kaleidoscope.

Notably, taking advantage of the digital form in the way that we have has helped us endure double binds that I outlined earlier in the chapter - that we need the system to preserve the meaning of data, while also perpetually enabling re-

searchers to bring new meaning to data, and that we need the system to structure a mess. The platform has been designed so that individual artifacts get attributed to their creators; no matter where artifacts are referenced in PECE, it is always possible to link to their metadata. This helps preserve the context of data throughout collaborative and experimental analysis. But the system also encourages ethnographers to continuously interpret the data in new contexts. Users see ethnographic data in new contexts as they navigate to annotation pages responding to the artifact in diverse ways. They see it in new contexts as they navigate to question pages that display how a user responded to a question for that artifact, alongside how other users responded to the same question for other artifacts. They also see it in new contexts as they navigate to PECE Essays that arrange the artifact alongside other artifacts to advance an analytic point. Hypertext helps lightly structure this; data can be reconfigured in ways that we, the designers of PECE, cannot yet imagine. Different researchers can pull different signal from the noise and make the data meaningful in different ways. Yet, that data always remains linked to the ethnographer that uploaded it to the platform and to the metadata that this ethnographer attributed to it.

For Kim Fortun (2004, 312), this capacity to configure data through processes of reordering illustrates the (often under-acknowledged) promise of informatics:

In allowing for and encouraging reorderings, informatics operate with what can be called a logic of supplementarity (following Derrida's conception), enabling substitutions and additions that have the potential to reconsider what a system can say and do, encouraging displacements and realignments. Informatics thus destabilize established systems by design. Though the context always matters, informatics can be conceived as a material cultural form that is valenced in particular ways. Through the facilitation of constant reorderings and revisualization of one's "object" of concern, informatics tends to push fields in which they operate into iterative rather than into reproductive modes.

PECE advances this type of informatics. It privileges reordering, supplementing, displacing, and destabilizing some things because others have been stabilized. It

has been designed to allow data's meaning to emerge in new and currently unimaginable ways. And importantly, it does so in order to disrupt and undermine systems and forms of language that tend to position an anthropologist's text as a singular authority and that tend to cast culture as static and fixed across time, geography, and scale. Genealogies of anthropological thinking that have been attentive to the politics of representation have been interwoven into the design of the system, impacting how ethnographic data will eventually be made meaningful.

### 2.3 Conclusion

Working towards designing the information architecture for PECE has led me to assume certain things about digital infrastructure, and these assumptions have guided the rest of the research presented in this dissertation.

First, I began my dissertation research presuming that information infrastructures have genealogies of thinking and assumptions about language built into their design. While, in designing PECE, we were particularly deliberate in unpacking the language theories that would guide the design of our system, we also recognized that every digital infrastructure that we encountered ordered and represented knowledge in particular ways. In my own dissertation project, I became interested in tracing the genealogies of thinking that led to these particular orderings and modes of representation. Thus, one of the earliest research questions I brought to my examination of Semantic Web communities was: *What cultural, historical, and infrastructural conditions have shaped the design and affordances of information infrastructure?*

Second, I began my dissertation research presuming that knowledge representation is often a complex, doubly bound task. To represent complex information, knowledge representation experts often have to figure out how to work at the limits of cleanness and messiness. To design a shared information system, they often need to implement standards and undermine them - so that not everyone using the system is required to interpret data in the exact same standardized way. Thus, another question that I brought to my dissertation research was: *How have the designers of information infrastructure engaged with the limits of knowledge representation - particularly those imposed by language or by other infrastructures?*

Finally, having had to learn to grapple with these double binds in designing the information architecture for PECE, I was particularly interested in examining how the designers of information infrastructure learned to work experimentally or "deviously" in the face of these limits.

Approaching the research with this framework I aim to better characterize how and why information architectures tend to order knowledge in the ways that they do and what this means for how diverse knowledge - about ideas, people, and problems - gets represented. I also aim to highlight the "faultlines" - the disjunctures in thinking about information infrastructure - in order to mark potential places where entrenched thinking about information infrastructure can evolve. In this dissertation, the genealogies of these aims can be traced back, in part, to my work on PECE.

### 3. TROUBLING TRADEOFFS IN SEMIOTIC INFRASTRUCTURE DESIGN

The<sup>1</sup> 2016 World Wide Web conference - the 25th International World Wide Web conference - marked several other important milestones for the Web. It had been 25 years since Tim Berners-Lee first published a summary of his invention - the World Wide Web - to the alt.hypertext Internet newsgroup, marking the hypertext system's public debut on the Internet. It had also been 15 years since the publication of the first article publically introducing the concept of the Semantic Web - a system for structuring Web data into machine-readability - (Berners-Lee et al., 2001), 10 years since the publication of the "Semantic Web Revisited" (Shadbolt et al., 2006) - an assessment on the progress and enduring challenges of the semantic web's development - and 5 years since the release of schema.org.

The line-up of keynote speakers for the conference was impressive. Sir Tim Berners-Lee himself presented the kick-off keynote, calling for a re-decentralization of the Web - for a breaking up of the Web silos that had been created by large companies like Facebook and Amazon.<sup>2</sup> When asked where he saw the Web in 20 years, he responded that, if we are to take seriously the push to keep the Web open, we should design the infrastructure recognizing that we can't imagine what people are going to do with it in 10 years. Martha Lane Fox, a crossbench peer for the UK House of Lords and Chancellor of Open University, delivered a keynote discussing the need for gender equality in the tech workplace, along with cultivating STEM education interwoven with liberal arts education. Many conference attendees tweeted out that it was the first standing ovation they had seen at a WWW conference.

---

<sup>1</sup>Portions of this chapter are in Press: Poirier, Lindsay. (2018). Making the Web Meaningful: A History of Web Semantics. In Niels Brügger and Ian Milligan (Eds.) *Sage Handbook of Web History* Sage Publications Inc.

<sup>2</sup>The theme of the conference was "ouvert"; at the opening ceremony, the introductory slides included the subtitle "Come in; we are open." Indeed, the need for increased openness was one of the key themes running throughout the keynotes and general discourse for the conference. With issues such as net neutrality and government censorship ripe in 2016, the need for increased "openness" and decentralization has been considered the most prominent contemporary issue for the World Wide Web (Hardy, 2016).

On the final day of the conference, however, Peter Norvig, Director of Research at Google, delivered the keynote that most caught my attention. Norvig is a distinguished researcher in artificial intelligence and co-authored what is often cited as the most popular textbook on AI (see (Russell and Norvig, 1995)). In his talk "The Semantic Web and the Semantics of the Web: Where Does Meaning Come From?," Norvig (2016) reflected on how Web researchers could best draw meaning from content on the World Wide Web.

In many respects, figuring out "where meaning comes from" has been the key challenge of Semantic Web research. Semantic Web researchers aim to design information infrastructures for adding machine-readable "meaning" to Web data in order to make the Web smarter - to enable a Web that not only presents information, but also understands it. They hope that in doing so, we can better interlink and share the data on the Web, improve Web search, and build smart agents that can sort and make sense of diverse Web data.

In his talk, Norvig recounted an infamous essay written by tech blogger Cory Doctorow (2001) - the same year the first article about the Semantic Web was published. Doctorow's article cast doubt on what he called the "meta-utopia," listing seven reasons why getting people to "mark up their data," or add machine-readable metadata to their digital content, could never produce a sufficiently smart Web. "1. People lie," he argued. "2. People are lazy," and "3. People are stupid" and don't always use technologies for marking up data correctly. He went on: "5. Schemas aren't neutral," and "7. There's more than one way to describe something." Norvig cited these as enduring challenges for the Semantic Web and went on to question how best to overcome them. Should it be with "highly trained logicians," well-versed in complex Semantic Web technologies designed to describe, link, and model web data? Or should it be with "lightly trained webmasters" presented with simple technologies for marking up their own Web data?

He went on to argue that, although highly trained logicians were able to add the most value to Web data (describing and representing it with detailed and complex semantics), in practice relying on intricate schemas and ontologies for semantifying the Web produced a great deal of spam (because people lie), a 40% error rate

(because people are stupid), and a low adoption rate (because people are lazy). However, newer semantic web frameworks like Schema.org (which will be described in greater detail in Chapter 4) aimed to simplify intricate schemas and ontologies, so that a webmaster could simply add a line of code around a data point on a webpage to let the search engine know that it was referring, say, to a restaurant name, and another line to let the search engine know that a data point referred to the restaurant's address. Norvig noted that when Schema.org was introduced, its adoption grew by a factor of more than 1000 in 3 years - a factor that he noted was similar to the viral growth of the Web itself.

This was not the first time that I had heard it suggested that Schema.org represented what was "winning out" in the Semantic Web. Yet, the release of Schema.org was controversial. In an effort to simplify the ontology, Schema.org was much less rigorously defined than previous Web ontologies, leaving more room for webmasters to interpret the classifiers as they saw fit. For instance, a schema.org classifier for SportsTeam left open to webmasters a great deal of interpretation on what counts as a "sports team." Should a chess team be included? Must a sport be competitive? This presented a "messy" problem for many Semantic Web practitioners, particularly the "highly trained logicians" that Norvig referred to in his talk. Without "clean" and "consistent" semantics, it would be much more difficult to produce "intelligent" systems for understanding and manipulating Web content. Yet, despite these early concerns, at WWW16, both Norvig's talk and general discourse suggested that Schema.org would be the technology that advanced the Semantic Web towards widespread adoption.

In my fieldwork, I have found that semiotic technologists have often relied on a series of oppositions<sup>3</sup> to characterize diverse approaches to knowledge representation. Notably, the line that Norvig draws between "highly trained logicians" and "lightly trained webmasters," along with the line that often gets drawn between "clean" and "messy" solutions to problems in Web semantics, mark two opposi-

---

<sup>3</sup>I use the phrase opposition to designate a set of paired terms that are understood to counterpose one another. In this sense, oppositions tend to represent contexts where an individual must make tradeoffs to achieve a goal. In certain contexts, when paired terms are understood to be mutually exclusive, I will refer to them as binary oppositions. Binary oppositions do not represent contexts where tradeoffs are possible. Instead, they represent double bind.

tions that structure how the Semantic Web community has talked about knowledge representation "thought styles" (Fleck, 1981). The answer to the question "where does meaning come from?" gets broken down into similar oppositions - either from a universal logic or from an aggregation of the wisdom of the masses. Approaches to knowledge representation have been described as formal or loosey-goosey, logic-based or crowd-sourced. Semiotic technologists have relied on such oppositions to characterize the disparate ways of thinking (about what constitutes knowledge and how best to represent it) and the distinct communities of practice that design semiotic infrastructure. In this chapter, I trace the genealogies<sup>4</sup> of these oppositions back to epistemological divisions that emerged in the artificial intelligence community in the mid-20th century.<sup>5</sup> Next, I draw on oral history interviews that I conducted with members of the Semantic Web community, along with archival research that I conducted on the World Wide Web Consortium (W3C)'s public forums to show how different styles of thought that guided knowledge representation work in the 1970s and 1980s have become interwoven in the design of the Semantic Web.

Notably, an opposition is itself a rather clean and formal way of describing

---

<sup>4</sup>For Foucault (1982), language, and more specifically "objects of discourse," are governed according to a *discursive practice* - a set of rules that establishes the relations and "order of things" in discourse. These rules do not emerge in a vacuum. Instead, they are rooted in and shaped by historical, cultural, and infrastructural conditions. Studying discourse "genealogically" means to study the historical, cultural, and infrastructural conditions that have shaped present discourse. There are two levels at which to study the genealogies of discursive practices in the design of semiotic infrastructure. First, semiotic infrastructures enact discursive practices. Rules about how data can be related are encoded in digital infrastructure like the Semantic Web and establish an "order of things." Second, discursive practices order the discourses that semiotic technologists use to describe their work. This chapter is focused on examining the genealogies of the latter discursive practices, which in turn implicate the former.

<sup>5</sup>During my fieldwork, I examined several primary sources (primarily journal articles, but also magazine articles and technical documents) that outlined agendas and programs for advancing AI. I should note that many of these documents included logical notations, intricate technical diagrams, and proofs that spanned several pages. While I was able to scan these aspects of the documents to get a sense for the shape and logics of the programs, my aim in "reading" these documents was not to verify a proof or to find its flaws. Instead, I focused on how AI researchers *framed* their programs - how they described their programs' contributions and shortcomings and how they positioned their ideas against others in the field. I think this is especially important to note because many critiques AI researchers have written of each other's work focus specifically on attacking their technical shortcomings; they use proofs to criticize proofs and to show why they could never model how things are in the "real world." Yet, as I will show, hidden beneath many of these critiques are different worldviews about how language should be encoded for a computer - worldviews that contradict each other in fundamental ways yet are necessary for building intelligent systems.

a context. For semiotic technologists that tend to position their work as clean, formal, and rooted in logic, it has been natural to rely on these oppositions to characterize their work. However, for those who view knowledge as never completely complete or consistently consistent - and who characterize their approaches to semiotic infrastructure design as messy and loosey-goosey in response to these knowledge conditions - there is contradiction at the very core of representing knowledge representation work with a series of paired terms. This chapter details how semiotic technologists (in both AI's knowledge representation community and the Semantic Web community) grappled with this contradiction - at times upholding the oppositions, at times calling into question their mutual exclusivity, and at other times deconstructing them.

I argue that, as semiotic technologists began destabilizing the signifiers that had habitually been used to describe their work, the phrase the "real world" cat-achrestically came to stand in for all the language conditions that they could not control (e.g. polysemy, diachrony, and paradox) - the language conditions that emerged outside of the carefully bordered worlds they engineered to test their ideas. I argue that this positioning of the "real world" as synonymous with messy language dynamics contributed to the deconstruction of oppositions in knowledge representation, revaluing messiness over cleanness, inconsistency over consistency, and contradiction over logic. In the conclusion, I show how this revaluing eventually led Peter Norvig to call into question the clean divisions between a real world and an artificial one.

### 3.1 Knowledge Representation Oppositions

On August 31, (1955), John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon initiated a proposal for a research project, to be conducted the following summer at Dartmouth College. The summer project would later be considered the birth of the field of artificial intelligence. The proposal began:

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture

that every aspect of learning or any other feature of intelligence can in principle be *so precisely described* that a machine can be made to simulate it. An attempt will be made to find how to *make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves*. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer. (emphasis mine)

It was not a coincidence that formalizing language for computer-consumption played such a prominent role in the proposal. The proposal was coming at the rise of theories in "structural linguistics" - theories that suggested that a coherent, bounded, rule-based system orders the various components of language. Leading up to the Dartmouth event, theory in structural linguistics had been increasingly viewing syntax (or the grammar of a language), as mechanized - ordered by rules of logic. Rudolf Carnap (1937), a logical positivist and member of the Vienna Circle, published a landmark book *Logical Syntax of Language*, which proposed that language could be formalized into a series of rules.<sup>6</sup> He wrote:

By the logical syntax of a language, we mean the formal theory of the linguistic forms of that language - the systematic statement of the formal rules which govern it together with the development of the consequences which follow from these rules. A theory, a rule, a definition, or the like is to be called formal when no reference is made in it either to the meaning of the symbols (for examples, the words) or to the sense of the

---

<sup>6</sup>Carnap claims his ideas in *Logical Syntax of Language* are greatly indebted to Wittgenstein's (1922) *Tractatus Logico-Philosophicus*. Wittgenstein refers to a "logical syntax" in *Tractatus* while setting forth his picture theory. Picture theory suggests that statements about the world, or propositions, are pictures of reality - "a model of reality as we imagine it." For Wittgenstein, language signifies what can be pictured, or, in other words, language signifies reality. Since language produces signification, it cannot itself be signified; instead language is a logical syntax for signifying reality. The Vienna Circle had been drawn to Wittgenstein's theory because it suggested that logic could still have a role to play in empiricism. Yet, when Carnap introduced "logical syntax," he described it as "a logical statement of formal rules." This is a departure from Wittgenstein's theory in that stating formal rules demands signifying logical syntax. Indeed, Carnap (1937, 1) introduces a "metalogic" for formally representing logical syntax, and for Wittgenstein, such meta-levels were not possible.

expressions (e.g. the sentences), but simply and solely to the kinds and order of the symbols from which the expressions are constructed.

In the months leading up to the proposal, Noam Chomsky (1957) had begun proposing his own rules of syntax, which he would eventually formalize into his seminal book *Syntactic Structures*. Alfred Tarski (1944) had been developing theories for determining the "truth" of any given sentence (or its formal semantics) by modeling the configuration of statements with formal logic. He proposed developing meta-languages, which would model "true" statements. For instance, one statement in a meta-language might be A satisfies (X and Y) if and only if ((A satisfies X) and (A satisfies Y)). Now, let's fill the variables, A, X, and Y with 'Max, 'brother,' and 'son' to form the everyday-language statement, "Max is both a brother and a son if and only if Max is a brother and Max is a son." With this as a model, when the statement "Max is both a brother and a son" is true, a model theorist, using principles of formal logic, can then deduce further statements to be true, false, or ambiguous. For instance, the statement "Max is a brother," would be logically true; the statement, "Max is either a brother or a son," would be false, and the statement, "Max is a brother and/or a son," would be ambiguous. Tarski's work would culminate with the emergence of model theory (which remains important to the design of the Semantic Web today).

Thus, at the time artificial intelligence was emerging as a discipline, the idea that language could be encoded via computerized algorithms was becoming increasingly viable. McCarthy et al. (1955) thus went on to suggest in their proposal one problem that advancements in artificial intelligence could solve:

## **2. How Can a Computer be Programmed to Use a Language It**

may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning and rules of conjecture. From this point of view, forming a generalization consists of admitting a new word and some rules whereby sentences containing it imply and are implied by others. This idea has never been very precisely formulated nor have examples been worked out.

This would give rise to the study of knowledge representation - a sub-field within AI. Yet, McCarthy and Minsky, two of the lead collaborators on the Dartmouth Project and often considered to be the "Fathers of AI," would eventually split considerably in their thinking about how best to approach representing knowledge to a computer.

In what follows, I introduce three oppositions that have ordered the way that semiotic technologists in the knowledge representation community tend to characterize their work: 1) whether knowledge should be encoded with depth or with breadth, 2) whether syntax should be encoded with neat logic or with scruffy procedures, and 3) whether semantics should be encoded to make knowledge systems complexly expressive or to make them able to produce results in a finite number of steps. Since the mid-1950s, artificial intelligence researchers have drawn on these oppositions to position themselves and their work within the knowledge representation community. Yet, as I will show in each sub-section, as these experts have attempted to make their semiotic infrastructures work in the "real world", they began encountering the limits of the semiotic systems used to describe their work. I argue that, in some cases, they responded by adopting a deconstructive practice - reversing the hierarchical oppositions that tended to privilege depth, neatness, and logic, and thus calling into question the stability of the semiotics used to represent their work.

### **3.1.1 Encoding Knowledge: Depth vs. Breadth**

Since the late-1960s, knowledge representation has been a significant goal in artificial intelligence research. Research in knowledge representation questions: How can we model the mind? How do we get machines to mimic the mind's recognition and understanding of everyday objects and events? How do we get machines to accurately understand information about a complex world? In order to build machines that can replicate human thinking, memory, and language, a great deal of research throughout the "golden years" of AI (1950s to 1970s) had been oriented towards understanding, logically and algorithmically, how these human processes work. At the time, defense organizations like DARPA had been funneling money

into the field, and as a result many digital systems were built to solve discrete problems. There was a great deal of optimism at the prospect of developing systems that could become just as intelligent as any human mind. Herbert Simon (1965) proclaimed, "machines will be capable, within twenty years, of doing any work a man can do." Marvin Minsky (1967, 2) proclaimed, "Within a generation [...], the problem of creating artificial intelligence will be substantially solved."

Almost paradoxically, knowledge representation researchers came to learn that the hardest challenges in their field were not getting computers to solve complex problems, but instead getting computers to solve very simple problems. Getting a computer to expertly play a game of chess turned out to be much easier than getting a computer to hold a coherent conversation for even just a few minutes.<sup>7</sup> Throughout the 1960s, a great deal of research in knowledge representation was thus directed towards getting computers to understand "common sense" knowledge and perform "common sense" tasks. Alan Newell and Herbert Simon (1961) worked towards building the General Problem Solver - a computer program that was designed to simulate human decision-making behaviors. Marvin Minsky, along with Seymour Papert, had attempted to tackle the challenges of teaching computers common sense knowledge through the construction of "micro-worlds" - or bounded environments with little complexity where computers could begin to learn. Upon mastering knowledge in a micro-world, computers could slowly be introduced to more complex thinking. These systems generated a great deal of excitement in the field and demonstrated the promise of knowledge representation at the time.

---

<sup>7</sup>Chess became an important metaphor for intelligence in AI research. As computing historian Nathan Ensmenger (2011, 9) writes:

Chess was, in this context, not just any game: the traditional province of kings (and scholars), chess had long been recognized as the pinnacle of human intellectual accomplishment, requiring both deliberate, carefully cultivated learning and strategy, as well as bold, creative, and courageous flights of inspired brilliance.

Thus, chess became a "model organism" for AI; it became AI's *Drosophila*. Yet, AI researchers would eventually come to lament that their systems could do little more than win a game of chess. Further, the way that their systems went about winning a game of chess was not very "intelligent." AI systems would use brute-force algorithms to calculate every potential move several steps ahead, whereas many expert chess players claimed to only think two to three moves ahead and instead relied on an ability to discern patterns. Elizabeth Wilson (2010) also notes that expert chess-playing computers lacked affect, which she argues to be just as significant to chess-playing as intellect.

Yet, the field of knowledge representation has since had a rocky history. In the late-1970s funding for AI began to dry up as researchers faced the fact that their systems, constrained to micro-worlds, could not solve more than very simple problems, such as picking up a block or moving it to a different room. This time became known as an "AI winter." Some of the most damning critiques of the AI agenda that emerged during this period focused specifically on issues of language. Herbert Dreyfus (1972), for instance, argued that computers will never be able to understand the meaning of statements because computers will never be able to discern the *contexts* in which statements are made - contexts that extended far beyond a micro-world.

In the early 1980s, AI began to make a come back with the design of "expert systems" that could solve complex problems in very narrow domains, or in other words, domains where the *contexts* of statements were less ambiguous. Yet, the challenge of getting computers to understand and produce *everyday* language continued to perplex knowledge representation researchers. In an interview, James Hendler described of the time:

One of the things that has always been a part of my work has been - there's kind of always a paradox in AI between what you might call depth and breadth, so the more narrow we make what we're trying to get the computer to know about, the better it does at knowing it. Of course, you end up with an idiot savant right? ... And sort of by the mid-80s people realized that computers are very powerful technology so you have expert systems and to this day [they] were sort of these narrow things.

Perhaps the most ambitious effort to encode common sense for computers was Cyc - a project led by Douglas Lenat and Ramanathan Guha from the mid-1980s to early 1990s (Lenat et al., 1985). One concept at a time, the project aimed to define and encode every piece of knowledge needed to write a one-volume encyclopedia. This included facts as mundane as: A thumb is connected to a hand, a bookshelf holds books, and to be on top of something means to be above it. The project

quickly became huge in scope; hundreds of thousands of concepts were defined for computer consumption. Today, many of my informants lament the effort that was put into this "ontology of everything," noting that it became so large that, while it aimed to codify simple reasoning, it became unusable to address simple tasks.

The history of knowledge representation is often described as a struggle to toggle between the breadth and depth of intelligence that can be encoded for a computer. While AI's knowledge representation systems became very good at answering complex questions in narrow domains, as soon as they were asked broader questions, the systems often broke down. Yet, knowledge representation systems designed to model broad, common-sense knowledge could never get enough knowledge to actually have common sense. Knowledge representation researchers working in this vein continuously found themselves dealing with the issue that in order to represent mundane knowledge to a computer, they would need to represent all of human culture - a challenge that was not only practically impossible, but with the computing power available at the time, also technically impossible. Dealing with this tension would lead many AI researchers to pose questions about the nature of knowledge that had troubled philosophers for centuries. Was knowledge based in universal logical truths, or was it based in our experiences? Could language be modeled with clean, logical formalisms, or was it inherently messy and illogical? It also led them to question how knowledge representation systems should be organized - or how the syntax of language should be encoded for computers.

### 3.1.2 Encoding Syntax: Logic vs. Procedures, Neat vs. Scruffy

"For there exists a great chasm between those, on the one side, who relate everything to a single central vision, one system more or less coherent or articulate, in terms of which they understand, think and feel-a single, universal, organizing principle in terms of which alone all that they are and say has significance-and, on the other side, those who pursue many ends, often unrelated and even contradictory, connected, if at all, only in some de facto way, for some psychological or physiological cause, related by no moral or esthetic principle."-Isaiah Berlin  
The Hedgehog and the Fox. (quoted in (Minsky, 1974))

In his paper, a "Framework for Representing Knowledge," Marvin Minsky (1974), often considered (alongside John McCarthy) to be the one of the fathers of artificial intelligence, introduced the concept of the frame. For Minsky (1974, 1), a frame is "a data structure for representing stereotyped information":

Here is the essence of the theory: When one encounters a new situation...one selects from memory a structure called a Frame. This is a remembered framework to be adapted to fit reality by changing details as necessary.

The "terminals" or slots of a frame have certain requirements that the values assigned to them must meet. In other words, in order to recognize and understand a situation, one must match the values of the situation to the terminals in various frames. Suppose, for instance, you enter a room and notice a sofa and a television. You may first try to assign these items into the terminals for a living room frame. But then you also notice bulletin boards, magazines, and a receptionist desk. These don't fit into the terminals for the living room frame so you select a new frame - a waiting room area.

Applying this to AI involved creating "frame languages" based on descriptions of objects, (rather than algorithms for how data should be manipulated). As a machine is exposed to a new object, it compares and tries to match the object to the description for frames it holds in memory. If the object's values can't be assigned to the terminals of a frame, the machine has to select other frames from memory or "de-bug" existing frames to create new ones. Frame languages provided an alternative to reasoning with first-order predicate logic. While first-order predicate logic presumed to know what was universally true about things in the world ahead of time, frame languages could approach new and mundane situations, draw on prior understandings of the world, and adapt. In other words, for frame languages *knowing how* to manipulate data was more important than *knowing what* was true about the data.

Concerned that knowledge representation techniques at the time could not model broad knowledge, Minsky critiqued his own micro-world approach as he introduced the frame concept in 1974. Minsky (1974, 74) argued that, while modeling

logic in a micro-world often produced favorable results, "...as we approach reality the obstacles become overwhelming." Minsky (1974, 78) also lamented that logic-based approaches to AI focused so intently on producing 'consistency and completeness'<sup>8</sup>:

I cannot state strongly enough my conviction that the preoccupation with Consistency, so valuable for Mathematical Logic, has been incredibly destructive to those working on models of mind. At the popular level it has produced a weird conception of the potential capabilities of machines in general. At the 'logical' level it has blocked efforts to represent ordinary knowledge, by presenting an unreachable image of a corpus of context-free 'truths' that can stand separately by themselves. This obsession has kept us from seeing that thinking begins with defective networks that are slowly (if ever) refined and updated.

Frames, on the other hand, focused, not on the internal structure of knowledge, but on how the mind came to recognize and structure external situations. While using first order logic to model knowledge assumed that concepts completely, consistently, and rationally followed the same set of rules, the architecture of the frame assumed that we can't possibly model this way since every situation is marked with a new set of components or conditions. Minsky argued that our assessments and expectations of any given situation can never be more than imperfect *approximations*; we can only adapt a concept or situation from frames we already possess - to what have already been exposed to, or to what we have already experienced. They will thus never be complete or consistent.

Minsky's frame-based approach to modeling knowledge along with others that were emerging in the early 1970s, such as Roger Schank and Robert Abelson's (1975) concept of scripts, represented what came to be known as a more 'procedural' approach to knowledge representation than the 'logician' or 'declarativist' approach,

---

<sup>8</sup>In mathematical theory, 'completeness' suggests a mathematical system where 'all true statements can be proven,' and 'consistency' suggests a mathematical system where 'no false statements can be proven'. While these tenets continue to be important in knowledge representation, Kurt Gödel's incompleteness theorems proved that no mathematical system exists where all true statements can be proven in 1931 (published (Gödel, 1962)).

which had been championed by researchers like John McCarthy and Patrick Hayes.<sup>9</sup> This binary opposition marked a schism in the knowledge representation community that emerged in the early 1970s. Logicians/declarativists tended to apply first order predicate logic to knowledge representation problems, whereas proceduralists tended to model knowledge representation problems by defining structures and then manipulating them with programming procedures.

Further debates over the value of consistency and the aesthetic of knowledge representation solutions marked another schism in the community - one that Roger Schank coined as the "neats vs. the scruffies" in the early 1970s (Crevier, 1994). According to those I've interviewed, neats, working in the tradition of John McCarthy and Nils Nilsson tended to seek formal, clean, consistent, and complete solutions to AI problems. They believed that the world should be modeled with neat and well-defined semantics to correctly characterize the internal workings of a system (see for example, (McCarthy and Hayes, 1969)). Scruffies, on the other hand, working in the tradition of researchers like Marvin Minsky and Roger Schank himself, asserted that the world was too messy to model formally or consistently; they tended to employ hacks (typically through programming procedures, since hacks are often illogical) in their work to get systems to perform. For them, building computer intelligence was less about modeling the internal workings of a system correctly and more about creating pragmatic structures and protocols for roughly but rigorously assembling AI.

These divisions were made more contentious as researchers on either side of the debates addressed the shortcomings of the approaches. John McCarthy and Patrick Hayes (1969), who may be characterized as the quintessential neats and logicians, described one such shortcoming to logic-based knowledge representation in the late 1960s. In what is now a classic AI paper defining the "Frame Problem" (separate from the frames that Minsky defines), they suggested the difficulty of using logic to model dynamic facts - or, in other words, facts that change over time. This problem led some logicians, like Drew McDermott (1987a, 116), to question the enduring

---

<sup>9</sup>The label "logicism" had also been used to describe the philosophical doctrines of thinkers like Gottlob Frege and Bertrand Russell - thinkers that positioned all mathematical truths as essentially logical (Hodes, 1984).

value of logical approaches:

At some point one has to ask, Why bother? When all the other boys and girls are out playing with their computers, why must the logicians stay indoors and practice finger exercises? Can we really believe that the insights gained will eventually allow logic to leapfrog other approaches to inference? It seems far more likely that logic will trail behind, struggling to stuff all sorts of inference patterns into its own view of the world, whether they fit or not.

Terry Winograd (1980), who had adopted a 'micro-worlds' approach to modeling knowledge in the early 1970s through a project called SHRDLU, also inclined towards a proceduralist and scruffy approach to knowledge representation in the mid to late 1970s after coming to terms with the inability of his micro-world system to scale to "real world" contexts. Describing the controversy between procedural and declarative epistemologies, Winograd (1975) argued that approaches to knowledge representation based solely in formal logic tended to separate "facts" from "processes" and ultimately went on to model facts as discrete units. Proceduralist approaches, on the other hand, were based more on "interactions" between data and the subroutines that manipulated them. Daniel Bobrow and Terry Winograd's (1977, 2) Knowledge Representation Language called KRL - "organized around conceptual entities with associated descriptions and procedures" - is considered one of the first knowledge representation languages to be based on Minsky's frames. Distinguishing this language from more logic-based approaches, Winograd (1980, 220) described, "There is a fundamental philosophical and mathematical difference between truth-based systems of logic, and process-based systems like KRL."

The distinction between declarative approaches based in formal logic and procedural approaches based in processes and programs emerged largely as knowledge representation researchers began to grapple with the inability of their systems to represent a breadth of knowledge outside of their engineered micro-worlds. Frame-based procedural approaches emerged in recognition that the context-free truths assumed in first-order predicate logic often did not hold outside of a cordoned micro-world.

Yet, proponents for logic-based declarativist approaches also argued that systems designed to manipulate knowledge without knowing anything about that knowledge also could never work outside of these carefully defined domains. As these debates advanced, the phrase the "real world" catachrestically came to stand in counter to micro-worlds, along with any other "world" engineered to test knowledge representation ideas. In the process, the "real world" came to signify messiness, paradox, and other conditions that opposed efforts to cleanly and consistently represent knowledge. As a result, knowledge representation experts began to question the mutual exclusivity of the binary descriptors they had been using to characterize their work - calling explicitly for systems that moved "towards a middle," incorporating elements of both declarativist and proceduralist approaches in order to overcome the limits posed at either end of the spectrum. For instance, Eugene Charniak (1981, 1083) introduced Frame-based Artificial Intelligence Language (FRAIL) that "use[d] both predicate calculus and frames." Drawing on the neat/scruffy distinction, Charniak (1986) went on to propose a "neat" theory for language interpretation. Winograd (1975) called for "steps towards a middle" and introduced a form of modular programming as a compromise.

Even Pat Hayes, often considered the archetypal "neat," eventually problematized the oppositions characterizing knowledge representation. For instance, in Hayes (1977) he critiqued the distinctions researchers like Winograd had drawn between proceduralist and declarativist approaches, arguing that the controversy was based in a false dichotomy. Logic, he argued, was not a programming language or even a style of programming. It was simply a set of ideas for justifying inferences. Proceduralists had been advocating for particular *data structures* and styles of programming, whereas the role of logic was to enrich these data structures and processes with interpretations of their meaning in the world (Hayes 1980). Perhaps this is why, during an October 2016 interview with me, he described it as a great compliment that Roger Schank once referred to him as a "neat with a scruffy heart."

For Pat Hayes (1979), getting systems to work in the "real world" meant that researchers needed to get over the debates around the "right" syntax. To overcome the problem of AI being full of "toy problems," knowledge representation

researchers would need to start developing mechanisms for formalizing how knowledge and meaning should be interpreted. Following this lead, in the late 1970s and early 1980s, more researchers began to position frames, first-order predicate logic, and other structures for modeling knowledge as providing a *syntax* for knowledge representation but lacking a *semantics* that would enable machines to *interpret its meaning*. In response, Hayes (1979, 244) went on to introduce a "naïve physics" that aimed to formalize a program for representing common-sense knowledge about the physical world.<sup>10</sup> He wrote of his approach: "The important point is that one *knows what it means*: that the formalism has a clear *interpretation* (I avoid the word 'semantics' deliberately)." As we will see in the next section, formal "semantics" too became a point of contention in the knowledge representation community.

### 3.1.3 Encoding Semantics: Expressivity vs. Tractability

In an effort to simulate human memory, Ross Quillian (1968) introduced a model for storing information about the world in computer memory. The model involved "recoding" the information found in a dictionary into a network of terms and their associated properties. Consider the example that Collins and Collins and Quillian (1969, 240) offered in their paper:

If what is stored with canary is "a yellow bird that can sing" then there is a pointer to bird, which is the category name or *superset* of canary, and pointers to two *properties*, that a canary is yellow and that it can sing. Information true of birds in general (such as that they can fly, and that they have wings and feathers) need not be stored with the memory node for each separate kind of bird. Instead, the fact that a canary can fly can be inferred by retrieving that a canary is a bird and that birds can fly. Since an ostrich cannot fly, we assume this information is stored as a property with the node for ostrich, just as is done in a dictionary, to preclude the inference that an ostrich can fly.

Quillian's model became known as a semantic network. In semantic networks,

---

<sup>10</sup>By "naïve physics," Hayes means to encode information that may be taken for granted about the physical world.

terms need not be completely and consistently defined in any one location. As terms are linked to other nodes in the network, properties are inherited through the link.

Ronald Brachman outlined the shortcomings of semantic networks in his (1978) dissertation, "A Structural Paradigm for Representing Knowledge." Brachman argued that semantic networks lacked consistency and precision in semantics, or, in other words, an explicit "epistemology" - defined by Brachman as "a set of primitive structures for encoding knowledge... and rules for combining those structures into well-formed representations of individuals and classes of individuals." Semantic networks could not distinguish between different types of relationships links may represent. For instance, what if another node in the network was "Serinus canaria domestica" - the scientific name for canary? The node for canary would likely be linked to this node, as well as the node for bird, but the relationships between these links would be different. Serinus canaria domestica would be linked to canary because they refer to the same real world thing, whereas serinus canaria domestica would be linked to bird because it was a subset of the category bird. With the former, both nodes should inherit each other's properties. With the latter, the term canary should inherit the properties of the category bird but the category bird should not inherit the properties of the term canary (since not all birds are yellow and not all birds can sing).

For Brachman, without a precise "epistemology," semantic networks may be able to express ideas about concepts, but they could not make inferences about the types of relationships between concepts. In other words, the semantic networks could describe the world, but they could not reason based on these descriptions. When asked discrete questions, these systems could not produce unambiguous answers.

Brachman aimed to design a system that could explain *how* sub-units composed wholes; *how* concepts were built up from representational units that related to each other. From this, Brachman introduced the structural inheritance network - a model for knowledge representation that would not only characterize the properties of objects, but would also establish what could be called a "neat" semantics for formalizing the relationships between objects. Perhaps, most notably, Brachman introduced the concept of "subsumption," or the ability to categorize nodes so that

a reasoner can eventually infer the relationships between concepts.

Other research emerging in the early 1980s was concerned with making object-oriented representation languages, such as frames, more "functional" (Levesque, 1984). In other words, this research aimed to design systems that could do more than simply "describe" knowledge through neat semantic structures; it also aimed to design systems that could make "assertions" based on these structures. Brachman and Levesque (1982) suggested that "competence" in knowledge representation required both "terminological adequacy" - where the structures of relations between terms could be characterized to a machine - and "assertional adequacy" - where the machine could interpret these structures in order to make assertions about what is known and what is not known in a knowledge base. However, frames did not make a distinction between these two forms of adequacy; according to Brachman et al. (1983), in many frame-based representation systems, it wasn't clear whether the frame was describing a concept or asserting the existence of a concept. They thus began to conceptualize systems that would differentiate between the descriptive and the assertional components of the language; the descriptive components were based on frames, while the assertional components were based on elements of first-order logic. Both subsumption and the distinction between descriptive and assertional components oriented the design of KL-ONE, a frame-based knowledge representation system that would become the primary model for description logics (DL) in the early 1990s (Brachman and Schmolze, 1985).

Yet, research in knowledge representation began to characterize tradeoffs to combining descriptive and assertional components into systems like KL-ONE. Brachman and colleagues aimed to build systems that could at once provide rich, or "expressive" descriptions of concepts, but could also make inferences about the relationships between concepts within a finite number of computational steps (or in other words, to build systems that were *tractable*). However, the more expressive knowledge representation languages are, or the more complexity they attempt to model, the more challenging it is to guarantee the system will be sound, complete, and capable of producing tractable results (Patel-Schneider, 1985; Levesque and Brachman, 1987). This problem has perplexed logicians for almost a century - since

Kurt Gödel's incompleteness theorem proved that no formal system of axioms is capable of proving all truths. Thus in order to guarantee completeness - to ensure the system could always produce results (or were *tractable*), - KL-ONE-like systems often restricted the number of constructors, or "primitives," of the language to a small set. It offered a substantially limited logic, in turn limiting the extent of "real world" complexity the system could represent. Thus, throughout the 1980s, a great deal of attention was devoted to "cleaning" semantics in object-oriented systems, advancing the inferential capabilities of these systems, and theorizing the computational limits of the work. In the late 1980s, this research merged into the DL framework, also known by some as terminological logics.

Throughout the 1990s, researchers went on to experiment with how far they could push the tradeoffs between expressivity and tractability in DL-based systems. Systems like CLASSIC (Borgida et al., 1989) considerably limited the expressivity of the language in order to ensure the system could produce inferences in a reasonable amount of time (Brachman et al., 1991). Yet, the designers of CLASSIC also recognized that they would need to make concessions to this limited logic so that users in the "real world" could take advantage of them. Deb McGuinness, who worked on CLASSIC in the early 1990s, described in a 2015 interview how providing such concessions made CLASSIC distinct from systems based on "pure theory" (or, in other words, systems that could always guarantee to produce results):

The description logics have a benefit because they're limited; they're computationally tractable, but it's almost impossible to use a limited language and meet all of your needs. So Classic had some extensions. [...] they had concessions to working in the real world. So we had these extra things that broke some of the nice properties [of] the theoreticians. In this part I wouldn't be considered a theoretician because I was arguing for the usability features of it.

Similarly, systems like LOOM "conced[ed] that it might be acceptable to deliberately build an incomplete system" (MacGregor, 1991, 90).

Notably, right around the same time Ronald Brachman was completing his dissertation, Pat Hayes had foreshadowed some of the "real world" challenges the

expressivity/tractability tradeoff would pose. Much like Brachman, Hayes was concerned with formalizing how a computer would interpret meaning from knowledge with greater precision in his 1979 "naïve physics manifesto." However, for Hayes, limiting logic and restricting domains in order to make programs tractable would repeat the same mistakes those working to develop 'micro-worlds' in the 1960s had made. Hayes argued that it was wrong to assume that combining a bunch of programs that work well in small worlds will produce a system that works well in large worlds. In other words, limiting logic and restricting domains would inhibit efforts towards common-sense reasoning or towards encoding a breadth of knowledge. He emphasized, "We are never going to get an adequate formalization of common-sense by making short forays into small areas, no matter how many of them we make" (Hayes, 1979, 265). As I will show next, Hayes's manifesto would provoke many in the knowledge representation community to confront the overlapping and at times contradicting implications of ordering knowledge representation thinking and work according to oppositions such as breadth vs. depth, logic vs. procedures, and expressivity vs. tractability.

### 3.1.4 Battle Grounds

Many of the divisions in the knowledge representation community converged with the publication of a 1987<sup>11</sup> issue of the *Computational Intelligence* journal. In it, Drew McDermott (1987b) published a paper he called "A Critique of Pure Reason." The paper was based on a talk that he had delivered two years earlier at the AI Society of New England Meeting. The paper aimed to tackle the "logicist" agenda put out by John McCarthy and Patrick Hayes. It was accompanied with 27 commentaries responding to his "critique."

The title of McDermott's paper, of course, was a reference to philosopher Immanuel Kant's book published under the same name in 1781. Although Kant was not cited anywhere in the paper, the influence of Kant's thinking and work on the argument made in the paper is evident. Following Kant, McDermott aimed to critique assumptions about truth and deduction that had been guiding prominent work

---

<sup>11</sup>1987 is often cited as the start of the second AI winter – a time when funding was again drying up as researchers were unable to put forward examples of successful systems.

in computer science and knowledge representation for decades - most specifically, that truth could be formalized prior to its use or prior to experience of it. McDermott (1987b, 151) summarized (in order to eventually critique) what he called the "logician argument":

The next step is to argue that we can and should write down the knowledge that programs must have before we write the programs themselves. We know what this knowledge is; it's what everybody knows, about physics, about time and space, about human relationships and behavior. If we attempt to write the programs first, experience shows that the knowledge will be shortchanged. The tendency will be to oversimplify what people actually know in order to get a program that works.

He went on to argue:

There is an unspoken premise in the argument that a significant amount of thought is deductive. Without this premise, the idea that you can write down what people know without regard for how they will use this knowledge is without foundation.

For McDermott, the logicist agenda starts with the premise that knowledge should be formally represented in a computer before the computer learns how to manipulate knowledge. McDermott traced the lineage of this argument through several researchers - John McCarthy, Robert Moore, James Allen, Jerry Hobbes, and himself. However, in his paper, Patrick Hayes was positioned as the archetype of the logicist camp, and his naïve physics manifesto was the subject of the most scrutiny. According to (McDermott, 1987b, 151), "Hayes is simply the most eloquent."

"Eloquent," I believe, is a fitting description of Hayes. Hayes is eloquent in both writing and speaking. While writing of logic and formalisms with precision and rigor, Hayes often writes quite poetically with drawn out metaphors. Hayes does not hesitate to jump into the heat of an argument over logic. He told me in an October 2016 interview, "Oh it's great winning intellectual battles. That's what life is all about."

"Battle" has been one of the more prominent metaphors that have emerged in my fieldwork on the history of knowledge representation. In Hayes's (1987, 183) response to McDermott, titled, "A Critique of Pure Treason," he wrote:

In the distant past, AI went through a bitter civil war between two camps, the Union of "neat logicians" and the "scruffy proceduralist" rebels. ... During the heat of the battle, the term "logic" came to have the emotive force of a war-cry for the neat faction, including myself, and was spoken by the proceduralists in the same sort of way that right-wing Republicans now speak of Communism, as an evil, cancerous blight whose spread should be resisted at all costs (Minsky 1975). This is now over, although the landscape bears scars, and it would be a shame if this paper were to start it up again. Unfortunately, terms such as "logician" tend to blur hard-won distinctions.

Yet, in the same issue, Ronald Brachman (1987), who had also been championing neater semantics for the past decade, lamented that the field had taken a "turn for the neat" - a "Great Formality Shift" that Hayes's manifesto had heralded. For Brachman (1987, 169), the war was not over, and the distinctions were not hard-won. Logicians were still shouting battle cries:

Back in the early days of the Shift, the impetus was (I think) to eliminate sloppy, inconsistent, and meaningless informal work by encouraging analysis in terms of a precise formal system. Yet, merely gentle suggestions that the tools of mathematical logic could be of use in AI ... have given way to dogmatic proclamations from the heart of the new Logician camp: "There is only one language suitable for representing information - whether declarativist or procedural - and that is first-order predicate logic" (Kowalski 1980).

He went on to characterize what he called the "myth of one true logic," or Logicism "with a capital L." For Brachman, the myth of one true logic had become a form of "imperialism." Papers were not being accepted to conferences because

they weren't formal enough. "Hackers" were being shunned from the field. Systems based on anything but first order logic were not considered legitimate.

Brachman's assessment of the time was later captured in historical accounts of the field, which also drew on "battle" metaphors to characterize the divide. In what is now the leading textbook on artificial intelligence, Stuart Russell and Peter Norvig (1995, 25), marked modern AI methodologies as leaning towards neatness:

Some have characterized this change as a victory of the neats - those who think that AI theories should be grounded in mathematical rigor - over the scruffies - those who would rather try out lots of ideas, write some programs, and then assess what seems to be working.

A few years further on, Pamela McCorduck (2004, 487) suggested, "As I write, AI enjoys a Neat hegemony, people who believe that machine intelligence, at least, is best expressed in logical, even mathematical terms."

The endurance of the "battle" metaphor as a dominant narrative of the history of knowledge representation suggests that the field, as a whole, never quite made steps "towards a middle." Approaches to knowledge representation were persistently pitted against one another - seen as "vs.," not as "and" or "neither." Yet, three decades of effort in knowledge representation had shown that, at the ends of any spectrum, knowledge representation systems hit their limits. Systems encoding deep knowledge could not work outside of their domains, and systems encoding broad knowledge could never capture enough common sense knowledge to reason broadly. First-order logic could not handle diverse contexts, yet procedural programs did not actually know anything about the world. Expressive programs could not reason, and tractable programs were too limited to capture the world's complexity. Semiotic technologists had come to understand their work along battle lines, but following approaches along either side of the oppositions could never adequately represent complex knowledge. This contradiction would become even more intense as the description logic framework "emerge[d] from ivory towers" into World Wide Web usage (McGuinness, 2001, 64).

## 3.2 Knowledge Representation on the Semantic Web

While working at CERN as a contract programmer in 1989, Tim Berners-Lee submitted a proposal to his boss Mike Sendall for an information system that would help organize information distributed across the multi-national laboratory without requiring researchers to agree on a standard technology or information model (Berners-Lee, 1999). This system evolved into what we today call the World Wide Web (WWW).

Today, the WWW is understood as an information space where documents, or Web pages, can be referenced and linked to from other documents using hypertext links. However, the system Berners-Lee described in his proposal aimed to do more than link documents; it aimed to link data, or specific nodes on Web pages. To illustrate the system in the proposal, Berners-Lee described a complex information system diagramed with a series of nodes and arrows. Nodes represent people, software models, groups of people, concepts, documents, etc, and arrows describe the relationships between nodes - for instance, that node A depends on node B or that node A is a part of node B. In designing a system that could link nodes and describe their relationships, it would not only be possible to organize information in more intuitive ways, it would also be possible to perform data analysis on the system, determining, for instance, when groups of people had few ties or when certain software had several dependencies. Importantly, Berners-Lee noted in the proposal, "Ideally, [each node] represents or describes one particular person or object."

However, the Web did not evolve this way. After releasing the WWW to the public in 1991, it began to proliferate rapidly and evolved towards a document-based Web. Rather than hypertext linking data points, the document-based web linked Web documents that each described many objects and people. This made it difficult to perform data analysis on Web data; computers could not distinguish between data points within Web documents and how they related to documents from which they were linked. Thus, in his keynote speech at the first International WWW conference in 1994, Berners-Lee lamented that Web documents were "flat" and "devoid of meaning" when in fact they "describe real objects and imaginary concepts." He called for adding "semantics to the Web" so that information within

documents could become machine-readable, and relationships could be described between information. This was the first public reference to what today is called the *Semantic Web*.

As the WWW has developed into one of the world's largest information repositories, there have been notable efforts to structure Web data with semantics - to design the technologies and standards to enable a Semantic Web. To enable a Semantic Web, webmasters give data points on Web pages unique identifiers and metadata and link them with other data points on other Web pages. Using terms defined in openly available ontologies to describe the links between data points, computers can logically model and reason with web data. In what follows, I will show how the battles that had ordered the way knowledge representation researchers thought about and approached their work in the 1980s continued to play out into the design of the Semantic Web in the 1990s and beyond.

### 3.2.1 A Syntax for Web Knowledge

When Ora Lassila, a computer scientist and research fellow at Nokia Labs, came to MIT as a visiting scholar in 1996, Tim Berners-Lee asked him what he believed was wrong with the Web. Lassila replied that he would like the Web to be able to do more things without human intervention; Berners-Lee agreed.

The origins of the Semantic Web are often traced back to the publication of the *Scientific American* article "The Semantic Web" in 2001 (Berners-Lee et al., 2001), but the work to formalize standards and protocols for applying knowledge representation on the Web had begun in the mid-1990s - with the development of the Resource Description Framework (RDF). RDF offers syntax for enabling computers to interpret Web content. Based on RDF, Web data gets organized into 'triples,' consisting of a subject, a predicate, and an object. The subject is a piece of data on the Web that has a universal identifier; the object can either be another piece of data on the Web that has a universal identifier or a reference to something that exists outside of the Web. The predicate describes the relationship between the subject and the object. For instance, one triple may be John Doe (<http://johndoe>) is a Person. Here John Doe would be a piece of data on the Web; 'person' would refer to

an entity that exists outside of the Web, and 'is a' would describe the relationship between the two. Another triple may be John Doe (<http://johndoe>) is married to Jane Doe (<http://janedoe>).<sup>12</sup>

Work to formalize RDF into a Web standard through the W3C began in 1997, and the standard was formalized into a recommendation in 1998. The group of practitioners that participated in the formalization of the W3C RDF standard had diverse backgrounds and agendas. Some represented major search engines, aiming to build tools that could enable machines to better interpret the content of a webpage to improve search. Some came from industry, seeing an economic benefit to enabling smarter browsing. Another group of practitioners came from the knowledge representation community, excited by the idea of building a worldwide knowledge representation system. On the working group many argued that RDF had to be simple - like a basic programming language with which more complicated programs could be written. Others did not believe RDF should be treated like a programming language - that knowledge representation was not necessarily reducible to simpler forms. The recommendation emerged as a compromise between these viewpoints.

However, many individuals in the knowledge representation community still considered RDF too messy and inconsistent to enable knowledge representation. In fact, without provocation or prior introduction, Pat Hayes, who was still revered in the knowledge representation community, emailed the working group chairs to complain about the recommendation's sloppiness. He was later asked to join the working group. Reflecting earlier concerns surrounding the limitations of frame-based systems based solely in procedural methodologies, RDF provided syntax for knowledge representation, but lacked a neat semantics (Lassila and McGuinness,

---

<sup>12</sup>RDF emerged from a lineage of work that aimed develop standard formats for making meta-data, or descriptions of data, machine-readable. An early predecessor to RDF was the Platform for Internet Content Selection (PICS) - a system that allowed users to associate metadata with Web content designed initially to help parents control what children could access. Perhaps the most notable pre-cursor to RDF was the Meta-Content Framework (MCF). Ramanathan Guha began developing MCF at Apple in the mid-1990s. Guha had long been involved in knowledge representation work, playing a key role in the Cyc project. MCF aimed to structure metadata so that information could flow between software products with different information models and data structures. When Guha left Apple for Netscape in 1997, he met Tim Bray, who had developed the first version of the W3C's XML specification - a standard for storing and transporting data on the Web. Together, they adapted MCF using XML, and this project became the basis for the Resource Description Framework (RDF).

2001). This played an important role in the development of OWL.

### 3.2.2 A Semantics for Web Knowledge

In the late 1990s, the U.S. Defense Advanced Research Projects Agency (DARPA) was becoming interested in designing systems that could control and coordinate autonomous software agents. One vein of this research involved designing languages that could identify, understand, and integrate information sources across distributed agents through "semantics." The DARPA Agent Markup Language (DAML) Program was introduced in 2000, and James Hendler, then a computer science professor at the University of Maryland, was appointed as the program manager.<sup>13</sup> When Hendler got to DARPA, he was given funding to distribute to three laboratories that would build out proofs of concept for DAML. One funded group was the Knowledge Systems Laboratory at Stanford University - where McGuinness had moved after being quite active in the description logic community throughout the 1990s (McGuinness and Wright, 1998). Another funded group was the Massachusetts Institute of Technology Laboratory for Computer Science and W3C, where Berners-

---

<sup>13</sup>At the time of his appointment, Hendler, along with colleagues Jeff Heflin and Sean Luke, had been working to develop a system that could "mark up" Web pages with semantics. SHOE (Simple HTML Ontology Extensions) enabled website designers to reference an ontology where a series of terms were defined and then annotate data on their HTML documents with these terms. The system would enable search engines to "look up" the meaning of annotated content on a Web page by referencing the ontology. Placing this project in the context of research being conducted in the field of knowledge representation (citing systems like KL-ONE, CLASSIC, and LOOM), Heflin et al. (1999, 2-3) argued that applying a knowledge representation language on the Web would require considerably rethinking what knowledge bases could enable:

Web systems simply cannot assume that all of the information has been entered solely under a knowledge engineer's watchful eye, and is therefore correct and consistent. As authority on the Internet is distributed, it cannot and does not make any such promise. This lack of central control leads to a number of serious problems. Since there is often no editorial review or quality control of Web information, each page's reliability must be questioned. Since a web page that was useful one day can disappear the next, there is no guarantee on the availability of information. Since there are no integrity constraints on the Web, information from different sources can be in disagreement, leading to inconsistency. Some inconsistencies may be due to error, others due to philosophical differences. In addition, there are quite a number of well-known web hoaxes where information was published on the Web with the intent to amuse or mislead - the computational agent typically cannot tell the difference!

The guiding design principle for SHOE became "a little semantics goes a long way." Rather than attempting to develop concepts to represent every difference, they would develop just a few concepts that could be useful to represent most (but not all) differences.

Lee and others were advancing research towards what they were calling the "Semantic Web." Several folks that had long been involved in the knowledge representation community also became involved in the project, including Drew McDermott, Pat Hayes, and Peter Patel-Schneider. Together, these groups and other collaborators formalized DAML into a language in October 2000. Soon after, they joined efforts with other researchers that had been designing the Ontology Interface Layer (OIL) in the European Union. DAML+OIL was released in January 2001.

The W3C working group for the Web Ontology Language (OWL) began in 2001, tasked with moving DAML to a Web standard. With over 40 participants, the committee was very large in comparison to other W3C committees. Consisting primarily of academic researchers, there were not so much different stakeholder positions on the working group, but instead schisms in what members believed the ontology should do and how it should be structured. In an interview I conducted with Pat Hayes, he described:

The RDF working group ... had sort of a collegiate atmosphere and eventually ... after some initial fights, we got each other calibrated, and were able to sort of work together. Occasionally people would get exasperated with one another, but that sort of thing happens. The OWL working group was much more of a real genuine schism. It had an on-going fight, which was still ongoing, between two very different points of view about what OWL should be.

Some of the early controversies in formalizing the mark-up language into a Web standard were identified in a 'Trends and Controversies' section of the March/April 2002 issue of the IEEE Intelligent Systems journal (Harmelen et al., 2002). In the section, eight participants on the working group - each from different backgrounds - were asked to comment on the process of defining a standard ontology language for the Web. Frank van Harmelen (2002, 72), a knowledge representation and reasoning professor at Vrije Universiteit Amsterdam, suggested that the styles of frame-based modeling that he had seen used in both academia and industry "directly conflict with the DL [description logic] style of DAML+OIL." He called on the working

group to reduce the complexity of the ontology's formal semantics. Guus Schreiber, also a computer science professor at Vrije Universiteit Amsterdam and co-chair of the OWL working group, wrote:

The debate about what a Web ontology language should look like is reminiscent of past neat-scruffy struggles. Knowledge modelers want expressiveness, logicians stress decidability [or tractability]. The main difference is that the Semantic Web actually forces us to make some choices: there is a strong need for real-world knowledge representation. (Harmelen et al., 2002, 78)

Indeed, many of the arguments that arose in the working group reflected earlier neat-scruffy struggles. As James Hendler in his role as chair of the W3C WebOnt Working Group (2002a), noted in one email to the group:

In fact, if one goes back to a famous talk given by Roger Schank...he referred to the neats and the scruffies and at that time, this particular debate we're still having today was one of the examples used as a differentiator! ...the issue being discussed here gets so much to the heart of things and the differences in how frame and DL folks see the world that it's hard to even know where to begin.

There was considerable concern held by some from the description logic (DL) community that designing first-order logic on top of RDF syntax would prevent the ontology from producing decidable (or tractable) results.<sup>14</sup> Pat Hayes went on to propose a proof that this was not a paradox as long as the logic gave up decidability.

Another issue that arose in the working group concerned the type of logic on which the ontology should be based. Should the ontology be based on DLs - first-order logics with considerably limited expressivity that could ensure the system could produce results in a reasonable amount of time - or should the ontology be based on full logic where any RDF statements could be combined in any way to

---

<sup>14</sup>Peter Patel-Schneider, a particularly strong proponent for moving away from RDF syntax, acknowledged that scruffier approaches could overcome this issue, but "it would be a gigantic mess, at best" (W3C WebOnt Working Group, 2001).

build complex statements without guarantees that there would be soundness and completeness?<sup>15</sup> Finally, there was concern about whether the system would be usable by everyday webmasters.

While early efforts attempted to sort out these controversies, eventually the working group decided to propose three versions of OWL - one emphasizing simplicity and usability (OWL-Lite), one emphasizing decidability (OWL-DL), and one emphasizing expressivity (OWL-Full) (Horrocks, Patel-Schneider, and van Harmelen 2003). OWL-Lite and OWL-DL were limited and thus could only encode knowledge in restricted, or deep domains (but could do so particularly well). OWL-Full, on the other hand, could encode knowledge in *broad* domains but could not necessarily reason with the knowledge. Ora Lassila told me in an interview, "I think the distinction between these two [OWL-DL and OWL-Full] represents the rift that exists between the knowledge representation community, or it did at the time anyway." In an email thread with many working group members throwing around ideas for naming each version, Frank van Harmelen offered one possibility:

OWL Light: ?

wimpy

OWL fast: OWL/FOL-style

Neat

OWL large: OWL/RDF-style

Scruffy(W3C WebOnt Working Group, 2002b)

Recommending three versions of OWL signified that the battle had hit a stalemate. OWL did not signify compromise, or a "move to the middle," but instead demonstrated how a semiotic infrastructure can order knowledge in different ways (and thus have different "real world" effects) when designed according to different language ideologies.

---

<sup>15</sup>Jim Hendler described to me that this is an issue about whether infinity is permitted in the logic. "Logic," he argued, "loves infinity." However, infinity is not decidable and will break a DL reasoner.

Many Semantic Web practitioners have described to me that recommending three versions of OWL was extremely confusing to webmasters. They often cite this as the reason that the Semantic Web did not take off in the early-2000s like the Web had in the early-1990s. The working group had been tasked with coming up with the Web standard - one recommendation. Why did they produce three recommendations? Was it because they couldn't come to consensus on a particular worldview - because the "data friction" (Edwards, 2010) had been so powerful that they were ground to a halt?

That may have been part of the reason, but I believe there is more to this battle that has now dragged on for more than half a century. "Intelligence" is not just broad, just deep, or somewhere in between. It has all of these characteristics. Intelligence is neat, scruffy, and everything in between. It is logical and can also understand things that are much more complex than what can be modeled with first-order logic. To encode human language, sacrificing any of these human traits is not an option. Knowledge representation experts thus face a paradox - that to build intelligent systems they would need to build semiotic infrastructures that are broad, deep, neat, scruffy, tractable, expressive, and everything in between, but in the "real world," they have ordered their thinking and work in such a way where these necessities contradict each other. They cannot move "towards a middle" because losing ground on one end of the spectrum invalidates the entire project. It's not just a battle between people, then; it's also a battle between conflicting demands. To think, work, and appropriately represent knowledge in spite of conflicting demands, they need learn to creatively endure double bind (which will be the subject of the next chapter). Perhaps this is why Ronald Brachman (1987, 171-172) closed his response to Drew McDermott's *Critique of Pure Reason*:

Thus, what is clear is that we need both. ...Yes, our progress is slow, but why should we have assumed that we could "solve" intelligence overnight? I don't think our progress is slow because the basic argument of logicism is flawed (and McDermott certainly hasn't *proven* that it is), but rather because the enterprise is just very hard. Perhaps we've all been a little too naïve about it.

It is this contradiction - that semiotic technologists understand their work as needing to simultaneously advance all aspects of system organized as a tradeoff - that makes knowledge representation work so hard to represent in the "real world."

### 3.3 Knowledge in the "Real World"

Towards the middle of Peter Norvig's keynote at the WWW2016 conference, he put up a slide that presented two books - the first Ronald Brachman and Hector Levesque's *Knowledge Representation and Reasoning*, the second Vinit Nayak's *Copying and Pasting From Stack Overflow*. With a click of the mouse, a large red prohibition symbol displayed over Brachman and Levesque's book. For Norvig, meaning did not come from highly trained logicians carefully building out ontologies but instead from scruffy hackers, marking up their data with available syntax. His talk did not reflect a "victory of the neats" or a "neat hegemony," as the field had been characterized (in his book) a decade before. Instead for Norvig, there was too much ambiguity on the Web. He noted that when Google's knowledge graph was asked the question "How tall is George Washington?" it returned "22.9cm," because George Washington can refer to the historical figure, or it can refer to the book titled *George Washington*. The knowledge graph returned results about the book.

The Web, as an information infrastructure, has come to be seen as a "real world" context for testing ideas in knowledge representation. People from all over the world contribute to the Web, so its data is scruffy, inconsistent, and ambiguous. In the next chapter, I will address how working in this "real world" context has prompted members of the knowledge representation community to learn to represent knowledge (and their work) in less formal and more experimental ways. I will describe how and why many have come to understand their work as having become more flexible and scruffy over time.

Notably, this retooling of the semiotics used to describe their work has been extended to how they think and talk about the "real world." At the end of Norvig's talk, an audience member asked him at what point logicians should be re-enlisted in the knowledge representation project. Norvig replied that he believed the Semantic Web of the future would open a job market for philosophers. He went on to say,

"Maybe the real world is less real than we think [...] It's not so much about ontology. It's more about epistemology."

For the past half a century, the way that the "real world" has been used in the knowledge representation community is akin to the way that "parasite" has been discussed in the philosophy of language. In J.L. Austin's (1975) speech act theory, a "parasite" signifies speech that deviates from "normal use" - speech that is non-serious, strictly performative, or mocking. These non-standard forms of speech can only exist in contrast to standard forms of speech, so while they infect them, they are also dependent on them. Concurring with Austin, John Searle (1977, 204-205) commented that, "Furthermore, in a perfectly straightforward sense such utterances are 'parasitical' on the standard cases: there could not, for example, be promises made by actors in a play if there were not the possibility of promises made in real life."

Notably, for semiotic technologists, "the real world" has come to displace the "parasite" in this binary opposition; it is in the "real world" that speech deviates from norms, and this creates a new paradox. Since the mid-1950s, knowledge representation experts have both attempted to encode the "real world" - to encode how human knowledge really works - and attempted to encode in spite of the "real world," which seems to persistently get in the way of their task. The concept of the "real world" too then provokes double bind. How can semiotic technologists encode the reality of knowledge when the reality of knowledge is what is preventing them from encoding?

Part of the paradox arises from the construction of the battle lines - neat lines along which parasites are always infecting. Accordingly, Derrida (1988, 89-90) - who may be considered the "scruffy" to Austin and Searle's "neat" - responded to Austin and Searle by asking where the line gets drawn between real life and non-real life:

...even in what Sarl calls 'real life,' that 'real life' about which Sarl is so certain, so inimitably (almost, not quite) confident of knowing what it is, where it begins and where it ends; as though the meaning of these words ('real life') could immediately be a subject of unanimity, without the slightest risk of parasitism; as though literature, theater, deceit,

infidelity, hypocrisy, infelicity, parasitism, and the simulation of real life were not part of real life!

Norvig's suggestion that the "real world" may be "less real than we think" makes a similar case about the distinguishing between the real and the non-real (or the real world and a micro-world) - both at an epistemological level and an ontological level. At the epistemological level, he deconstructs the phrase "the real world" as a phrase that cannot be defined completely, consistently, or "neatly" against the "non-real world," the "micro world" or the "engineered world." In doing so, at the ontological level, he calls into question the extent to which it is "real world" conditions that are getting in the way of encoding. Norvig's response suggests a deconstruction of the functional semiotics that has ordered semiotic technologists' work. Epistemology may not be so separate from ontology or the design of ontologies. The "real world" may not be so separate from an "artificial world."<sup>16</sup>

This deconstruction of the phrase "the real world" (both at the epistemological and ontological level) has several important implications for how semiotic technologists think about their work. First, it casts doubt on any remaining justifications that knowledge is simply "out there" - that there are context-free truths waiting to be encoded. It acknowledges that semiotic technologists enact real worlds through their practical modes of representation - through their decisions about which differences should make a difference in their ontologies.

But perhaps even more notably, this deconstruction puts the "reality" that keeps getting in the way of encoding on trial. It acknowledges that the oppositions by which semiotic technologists order their work (depth vs. breadth, knowing what vs. knowing how, and expressivity vs. tractability) are also encoded through practical modes of representation. Semiotic technologists decide that there are differences that make a difference between these concepts as they use them to describe their work and to position themselves against each other. In this sense, semiotic technologists, through their language and their practices, play a part in enacting the "reality" that

---

<sup>16</sup>Reviewing Derrida, Spivak (1980, 38) remarks that the iterability of marks create the conditions in which the binary between real life and non-real life can be deconstructed and that such deconstructions demand a "revolutionary change of mind." I interpret Norvig's quote to signify such a revolutionary change of mind.

provokes double bind in their work. I argue this not to suggest that the limits don't exist - that the double binds are artificial; they may just be more (and less) *real* than we think.

## 4. PURSUING THE LIMITS OF KNOWLEDGE REPRESENTATION ON THE SEMANTIC WEB

In<sup>1</sup> October 2016, Pat Hayes and I sat down for an interview to discuss his involvement in the design and development of the Semantic Web. After describing the figures that had most influenced his thinking in AI, as well as how he had perceived his role in the knowledge representation community, he went on to discuss how "battles" (see Chapter 3) played out in the design of standards for the Semantic Web. Hayes noted:

The thing is it used to be considered sort of almost an intellectual war in AI. And now I think it's settled down into...we all agree *it's a matter of style...* almost a personality rather than a content. But certainly both styles are clearly visible in the way things happen. And released a friction on committees of course. (emphasis mine)

The way that Hayes' uses the concept "style" here is akin to Ludwig Fleck's (1981) "thought style," with 'neat' and 'scruffy' used to construct and label particular "thought communities." I've come to understand the development and evolution of the Semantic Web as oriented according to how researchers have crafted their relationships to neater or scruffier styles. Semantic Web researchers and practitioners use the classic distinctions to position themselves against one another, telling me so and so is a neat, or so and so is a scruffy. Yet, today, while such binaries are useful for characterizing the "style" or "personality" of Semantic Web work, the researchers and practitioners I've interviewed are also reflective in using them, almost always noting that their characterization is an over-simplification and that they don't know any one researcher that is perfectly (or "neatly") neat or perfectly (or "neatly") scruffy.

---

<sup>1</sup>Portions of this chapter previously appeared as: Poirier, Lindsay. 2107. A Turn for the Scruffy: An Ethnographic Study of Semantic Web Architecture. In Proceedings of WebSci'17. Troy, NY: ACM. <https://doi.org/10.1145/3091478.3091505>.

Historically, communities seeking to advance formal logic have attempted to dispel "style" from logical languages. Reviewing how early logicians conceptualized the marriage of logic and language, feminist philosopher Nye (1990, 154) writes:

A logical language is a *language without style*, a language purged of the coloring, nuance, rhythm, metaphor, rhetoric that mark an individual voice. These effects, characteristic of poetic or literary language, produce a series of private idioms particular to each speaker, and must not be allowed to interrupt the commonality of a language in which truth can be expressed. (emphasis mine)

Yet, Hayes (who may be considered an archetype for the logicist camp in AI) went on to tell me that, in the working group for the Web Ontology Language (OWL), he surprisingly played the role of the scruffy. In doing so, Hayes acknowledged not only that style was an important part of knowledge representation, but also that "neatness" - what may have historically been considered the "style of no style"<sup>2</sup> - deserved to be called into question. Up until the early 2000s, neatness had served as a privileged signifier in the functional semiotics ordering knowledge representation work (similar to the way that rationality tends to be privileged over irrationality in techno-scientific discourses). This chapter examines how semiotic technologists learned to deconstruct the oppositions that had ordered their thinking and work as they attempted to encode meaning in the "real world" - a real world that resists being reduced to an all-encompassing set of categories and logics.

Continuing my interview with Hayes, I asked him if his role as a scruffy was comfortable for him. He replied, "It was then because I enjoyed making my fellow neats squirm. I think actually in Roger's terminology, I started off as a neat, and I've gotten scruffier as I get older." I told him that he's not the first person I've interviewed to tell me that. He laughed. "Yeah, well...your youthful idealism gets worn off by life when you get to my age."

---

<sup>2</sup>This is akin to what Sharon Traweek (1992, 162) refers to as the "culture of no culture." In examining the culture of high-energy physics, Traweek described how physicists "long[ed] passionately for a world without loose ends, without temperament, gender, nationalism, or other sources of disorder." "Neats" similarly exhibit a thought style that longs for a world without disorderly styles.

This chapter examines how the culture and the style of knowledge representation has shifted and iterated (parallel to the way that meaning and semantics shift and iterate) as semiotic technologists have worked to encode scruffy knowledge. I argue that many Semantic Web designers have come to see paradox, contradictions, and incompleteness as an inevitable component of "real world" language. While historically approaches to knowledge representation have sought to restrict the design of semiotic infrastructures to prevent these inconsistencies from emerging, more recent efforts in the Semantic Web community have experimented with designing more open systems - systems tolerant of paradox and drift. In doing so, Semantic Web practitioners have approached their work with a trickier expertise - an expertise in encoding meaning that is perpetually taking on new forms. This, I argue, has not only changed the way that knowledge gets represented in digital systems; it also has also changed the way that Semantic Web designers perceive how language and logic operates and what it means to think, work, and communicate in the "real world."

## 4.1 Genealogies of Logical Restrictions

The Web Ontology Language (OWL) is a language for modeling data. Using the language, webmasters can order their data into certain groups (or *classes*) and describe the relationships between data and classes. When asked questions about a dataset modeled with OWL, computers can follow the logic of the language to draw conclusions.

In the working group for OWL, Patrick Hayes and several others had been calling for building an "expressive" logic - a logic that would allow researchers to "say anything about anything." One of the primary differences between OWL-Full (the expressive version of OWL that enables webmasters to "say anything about anything") and OWL-DL (the restricted version of OWL that ensures that a reasoner can always produce an answer in a finite number of steps) is around the issue of meta-modeling, or creating models of models. OWL-Full allows meta-modeling; more specifically, it allows modelers to treat classes as instances of a meta-class. Let me explain. Classes typically serve as encompassing structures, and instances

are members of those encompassing structures. You can compare this to the folder structure on a computer: folders (like classes) contain files (like instances). On a computer, you cannot treat a file as if it were a folder; files cannot *contain* anything else. The same is the case for OWL-DL; in OWL-DL, meta-classes are not allowed. Classes and instances are absolutely separate entities. So perhaps, I have a class called "endangered species," and it contains the instances "bald eagle" and "white leopard." In OWL-DL, since "bald eagle" is an instance, it cannot also be a class, and since "endangered species" is a class, it cannot also be an instance. Yet, in OWL-Full "bald eagle" can serve as both an instance and a class - perhaps an instance of the class "endangered species" and a class of individually named bald eagles. For a computer to be able to distinguish that "bald eagle" serves as both a class and an instance, a meta-model would need to be constructed (Allemang and Hendler, 2011).

To understand why this distinction is so important, we need to jump back about a century - to a time when formal logic was first emerging as its own form of semiotic infrastructure. In the early 20th century, Bertrand Russell had been closely following set theory - a branch of mathematics that studied how operations could be applied to objects ordered into sets. In set theory, objects and sets could be combined in any way; sets could contain objects, other sets, or even themselves. Russell recognized that a paradox could emerge when set theory allowed for sets to contain themselves. More specifically, paradox emerged when he tried to define a set of all the sets that do not contain themselves. Let's say I have a set P, which represents a set of all the sets in the world that do not contain themselves. Should P contain itself? As soon as it did, it would no longer be a set that doesn't contain itself. However, without including P within itself, it would not in fact be a set that contains all sets that do not contain themselves. A no-win situation.

To prevent this paradox, Russell introduced a Theory of Logical Types (Whitehead and Russell, 1925). The Theory of Logical Types posited that there needed to be a separation between classes and the members of classes - that classes and their members constitute different levels of abstraction. In unearthing this paradox, Russell put a kink in entrenched beliefs that sets could be constructed to "say anything

about anything." He showed that logic had to be *restricted* in some way to prevent paradox.

In advocating for making classes and instances disjoint (for making the two entities different Logical Types), the Semantic Web practitioners backing OWL-DL were not specifically concerned about reproducing Russell's paradox. Logic has come a long way since the early 20th century, and many new logical restrictions have emerged that ensure that Russell's paradox won't get in the way of formal reasoning. However, the concern about other forms of paradox emerging with meta-modeling came up often in the development of OWL. Further, meta-modeling adds layers of complexity to computer models. Restricting that complexity helps to guarantee that, when asked a question about some knowledge encoded in a knowledge base, a computer can produce an answer in a reasonable amount of time. In disallowing meta-classes, OWL-DL makes this guarantee. In allowing meta-classes, OWL-Full does not.<sup>3</sup>

The proponents of OWL-Full had been arguing that the restrictions in OWL-DL would disallow webmasters from representing knowledge with greater complexity. Sorting knowledge into neat logical types may ensure that a computer can produce an answer in a reasonable amount of time, but it would preclude many knowledge orderings (including orderings that may produce paradox). To make OWL work for webmasters worldwide (webmasters that may order their data into diverse logical types) they argued that OWL needed to be looser, scruffier, and more tolerant of inconsistency. OWL needed to be able to model knowledge in a world riddled with paradox and, at times, double bind.

---

<sup>3</sup>OWL 2 introduced a new feature called "punning," which allowed a Web resource to be treated as both a class and an instance. Yet, Dean Allemang and Jim Hendler (2011, 329) advised against the use of punning for meta-modeling in their book *Semantic Web for the Working Ontologist*. They write:

There really is a difference between a species and the set of animals of that species; there is a difference between the desktop and the applications that run on it. The relationship between a bottle of wine and its vintage is different from the relationship between an eagle and its species, and these distinctions could be important to someone who wants to reuse a model. Keeping them distinct in the first place will often enhance the model's utility.

Part of the expertise of a semiotic technologist is demarcating which differences make a difference in their work.

Anthropologist Gregory Bateson, in writing double bind theory, was actually responding to Russell's Theory of Logical Types. Citing how Russell's theory required an abstraction between classes and their members, Bateson (1972, 202) wrote, "Although in formal logic there is an attempt to maintain ... discontinuity between a class and its members, we argue that in the psychology of real communications this discontinuity is continuously and inevitably breached." Double binds emerge from breaches in the discontinuity of Logical Types - breaches that emerge continuously and inevitably in the "real world." As different orders of communication collide, individuals are exposed to situations where they must meet conflicting demands. Yet, Bateson (1972, 208) goes on to argue of double bind, "...without these paradoxes the evolution of communication would be at an end. Life would then be an endless interchange of stylized messages, a game with rigid rules, unrelieved by change or humor." As I will go on to show, concern for the ways rigid rules of representation foreclose possibilities for ordering knowledge in alternative ways has become increasingly prominent in Semantic Web work. Learning to creatively endure double bind is enabling evolutions in communication.

## 4.2 Shape-Shifting Knowledge

During his keynote at the World Wide Web conference, Sir Tim Berners-Lee noted that keeping the Web open and decentralized demanded that Web engineers acknowledge that they cannot predict what it will look like a decade from now. The Web was designed with a commitment to radical decentralization (Berners-Lee, 1999). In the Web's initial design, there were no central authorities controlling what ended up on the Web or how Web content would be organized,<sup>4</sup> and still today there is no centralized computer system controlling how documents get added to or deleted from the Web. A user anywhere in the world can upload a document to a server and then start linking it to other documents on other servers with the proper hypertext protocols. While there are still considerable divides in terms of who has the resources and expertise to upload content onto the Web, contributions to the

---

<sup>4</sup>While this is still technically the case, content organizing sites like Google and Facebook are, to a certain extent, recentralizing the Web (see for example (Gerlitz and Helmond, 2011))

Web come from a broad range of geographies and represent a broad range of cultural and political viewpoints. Because of this, both the Web and the knowledge contained within it are quite dynamic. The Web goes through waves of decentralization and recentralization;<sup>5</sup> Web knowledge iterates as new communities contribute to it.<sup>6</sup>

The Web itself has been characterized to me as a scruffy architecture - cobbled together rather than centrally formalized, tolerant of error rather than clean and consistent. Without a central system controlling what gets added, deleted, and edited, there is little preventing users from using protocols incorrectly or from deleting documents that have already been linked to from elsewhere on the Web. This is why you sometimes navigate to a page and receive a "404" or "Not Found" error - because someone created a link to something that did not exist or was later changed or deleted. Berners-Lee's decision to allow "404" has become a sort of urban legend in the Web community - a story referenced to acknowledge that the Web "won out" against other hypertext systems because it was completely decentralized and tolerated inconsistencies. As Deb McGuinness noted during an interview:

I think Tim Berners-Lee would even say, the Web...embraces hacks. The Web actually embraces 404, you know, not finding something. I mean that's one of Tim's brilliant contributions to the Web - that it's fine to not find something; you just know that you didn't find it. So actually I think Tim Berners-Lee would be happy with this characterization that the Web needs to work for hackers, and work in an imprecise, somewhat sloppy world.

The Web can be referred to as what Michael Fischer (2012) calls a "soft infrastructure." In soft infrastructures, architecture is underdetermined,<sup>7</sup> enabling the

---

<sup>5</sup>Understanding the Web as a socio-technical system has been the aim for the Web Science community since 2006 (Hendler et al., 2008). This community understands macro-level phenomena to "emerge" from micro-level architectures and social practices.

<sup>6</sup>Cultural theorists have suggested that hypertext - a semiotic infrastructure foundational to the Web - embodies a post-structuralist approach to language (Landow, 2006).

<sup>7</sup>The concept 'underdetermined' is used in the philosophy of science to mark when there is not enough empirical evidence to support a claim. For Quine (1975) since scientists make claims in spite of their inability to observe all possible phenomena, science is perpetually underdetermined. In mathematics, the concept 'underdetermined' marks when a system is under-restricted making it possible to have an infinite number of solutions. Perhaps most pertinently, Mark Poster (2001,

infrastructure to transform and evolve as it is exposed to new contexts. In this sense, Fischer argues, soft infrastructures are open to "emergent forms of life" (also known as culture, expertise, politics, subjectivities, temporalities, etc). Because soft infrastructures are underdetermined, they also tend to be quite scruffy; there are fewer rules dictating how content will be ordered and fewer restrictions preventing paradox. With and within soft semiotic infrastructures users have more freedom to "say anything about anything." Noting that the philosophy of the conventional Web inhibited it from ever being a "well-organized library," Berners-Lee et al. (2001) wrote in the publication introducing the Semantic Web, "Semantic Web researchers ... [also] accept that paradoxes and unanswerable questions are a price that must be paid to achieve versatility." In what follows I describe some of the challenges that have emerged in encoding Web knowledge.

#### 4.2.1 Encoding World Wide Knowledge

Even amongst the most traditional knowledge engineers, knowledge on the Web is not considered universal and absolute. Diverse people in diverse settings with diverse aims and perspectives upload data on the Web. The data can therefore be quite inconsistent; people disagree and thus can upload conflicting information; they make mistakes and typos when uploading information. And we've come to learn that people often deliberately upload false information to the Web. Consider how knowledge would be cleanly and consistently encoded on The Onion, where most of the "data" is satire, or certain news stations that create "fake news." How can a clean model make sense of such scruffy and inconsistent data? As James Hendler wrote in one commentary:

---

17-18) describes an 'underdetermined' social object as follows:

With the term *underdetermination*, I contend that certain social objects that I call virtual (hypertexts, for example) are overdetermined in such a way that their level of complexity or indeterminateness goes one step further. ... they do not direct agents into clear paths; they solicit instead social construction and cultural creation. ... A type of object thus emerges into social space that is overdetermined in the sense of being structured through multiple contradictory practices but is also underdetermined in the sense that it remains an invitation to a new imaginary.

Poster's distinction between overdetermination and underdetermination is related to the distinction Derrida draws between polysemy and *différance*.

... on the Web there is no way to guarantee consistency. It ... contains information that is inconsistent, incorrect, lacking reliable sources, combined with other information without author approval, and much more. Even worse, if I point at terms in your ontology, and then you change it (or move it), my representation becomes ungrounded. [Knowledge representation] has never before had to deal with the AI equivalent of an Error 404! (Harmelen et al., 2002, 74)

For Christian Bizer (co-founder of DBpedia, who I interviewed in April 2016), this means that it is important that Semantic Web designers view all data on the Web as *claims* and not as *facts*. He told me:

...[a] large part of the community for quite a long time [was] not willing to accept that the Web contains claims and not facts. This was [a problem in] the community for quite a while. And the problem was why didn't they accept the obvious. Of course, if I ask my mother does the Web contain claims or facts, she would say of course it's claim.

For Bizer, this distinction was important because often on the Web, semiotic technologists would encounter cases that violated claims they assumed to be factual. When I asked Bizer later in the interview if he saw any limits to achieving global data integration with linked data technologies, he referenced one of his "preferred examples." "Usually as a human, if I [ask]: Is a village and a tunnel the same? Or is a populated place and a tunnel the same? You would say no."

I confirmed. He noted: "It's class disjointness." Class disjointness refers to a rule in many ontologies (including OWL) that specifies that two categories will never overlap, or that a data point will never be an instance of both categories. Marking disjointness *restricts* the ontology; it lets a reasoner know that it need not look through a class called *populated place* when asked a question about a tunnel. Because of this, specifying that classes are disjoint can make it easier for knowledge representation systems to infer relationships and produce results in a reasonable amount of time. In DBpedia, *architectural structures* are disjoint from *populated places*. Bizer continued:

A tunnel is not a populated place. [But] if you look at reality, or even if you look into Wikipedia, you find that there's a tunnel in India that contains a slum, so a tunnel is a populated place. It violates your logical assumption, but still the logical assumption is quite useful. So if you want to cleanse Web data, even though it's only 98% or 99% true, the class disjointness helps you. But there are cases, which are true which violate the axiom. So basically, I think the Semantic Web community thought for a long time that things would be easy, but now as we look at reality, as this example nicely illustrated, it turns out that things are not as easy as we hoped.

Because data on the Web is contributed from diverse geographies and cultures and represents diverse points of view, it cannot easily be ordered into classes and their instances. In other words, data on the Web is already riddled with scruffy style; there are already cases on the Web (and "in reality") where claims that are made cross logical types and violate logical axioms. Thus, designers of the Semantic Web have acknowledged that modeling Web data will require sacrificing neatness and formal logic. For instance, in one email exchange discussing how logic could be applied to data formatted with RDF, Pat Hayes (2000) wrote:

Being a good oldfashioned logical type, I'm all for good oldfashioned logics (as 'full' as we can manage), but I think that we will need to modify our old ideas about semantics to accomodate to the web's messiness. In particular, we have to come to terms with the fact that logical names aren't just logical constants any more ... and may even get used in 'nonlogical' ways, eg as words in [natural language] text... We will have to be able to deal with the fact that inconsistencies will arise involving assertions from disparate sources, and the processes of resolving them may need to take into account the nature of these sources and some kind of notion of their warrantability.

Importantly, Web knowledge is not just already messy; it's also perpetually messy. Because the architecture of the Web is open and decentralized, Semantic

Web practitioners cannot predict when new content will appear on the Web that could contradict existing content. They cannot predict how people will describe their data and link it to other data, and they cannot predict when content will be taken down, breaking links. They also cannot predict how language itself will change as cultural and political evolution mark the emergence of new words and the retirement of others. Because the Web is a soft infrastructure - because its architecture is underdetermined - meaning on the Web is also underdetermined. Meaning too can iterate and evolve as diverse populations in diverse settings with diverse perspectives bring new or different interpretations of the world to the Web. Encoding worldwide knowledge on a soft infrastructure thus requires figuring out how to build tools for encoding knowledge that is always already incomplete.

#### 4.2.2 Encoding Knowledge Worldwide

Notably, the designers of Semantic Web infrastructure are often not the individuals modeling knowledge on the Web. The designers of Semantic Web infrastructure design RDF, OWL, and other Semantic Web technologies for "everyday webmasters." They hope that these webmasters will use the tools that they build to model data on their own Web pages. Using this crowd-sourcing technique, they hope, will enable the Semantic Web to very quickly grow into a massive knowledge base.

This means that tools like RDF and OWL, as well as taxonomies for describing data, need to work for webmasters worldwide. Unlike in communities like the geosciences and genetics (where ontologies are designed with a specific domain in mind) RDF and OWL (which serve as the building blocks for many ontologies in specific domains) need to work for every person in every domain; they need to be structured in such a way that they can model any form of knowledge. However, people across the globe, in different domains and cultures, define, structure, and relate concepts differently. Since the early 2000s, the designers of Semantic Web infrastructure have attempted to design the infrastructure in such a way that webmasters will use it "correctly." They have written out extensive documentation, primers, and formal definitions that outline exactly how the taxonomies, data formats and ontologies

should be used. Yet, they found that webmasters often used the tools in ways that didn't align with the semiotics they (the Semantic Web designers) sought to encode.

Take for instance, OWL:sameas, a primitive of OWL that is used to mark when two pieces of data on the Web are absolutely identical - for instance, that a Web page describing "Rensselaer Polytechnic Institute" is referring to the same real world thing as another Web page describing "RPI." Designers of sameas have lamented that the primitive is not used correctly "in the wild" (or by your average webmaster); in other words, people have used the primitive to mark things that are sort of the same - perhaps that information studies is the same as information science (Halpin et al., 2010). Neater semiotic technologists have argued that OWL:sameas should be stricter - that the "misuse" (305) had led to a philosophical "identity crisis" (307) that was turning the interconnected web of data into "the semantic equivalent of mushy peas" (308).

Yet, in an interview with James Hendler, a self-declared "scruffy" Semantic Web expert, he described of OWL:sameas:

So its usage - its pragmatics - don't actually concur with the formal specification, and that's driving a bunch of people crazy because the formalists are saying "make all of those people stop doing something wrong." And us pragmatists are saying, "well, let's see there's a hundred million of them and 6 of you."

He laughed, and continued, "Please go ahead and convince them that they're wrong."

Semantic Web designers do not just have to deal with being unable to predict how knowledge will iterate; they also have to deal with being unable to predict how the tools they build for modeling knowledge will be used.

Designing knowledge representation for the WWW (a scruffy base infrastructure) is thus pushing Semantic Web researchers to bring different assumptions about language to their work. They cannot assume that meaning will stabilize, and they cannot build tools that assume that meaning is already stable. They have had to acknowledge that they may never produce a functional or neat semiotic infrastructure.

Instead they have had to learn to design components for a semiotic infrastructure that (like the Web) is a soft infrastructure - an underdetermined infrastructure that can evolve as new meaning emerges or as tools are used in new ways. Working in the "real world," on top of an infrastructure that may not be quite "worldwide" but certainly represents a wide world of knowledge, may provoke paradox in Semantic Web work, but it also enables experimentation - experimentation that (I will argue) can help bring about new styles of semiotic infrastructure design.

### 4.3 Delaying Semantic Commitment

As I was wrapping up an interview with Ora Lassila in October 2016, he asked me to describe what I'm trying to say with my research. I summarized that I'm interested in how knowledge representation communities come together and learn to work in the face of limits. He told me he was glad I brought this up and went on to describe what he believed made Semantic Web work different than other standardization work:

Standardization as a means of achieving interoperability is that we decide in advance what are the possible things to say and what do they mean - so that two systems can then talk to one another. ... The Semantic Web takes a different approach. We don't actually try to standardize what are the possible things that you can say. We only standardize how to say them. And we give a framework that allows me to give you some clues as to how to interpret the things that I'm saying that you have not heard before, which I refer to as *delayed semantic commitment*. So standardization has this disadvantage in that it *tends to be a limiting thing for technology because you cannot anticipate everything in advance, and if you haven't anticipated those things then maybe they cannot be done because the standard precludes them*. And Semantic Web, particularly RDF was constructed to be free of this limitation. (emphasis mine)

For Lassila, standardization is a neat approach to encoding semiotics. It stabilizes meaning by laying out the ground rules for what people can say and how

they can say it. Standardization *restricts* what can be said about data and how meaning can be interpreted from data. However, for Lassila, this was not the goal for the Semantic Web. A strategy of "delaying semantic commitment" emerged in recognition that, at least on the Web, it would not be possible to anticipate how meaning would evolve or for what purposes people would use the semiotic infrastructure. It emerged because the semiotic technologists working on the Semantic Web did not want to preclude potential divergences in meaning. The strategy that Lassila describes here does not attempt to delimit what becomes knowable, nor does it assume that knowledge or even logic is universal and absolute. It acknowledges that achieving signification is always a process of deferral; achieving signification is indefinitely postponed, or "delayed" (Derrida, 1970).

Many Semantic Web designers have told me that early attempts at designing Semantic Web infrastructure failed to take off like the Web had in the early 1990s because they attempted formalize a set of rules (or in other words to standardize meaning) too soon. Dame Wendy Hall, a computer scientist and founding director of the Web Science Research Initiative, described this in a June 2015 interview:

Tim painted his vision of linking data, and he called it the Semantic Web. And people [in] the AI community picked this up, because of the semantics, and I talk about the Semantic Web going down an AI rat hole in that it got quite distracted by the AI community trying to sort out the issues before there was any data out there. [...] A lot of theory was talked about - a lot of upside down As and backward Es to try and prove things about ontologies and try and work out the theory of a Semantic Web before there was any data to experiment with.

Hall was referencing the precursor to what I've come to recognize as a notable discourse shift by some in the Web community around 2006 - from the "Semantic Web" to "linked data" - a shift heralded by the publication of a paper entitled "The Semantic Web Revisited" (Shadbolt et al., 2006). Those endorsing the shift argued that it was impractical to try and define schemas and ontologies before there was any data to model. Taking to heart that the Web would never be well-organized,

rather than mapping pre-formalized ontologies onto Web data, they instead built ontologies from Web data.

Consider DBpedia (referenced earlier in the chapter) - an initiative to structure data on Wikipedia pages with "semantics." DBpedia developed an ontology to describe data on Wikipedia pages. For instance, take the Wikipedia page for Rensselaer Polytechnic Institute (RPI). The DBpedia project extracts the string "Troy, NY" from the RPI Wikipedia page and wraps it with code that tells a computer that this phrase refers to RPI's location. Using RDF, it creates a "triple" "(subject: RPI) / (predicate: is located in) / (object: Troy, NY)" and stores this in a knowledge base. Similarly, the phrase "Puckman" gets extracted from the RPI Wikipedia page, wrapped with a bit of code to tell a computer that this phrase refers to RPI's mascot, and stored in a knowledge base as "(subject: RPI) / (predicate: has mascot) / (object: Puckman)." "Location" and "mascot" are defined in the DBpedia ontology as properties of "Universities." Through this effort, users can query the knowledge base to get a list of all the universities located in Troy, NY (among many other things).

Rather than building an ontology from scratch (based on the properties the designers presumed universities to have), the DBpedia ontology was built from data that was already listed in Wikipedia pages. All Wikipedia pages contain an "Infobox" - a box on the right hand side of the page that summarizes the data on the page. The designers of DBpedia's ontology extracted common properties listed in the Infoboxes for thousands of Wikipedia pages and then mapped these properties into a DBpedia ontology. This is how they identified "location" and "mascot" to be properties of universities. Notably, Wikipedia itself is a decentralized, crowd-sourced initiative. Contributors from across the globe can add to and edit Wikipedia pages. Because of this, the properties listed in Infoboxes are constantly evolving. To keep the ontology dynamic, the creators of DBpedia have opened the ontology's maintenance to the community; anyone (with the proper access and skills) can suggest edits to the way properties listed in Infoboxes get mapped into a DBpedia ontology. The ontology iterates as diverse users modify it, and so the designers cannot predict what it will look like in the future.

With linked data, the aim was not to get the semantics "right" ahead of time but instead to provide the scaffolding to enable diverse communities to describe the relationships between their own data. In this sense, rather than imposing a particular ontology language or schema on webmasters ahead of time - rather than pre-formalizing a set of rules for how data could come together - linked data ontologies would be emergent. They could evolve organically as webmasters linked their data together in new ways; they embodied delayed semantic commitment.

As a result, different assumptions guided the way that reasoning was applied to data. With earlier attempts to formalize ontologies ahead of time, computers could reason on data because they already knew how to logically follow the rules. They could interpret meaning from the data because the meaning had been encoded ahead of time. However, rather than formalizing rules ahead of time, many in the linked data community advocated for statistical reasoning to be applied to linked data after it had been linked. Statistical reasoners would interpret meaning from data, not based on what they already knew about how data could come together, but instead based on trends that emerged in the way data was linked. This is how Google's knowledge graph works today. It does not sort all of the knowledge on the Web into a complex ontology. Instead, it uses statistics to determine which data points tend to be linked on the Web. For instance, when I typed "when was the World Wide Web invented" into a Google search bar in September 2017, it returned "1989." There are many, many pages on the Web that suggest that the Web was invented in 1990 or in 1991. It would not be possible to sort all of this data into a neat and consistent ontology. Instead, Google returned "1989" because most Web pages Google has deemed authoritative on the subject mark the date the Web was invented as 1989. Google applies a form of statistical reasoning to Web data to infer this result.<sup>8</sup> Christian Bizer described this during our interview:

If you ask Tim [Berners-Lee], kind of everybody is shifting. But basically there's strong deployment on the distributed data provision side, which

---

<sup>8</sup>Paul Cilliers (1998, 35) reminds that statistical reasoning does not imply capturing the complexity of differences within a system. While statistical reasoning has been a response to scientific unease with deterministic methods, "It remains a tool in the process of establishing the true mechanisms of the phenomena being investigated. The heavy price paid in the process - that of averaging out the complex internal detail - is usually glossed over."

is an integral part of the Semantic Web. There's strong deployment on the ontology side because basically people agree on the global ontology, which is Schema.org, which is not as formal. You can't do reasoning with it. ... but you do some other kind of reasoning. You [can] do more statistical reasoning, not so much based on ontology constructs. More based on the data, which is a trend that you might describe in your PhD.

The linked data approach to knowledge representation on the Web is much scruffier than earlier approaches (such as formalizing OWL). It considerably "softens" the rules for how data can be modeled on the Web. Actually, since 2006, it seems that most new Semantic Web tools keep getting softer and softer. These tools (such as FOAF, JSON-LD, Schema.org) allow webmasters to "say anything about anything" - to link and describe their data in any way that they want. They give some clues to computers and folks with whom they share their data as to how to interpret the data, but they do not preclude people from bringing different interpretations to the data, and they do not break formal reasoners when people inevitably do bring these different interpretations. They remain open to emergent forms of meaning.

Fischer argues that soft infrastructures buttress "experimental systems." For Hans-Jorg Rheinberger (1998, 287), experimental systems are "future-making machines;" they produce results that are "beyond our present knowledge." Rheinberger (1998, 291) describes the architecture of an experimental system as a "labyrinth":

...one never knows exactly where it leads. As soon as one knows exactly what it produces, it is no longer a research system. An experimental system in which a scientific object gathers contours and becomes stabilized, at the same time must open windows for the emergence of unprecedented events. While becoming stabilized in a certain respect, it must be destabilized in another. For arriving at new "results:" the system must be destabilized-and without a previously stabilized system there will be no "results." Stabilization and destabilization imply each other. If a system becomes too rigid, it is no longer a machine for making the future; it becomes a testing device, in the sense of producing standards or replicas.

Standards produce "testing devices" - devices that "harden" or "stabilize" the contours of working with and within them. On the other hand, soft infrastructures enable experimental future-making *because* they are underdetermined. Because they lack "rigid rules." Because the conditions for working with and within the infrastructure have not been completely formalized ahead of time. This is the case for many linked data initiatives; these designers have had to learn to navigate a labyrinth - designing systems that loosen rigidity, while acknowledging that this means they won't be able to predict where the systems will lead.<sup>9</sup>

Approaching knowledge representation experimentally requires a different kind of "reasoning" and a different kind of "logic" than the reasoning and logic that has tended to fortify entrenched techno-scientific epistemologies. Softness does not just mark an evolution in the style of semiotic infrastructure; it also marks an evolution in the style of knowledge representation thinking and work. In building knowledge representation on top of a soft infrastructure (the WWW and language itself), the Semantic Web community has become entwined in an experimental system; their culture and their style has iterated and evolved in this new context.<sup>10</sup>

#### 4.4 Learning to Swim in Troubled Waters

Building a soft infrastructure demands figuring out how to stabilize the contours of an experimental system just enough to create possibilities for working and communicating, but destabilizing them just enough to enable evolution and iteration. For Rheinberger (1998, 291), there is no neat or logical way to do this: "It forces one to move by means of checking out, of groping, of tatonnement." Neat design styles (or standards for designing) restrict design logic and design work. But logic and standards are also enabling. It is much easier to move through a system

---

<sup>9</sup>Kim Veltman (2006, 14) argues that Semantic Web designers "continue to assume that words, topics and disciplines are fixed and unchanging, not unlike thinkers prior to the time of Linneaus, who assumed that knowledge and its categories were fixed and static. Until we develop dynamic data structures and databases to address these dynamic dimensions of knowledge, we cannot hope to understand how contemporary categories have evolved, let alone develop frameworks to stimulate their future evolution." I argue, that in working with a soft infrastructure, this assumption has iterated.

<sup>10</sup>Fischer (2012, 3) argues that culture (defined as a "methodological tool") can be understood as "refined into a series of experimental systems." Culture too changes as it is exposed to new realities and new contexts.

when the path through it has been clearly and unambiguously set up ahead of time. It is much easier to move through a system when there are rigid rules dictating how to do so. However, when building a soft infrastructure, there are often multiple possible paths, and designers cannot know where they will lead or if they will lead to a dead end.

In this section, I show how designing soft infrastructures demands learning to work at, against, and in spite of various limits. For instance, it demands designers to push the tradeoff separating stability from instability, the tradeoff separating premature rules from emergent meaning, and the tradeoff separating neatness from scruffiness. Designers of soft Semantic Web infrastructures - infrastructures like Schema.org - have had to learn to work both with and against these conditions.

Schema.org is a project that seeks to develop a very simple schema of properties that webmasters can use to describe content on their Web pages. It is a newer linked data initiative that many informants have described to me as what is "winning out" in knowledge representation today or as the first truly "global" ontology. Webmasters can use properties defined in Schema.org to "mark up" data on their Web pages. Imagine that I have a Web page about the "Society for the Social Studies of Science" (4S). Using properties defined in Schema.org, I can wrap the string "1975" with the Schema.org property "foundingDate" to let search engines know that this piece of data on the page refers to the date that 4S was founded. I can also wrap the string "International Social Science Council" with the Schema.org property "parentOrganization" to let search engines know that this piece of data on the page refers to 4S's parent organization. Both of these properties are defined in Schema.org as properties of organizations. Today more than 10 million Web pages use Schema.org to mark up data.

The initial release - the core of Schema.org - was created in 2011 as a joint collaboration with representatives from Google, Yahoo, and Bing. A year later, the Russian search engine Yandex, joined the collaboration.<sup>11</sup> Schema.org initially drew

---

<sup>11</sup>While these search engines are typically in competition with each other, they agreed to collaborate on this initiative because they saw a significant benefit to getting everyone on the Web using the *same* ontology. Had they each developed their own ontology language, it could have further fractured efforts to integrate data on the Web.

heavily from other established, yet more robust, Web vocabularies such as Microformats, FOAF, and OpenCyc to select properties to include in the ontology. Shortly after this release, the development, maintenance, and extension of Schema.org were vastly opened to the Web community. Now, anyone on the Web can collaborate on the project through a W3C sponsored mailing list. A steering group comprised of representatives from sponsor companies, a W3C representative, and individuals that have contributed substantially to Schema.org over the past five years make all final decisions about the schema (W3C Team 2015).<sup>12</sup>

Taking this participatory approach to designing the schema was, in part, in effort to democratize knowledge representation. Rather than having highly trained logicians define the ontology in advance, lightly trained webmasters could ask for properties to be added to the schema as they needed them. In other words, the ontology responded to the data, not the other way around; it delayed semantic commitment. The creators of Schema.org, in line with many in the linked data community, were incredulous towards knowledge representation efforts that attempted to formalize ontologies before there was data to experiment with. However, they also recognized that in a knowledge representation system where webmasters could use the properties however they wanted, there would likely need to be some cleanup. Guha et al. (2016, 50), the creators of Schema.org describe:

Schema.org also shares the linked-data community’s skepticism toward the premature formalism (rule systems, description logics, and so on) found in much of the academic work that is carried out under the Semantic Web banner. While Schema.org also avoids assuming that such rule-based processing will be commonplace, it differs from typical linked-data guidelines in its assumption that various other kinds of cleanup, rec-

---

<sup>12</sup>Throughout its history, various stakeholders have become involved in public discussions. To provide just a few examples, just a few months after the launch, rNews, a schema developed by the International Press Telecommunications Council, in collaboration with the New York Times, was folded into Schema.org (Miller, 2011). The U.S. Office of Science and Technology Policy also became involved in defining the schemas for job postings. At Rensselaer Polytechnic Institute, the Tetherless World Constellation spearheaded the effort to include schemas for describing datasets. The contributions of such stakeholders are notable; the vast majority of changes and extensions to Schema.org have been backed by an individual or institution with a large degree of influence in the Web community.

conciliation, and post-processing will usually be needed before structured data from the Web can be exploited in applications.

Designing Schema.org required figuring out how to encode just enough rules for the system to enable communication and collaboration, but remain open enough to prevent the schema from precluding meaning that had not yet emerged. In what follows, I describe two scenarios where Schema.org designers had to work at this limit. The examples I provide demonstrate how, as Rheinberger (1998, 291) argues, Schema.org designers have had to "move by means of checking out, of groping, of tatonnement" in designing this soft semiotic infrastructure.

#### 4.4.1 Scoping Schema.org

In October 2011, one Schema.org contributor proposed to the mailing list adding a new 'type' to Schema.org called 'activity' after being unable to structure data on tourist activities on his website using existing classifiers. The resulting email thread went on for days. Immediate responses to the contributor's email suggested that web developers wishing to characterize a tourist activity could reference other, more robust schemas, such as Cyc or DBpedia, to include this type of information. They argued that an enumerated list of all activities would quickly become too large to manage. The contributor to the W3C Public Vocab Group (2011) responded:

I understand adding all these subtypes pose all kinds of problems such as ontological and perhaps even political issues. Anyway if you at least just add the new type Activity webmasters could use extensions to define the subtypes so you (Schema.org) don't have to take a stance on what subtypes to define.

Some individuals on the mailing list expressed support for the request, arguing that, while it may not be possible for Schema.org vocabularies to categorize the whole world, they did need to cast a wide enough scope to support the needs of "ordinary webmasters":

Sure, neither schema.org nor another vocabulary will encompass all terms that are required by webmasters - well-identified by many commentators

as the problems implicit in trying to create an "ontology of everything" - but in that case what are the boundaries of schema.org? The limits of schema.org are not ill defined, they are not defined at all. And in this there is a disconnect between enterprise outfits (and semantic web developers) and ... "ordinary" webmasters which threatens to alienate the latter. (W3C Public Vocab Group, 2011)

There is general consensus in the Schema.org community that the schema should not become an "ontology of everything," listing and defining every property that may be used to characterize web content. It does not enable webmasters to "say anything about anything" (though it does provide mechanisms to extend the core schema for new domains). Schema.org participants are very wary of the schema repeating the Cyc project (see Chapter 3) - an ontology that took almost a decade to build and that became complex to the point of inaccessibility. Furthermore, an ontology of everything is not particularly useful to search engines that are trying to figure out how to return better search results. Of course it is possible to divide the differences between concepts and properties infinitely, but without setting some limits to when these divisions should be arrested, search engines cannot discern when a user's search query is *similar* to particular Web content. Imagine, for instance, that within Schema.org there were different properties for every type of activity in which someone could engage, or perhaps different properties for every type of tourist activity, or perhaps different properties for every type of tour, or perhaps different properties for every type of walking tour, or perhaps different properties for every walking tour within a specific location... The divisions are endless, but describing data on a web page with this granularity will not be particularly useful when someone types into the Google search bar "things to do in Boston." Limiting the schema ensures that Web pages are encoded with some shared semantics.

Instead of an ontology of everything, the schema.org community has aimed to build out a "middle ontology" - one that most webmasters will find accessible and useful (Ronallo, 2012). As of today, there are no definitive limits to the schema's scope. As Dan Brickley, the chair of the W3C community group for schema.org, has noted, there is no "iron-clad policy" for when the schema exceeds a middle

ontology (W3C Public Vocabs Group, 2013a). Scanning through the email archives documenting Schema.org's design, the way the ontology gets delimited at first seems arbitrary. Should the vocabulary's size be limited to "10,000? 100,000?" properties, asked one contributor (W3C Public Vocabs Group, 2013a). Despite there being no explicit definition of where the ontology starts and ends, adding to and revising the ontology is by no means a free for all. Every suggestion for an addition to the schema must be defended fiercely against wariness that the schema will become too complex.

In delimiting the scope of the schema, designers have been forced to consider how to include enough classifiers to satisfy webmasters in diverse domains and settings, yet limited enough to not become overwhelming and to remain useful to search engines. In other words, they are working at the limit of defining worldwide knowledge without producing an ontology of everything. This is in part a challenge of deciding, collaboratively, which differences make a difference - which distinctions should be encoded in the schema and which should be left general. There are no steadfast rules about how to do this.

#### 4.4.2 Modeling Schema.org

Skeptical of premature formalism, Schema.org directors write in their wiki that an overarching design approach for the schema is to remain flat - to avoid nesting concepts and modeling with hierarchy: "there is no single right way to model anything. For our purposes, we have a bias towards flat models" (W3C, 2012). Designing semiotic infrastructure with flat models is another experimental strategy that semiotic technologists have employed in order to delay semantic commitment. When concepts are modeled in hierarchies - for instance, when walking tour is seen as an instance of day trips are seen as an instance of tourist activity - the semantic commitment designed into the system is hardened. It becomes much more difficult to use the property "walking tour" to describe anything other than a day trip or a tourist activity. Schema.org designers aim to keep the schema as flat as possible in recognition that, worldwide, different communities will organize their concepts into different hierarchies. Different communities will bring different meaning to their

concepts, so encoding these meanings prematurely would likely preclude future use cases.

However, in designing a flat schema, they introduce many inconsistencies into the system. Consider one issue that emerged as the Schema.org community attempted to develop a property for the concept 'abdomen'. Abdomen can at once refer to a body part and also refer to a type of physical medical exam. In a flat schema, however, there is no way of disambiguating these two uses of the same word. Disambiguating them would require finer levels of granularity - noting that 'abdomen' meant body part in this context, and medical exam in another. Yet imposing such levels would stratify a flat ontology - an ontology designed to be flat in acknowledgement that "there is no single right way to model anything." In other words, the Schema.org community has to work at a tricky limit: they must define the infrastructure "flatly" in order to enable concepts to take on new meaning as they move in diverse and unpredictable settings, yet in doing so, they enable concepts to take on a surplus of meaning (or to be polysemic) in any given setting - at times in instances when there is a difference that makes a difference between two uses of a word. Referencing this issue, one W3C Public Vocab group designer (2013b) responded:

I think you are rightly pointing to the implicit initial (and naïve) assumption of schema.org, which is that the whole world can be represented under a single flat namespace at arbitrary level of granularity, with natural language words as identifiers. Obviously, this does not scale and hits quickly the wall of polysemy, as the Abdomen example perfectly illustrates, and we are bound to have more of the same with the schema growth (which is, remind you, potentially unbound[...])

That said...other problems you point at (lack of documentation, semantic glitches etc) will always be present in this scruffy-work-in-progress called "Web semantics" (read: fuzzy, plural, inconsistent etc). I'm sure you will ever ever fight it with all your will and strength given where you come from, but I'm afraid this battle has been lost for quite a while now. As Pat Hayes told me a while ago "My ivory tower has been seriously shaken

these days, waters of real world are slowly rising around us." Time to learn swimming in troubled waters ...

To design infrastructure that encodes meaning and enables it to travel into diverse settings, semiotic technologists need to tolerate inconsistencies - to allow data to simultaneously be polysemic (having a surplus of meaning) and underdetermined (marked with an absence of full signifying closure). The designers of Schema.org need to model the ontology with enough formality to make it useful for interpreting the meaning of data in diverse contexts, while leaving the ontology soft enough to enable shifts and displacements in meaning. In other words, they need to push the tradeoff of formality and openness. Again, there are no steadfast rules for how to do this. It demands a mode of work that is different from standardizing, ordering, and logically typing. It demands a mode of work that is learning to swim in troubled waters. In what follows, I characterize the expertise that pursues working at these limits.

## 4.5 Expertise in Encoding Différance

Emerging literature in information studies has ethnographically examined how data communities produce standards for structuring data and encoding its meaning. Producing such standards has been important for advancing global and interdisciplinary research agendas (such as climate science, genetics, and oceanography). Since often scientific communities in different settings or disciplines use different terms to describe the same phenomena (or alternatively terms used in one community mean something entirely different in another community), capturing and encoding unambiguous meanings of data is extremely challenging. Edwards and colleagues (2011, 669) refer to this challenge as "data friction." They argue that every time a piece of data interfaces with a new person, digital system, or organization, there is greater opportunity for data to be misinterpreted: "In social systems, data friction consumes energy and produces turbulence and heat - that is, conflicts, disagreements, and inexact, unruly processes." Data friction pervades the Web - a digital space where data interfaces with new people worldwide.

The expertise at overcoming data friction involves helping communities reach semantic agreement or helping them "find common ground."<sup>13</sup> Ribes and Bowker (2009) provide perhaps the most robust characterization of this expertise. Ribes and Bowker describe the expertise of "ontologists," or experts who are hired to help interdisciplinary communities design ontologies for ordering, describing, and linking the data that will be produced in a collaborative project. Because data friction tends to make sharing data across disciplines challenging, ontologists design ontologies in an effort to reduce data friction in interdisciplinary, collaborative projects. The ontologists Ribes and Bowker studied moved between different domain communities, resolving epistemic disagreements about how data should be modeled and defined. For instance, they sometimes encouraged communities that couldn't come to agreement on the meaning of a particular concept to try to "decrease the granularity;" in other words, they encouraged communities to move to a higher conceptual level, focusing less on the words and more on the underlying meanings to come to consensus about the right way to describe and order their data.

"Decreasing the granularity" or moving to a higher level of abstraction to sort out conflicting meanings is a form of Logical Typing. Semiotic technologists introduce this sort of strategy to prevent paradox from emerging in information systems; they separate out meanings to restrict practitioners from "saying anything about anything." Through negotiation and consensus, they order meaning hierarchically so that new meaning cannot enter the system and produce more meaning - so that meaning cannot iterate and evolve.<sup>14</sup> The expertise at addressing data friction is both a logical and structural-functionalist expertise;<sup>15</sup> it seeks to stabilize - to find

---

<sup>13</sup>Sabina Leonelli (2010) argues that as data curators in model organism biology "package" data so that it can travel to and be understood in diverse settings, they have to anticipate where facts will travel. Their work involves both decontextualizing and recontextualizing knowledge.

<sup>14</sup>For Derrida (1970, 150) this structuring can be described as limiting play:

Nevertheless, up until the event which I wish to mark out and define, structure – or rather the structurality of structure – although it has always been at work, has always been neutralized or reduced, and this by a process of giving it a center or of referring to it to a point of presence, a fixed origin. The function of this center was not only to orient, balance, and organize the structure – one cannot in fact conceive of an unorganized structure – but above all to make sure that the organizing principle of the structure would limit what we might call the play of the structure.

<sup>15</sup>Paul Cilliers (1998, 42) describes how Saussure, often considered the Father of structuralism,

a stable center.

Standards (such as ontologies) become the central object around which the different meanings and interpretations diverse communities bring to their data revolve. Yet, Derrida (1970) reminds us that the center of a text is perpetually escaping its structure. Identifying a temporary central common ground may reduce data friction (or reduce polysemy), but that ground is also always shifting, proliferating, and evolving. Polysemy can only be sorted and logical types can only be ordered in a singular context - "within a semantic horizon" (Derrida, 1983, 350). The challenge of representing knowledge does not only involve figuring out how to reach consensus about definitions and schemas; it also involves grappling with the infrastructural conditions that make anything but temporary consensus impossible.<sup>16</sup> It involves figuring out how to coordinate meaning (so that we can communicate, share information, and solve problems collaboratively), while also acknowledging that meaning is endlessly deferred. Double binds emerge in semiotic technologists' work because of *différance*; we need ontologies to share information, but ontologies cannot (at least permanently) exhaustively encode and represent meaning.

Members of the Semantic Web community are coming to terms with this. They are witnessing how meaning is constantly escaping its temporary center - that knowledge in the "real world" is perpetually shape-shifting and semantic commitment is perpetually delayed. In the process, they are learning that, in order to approach semiotic infrastructure design, they will need to bring new assumptions and design modalities to their work. They are learning that they cannot assume there to be a standard formula or a "best practice" to advancing knowledge rep-

---

"understands language as a system in which every word has its place, and, consequently, its meaning." However, language, a complex system, does not operate this way. "Words, or signs, do not have fixed positions." Cilliers (1998, 114) goes on to argue that while some long for "a grand narrative that unifies all knowledge," "those who embrace postmodernism find it challenging, exciting and full of uncharted spaces. ... Which of these two evaluations apply is often determined by whether one feels comfortable without fixed points of reference."

<sup>16</sup>Bowker (2000, 670) argues that keeping data structures "open" - "though none will ever be completely so" - requires two things. First, it requires documenting in detail the development of a given database (creators, social and political contexts, etc.). The second step "is to admit that the goal of metadata standards should not be to produce convergent unity. We need to open a discourse - where there is no effective discourse now - about the varying temporalities, spatialities and materialities that we might represent in our databases, with a view to designing for maximum flexibility and allowing, as possible, for an emergent polyphony and polychrony."

resentation because the work itself is characterized by a paradox - the charge to encode *différance*. Instead, it will require a more creative and experimental expertise - one that is both attentive and responsive to the semiotic dynamics of specific knowledge domains, and one that is constantly testing and pushing the limits of knowledge representation tradeoffs in order to appropriately represent knowledge in those domains. Enacting such expertise demands letting up on legacy battles in favor of something more interesting and flexible (but also rigorous). As one participant on the Schema.org mailing list wrote, "Let's not revisit the tiresome RDF v OWL wars in this forum." Instead, he argued, "The trick is to be consistent in the face of pragmatic realities." The expertise of this type of semiotic technologist (a critical and sometimes devious semiotic technology) may be best characterized as a pursuit - a pursuit to respond affirmatively to conflicting demands, to be consistent in the face of inconsistencies, to creatively endure double bind.

## 4.6 Semiotic Tricksters

The divisions between approaches to knowledge representation (or the oppositions that have ordered knowledge representation thinking and work) have historically mirrored many of the ordered pairs that have propped up techno-scientific epistemologies. For instance, the divisions between *formal logic/pragmatism* and *neatness/scruffiness* are similar to divisions between *rationality/irrationality* and *universality/particularism* in techno-scientific discourses. Similarly, according to the assumptions of Analytical Philosophy and more specifically of logical positivism, *hard* sciences are seen as producing more empirical Truth than *soft* sciences. Logical positivists order these binaries according to a logic of negation (where p serves as a privileged signifier over (not)-p); the first half of the slash represents what is, and the second half represents what is not.<sup>17</sup> Because of this, logical positivists characterize irrationality, particularism, pragmatism, scruffiness, and softness in techno-scientific discourses as needing to be tamed, cleaned, hardened, or restricted in order to produce Truth. On the other hand, they tend to see rationality, uni-

---

<sup>17</sup>Feminist philosopher Val Plumwood (1993) describes how this dualist logic tends to naturalize domination, leading to further binaries of domination/subordination, such as male/female, mind/body, civilized/primitive, and master/slave.

versality, logic, neatness, and hardness as more precisely representing reality - as representing what *is*. For Donna Haraway (1988), science that operates according to these rational assumptions plays the "God trick" (where knowledge and the rules for ordering it are assumed to derive from an omnipresence that exists nowhere in particular, rather than from situated perspectives).

However, while rooted in logical positivist traditions, more recent work in knowledge representation has continuously situated the "real world" as marking the other half of the slash - the irrational half of the opposition. Working in scruffy domains where paradox has come to be considered inevitable, knowledge representation experts have described illogical conditions as "real world" conditions that get in the way of clean encodings. They have described them as "real world" conditions that challenge the universal applicability of their semiotic infrastructure designs. The *parasitic* (that which is typically seen as beyond the real world and that which renders the real world strange) has worked its way into the real world for these practitioners. In the process, the logic ordering semiotic technologists' work has been flipped. For this community, rational conditions have been characterized as what reality is *not*. What Haraway (1988) calls a "trickster nature" - "a shape-shifting provocative world which can never quite be pinned down" (Fortun and Bernstein, 1998, 31) has displaced dominant orderings of the real world.<sup>18</sup>

Haraway (1988, 593-594) figured the trickster, or Coyote from American Southwest Indian symbolism, to challenge Euro-American nature/culture divides. Coyote is cunning, persistently evading human control. Coyote places a check on God-like accounts of what constitutes the real world by suggesting that the real world may play tricks on us:

Acknowledging the agency of the world in knowledge makes room for some unsettling possibilities, including a sense of the world's independent sense of humour. ... Richly evocative figures exist for feminist visualizations of the world as witty agent. We need not lapse into an

---

<sup>18</sup>Haraway (1997, 127) has argued that in the 1990s, in part due to the explosion of hypertext through the World Wide Web, "meaning-in-the-making - the physiology of semiotics" has become "a more cyborg, coyote, trickster, local, open-ended, heterogeneous, and provisional affair. Sign interpreters are ontologically dirty; they are made up of provisionally articulated, temporally dispersed, and spatially networked actors and actants."

appeal to a primal mother resisting becoming resource. The Coyote or Trickster, embodied in American Southwest Indian accounts, suggests our situation when we give up mastery but keep searching for fidelity, knowing all the while we will be hoodwinked.

Encoding ontology is paralogical (a term typically defined to mean illogical). It demands encoding a real world that is constantly, cunningly getting in the way of encoding a real world. However, the prefix "para" can describe more than just contradiction. "Para" can also describe something as beside (as in parallel) and something as beyond (as in paranormal).<sup>19</sup> Semiotic infrastructure designers have had to work both in parallel to logic and beyond logic.

Fortun and Bernstein (1998, 270) describe the pursuit of science as working at the "middle" - "an unhappy medium: forever restless, questioning, insatiable, looking for trouble and almost always finding it." The sciences work between concepts ordered according to a logic of negation. In doing so, scientific work troubles formal logic's law of the excluded middle. "The middle, the mess, is real - the realest of the real" (more real, perhaps than the real world Peter Norvig begins to question at the end of Chapter 3). The middle is perhaps where a trickster nature disrupts natural orders. Figuring out where and how to work within this real mess is the trick of designing a soft infrastructure, the trick of working experimentally, and the trick of encoding a scruffy, trickster nature. It is a style of expertise that perpetually pursues the limits of knowledge representation.

---

<sup>19</sup>For Derrida (1994, 161), being is a hauntology, a specter, or an always already absent presence. Ontology is never stabilized, settled, or centered, because time continuously renders presence "out of joint." Thus, encoding ontology expels the spectrality inherent in being; it renders present what is always already absent-present: "Ontology opposes [hauntology] only in a movement of exorcism. Ontology is a conjuration." In other words, encoding ontology - or encoding what is - is a "trick" - a paranormal activity.

## 5. CATACHRESTICALLY DESIGNING SEMIOTIC INFRASTRUCTURES FOR THE HUMAN SERVICES

On June 9, 2016, Greg Bloom was the third speaker on a panel entitled "Solutions at Hand" at the Personal Democracy Forum - a conference that brings together technologists, hackers, government officials, academics, and journalists to discuss how technologies interact with government and society. At the time, Bloom was a Civic Imagination Fellow with Civic Hall Labs - a New York City-based collaboration center that designs technologies to advance the public good. Bloom's talk was entitled "Building a Safety Net for the 21st Century," and in it, he described a project he had been working on for several years to enable an "Internet of help."

By "help," Bloom was referring to the help offered by a variety of human services available in communities throughout the United States - homeless shelters, food pantries, domestic abuse shelters, and immigration services, for example. Bloom's talk addressed how citizens find these services and how they determine whether they are eligible for these services - information, he argued, that is not easy to track down. During the talk, Bloom quoted his sister, a public defender that "helps poor people facing criminal charges in court," as saying, "When it comes to things that are really needed by people who are really in need Google is a *ghost town*" (emphasis original).

Bloom went on to describe the work done to address this issue at Bread for the City - a non-profit agency that offers help and services to low-income residents in Washington D.C. Bloom had worked for Bread for the City as a "Communications Guy" for almost 4 years. He described how part of the agency's work involved referring people that came to them out to organizations that offered human services throughout the city. This meant needing to know and keep track of where those organizations were, how to contact them, which services they offered, and who would be eligible for the services. None of this information was readily available. Collecting and organizing this information took a great deal of work.

Thus Stacey Johnson, a social worker at Bread for the City, set up a resource

directory database in Microsoft Access to store this information. Bloom noted that the database kept track of more than 500 organizations, which offered more than 1000 services throughout Washington D.C. In setting up this database, Bread for the City became a hub of knowledge about human services. Other organizations would come to Bread for the City asking for copies of the database, and the agency would freely share their database with them. However, once they shared it, the knowledge hub became a mess. Bloom acknowledged, "As soon as that data left our doors, it started decaying, and it was incompatible ... We counted as many as 12 different resource directories across D/C. It got to the point where I gave up trying to update our own information and all of them over and over again." Bloom continued:

This is an endemic systemic pattern. Information about who is providing what to whom, information that should be freely available and a public good and abundant is more like a commodity - scarce, and scattered and redundant, and that pattern is a source of waste, misery and chaos for people who need help and people who are trying to help them; it stymies the ability of people like you, civic technologists, to build tools that can help them. And it also makes it hard for decision-makers, funders, and policy makers to evaluate how our safety net is doing at actually meeting community's needs.

Folks that I have interviewed in the human services community share Bloom's concerns about the state of referral data. While there are numerous software packages and data models for organizing referral data into databases, these resources are scattered and often times behind pay walls. There is no one standard that every agency is using to organize its data, and thus the data is very difficult to share and keep up to date. Further, non-profits that maintain referral lists often do not have funding, resources, or expertise to implement software and data standards. Many non-profits keep their referral lists in Microsoft spreadsheets or even Word documents. These issues are understood to derive from and contribute to a much larger cultural dynamic in the human services - that *there are lots of stakeholders in*

*the mix (funders, service providers, non-profits, help seekers, government agencies, software developers), and they all use different vocabularies in their work.* With language so diverse in the human services, building information systems to support sharing data between diverse stakeholders is very challenging.

The project that Bloom went on to describe in his talk - an "Internet of help" - aimed to tackle this problem by creating what he elsewhere referred to as an "interlingua" for the human services. His project, Open Referral, aimed to enable cross talk between different information systems and data standards that organize referral data. This would enable referral providers, service providers, and other stakeholders to more readily share referral data and keep it up to date. He noted, "Just like the World Wide Web, we started by developing an open data format" - a format for structuring referral data in such a way that it could easily be mapped into diverse existing systems and standards.

In this chapter, I shift away from the World Wide Web to a different world - a world where the stakes for encoding semantic commitment are much higher in a day-to-day way. This chapter discusses the design of several semiotic infrastructures that structure and encode the meaning of referral data so that the data can be shared between diverse stakeholder communities in the human services. I argue that semiotic technologists in the human services have put a great deal of faith in semiotic infrastructures to sort out the semantic heterogeneity in the domain. They believe that in designing data standards, taxonomies, and ontologies that "logically" reflect the underlying structure of the human services, they will not only technically support data sharing, but they will also provoke everyone in the community into using a common language - a common language that they deem necessary for ensuring that individuals seeking help are directed to the right information. Semiotic technologists have thus worked to design semiotic infrastructures that sort out the "hidden structure" of meaning in the human services - to figure out how words, concepts, and relationships should logically be ordered.

However, data standards in the human services never seem to become "standard."<sup>1</sup> As I will show in the chapter, once a community in the human services

---

<sup>1</sup>Sociological literature on standards has been summarized in detail in (Timmermans and Epstein, 2010).

decides that a certain standard does not meet their needs or reflect their commitments, the tendency has been to design a new standard. In this sense, despite efforts to design a common language, the development of new semiotic infrastructures has tended to further splinter the domain's language. I thus conclude the chapter advocating for certain semiotic infrastructure design strategies I argue to be most appropriate for advancing an "Internet of help."

## 5.1 Information and Referral as a "Real World" Search Engine

Navigating to the "About" page for United Way Worldwide's 2-1-1 service in October 2017, you would be greeted with the headline "Real People, Real Help" (United Way Worldwide, 2017). The text on the Web page continues:

You are not alone. Every day thousands of people across the U. and Canada turn to 2-1-1 for information and support - whether financial, domestic, health or disaster-related. 2-1-1 is a free, confidential referral and information service that connects people from all communities and of all ages to a specialist who will help you find local health and human services, 24 hours a day, seven days a week.

211 is a dialing service that links citizens in a community to specialists trained in "information and referral" or "I&R." The Alliance for Information and Referral Systems (AIRS), the main standards setting and certifying body for I&R, describes I&R as the "art, science, and practice of bringing people and services together."

I&R services are decades older than 211. Reviewing the emergence of I&R in the human services domain, Nicholas Long (1973) suggested that the most prominent antecedent to I&R services was the Social Service Exchange. The Social Service Exchange emerged as part of the 1870s charity organization movement (Towle, 1949). As both public and private organizations in the United States began offering a wide variety of services, it became much more difficult to ensure that different organizations did not offer duplicate services to the same family.

In cities across the United States, social service exchanges were set up to act as information clearinghouses. Much like a library index, inquirers could look up case records that detailed which organizations in the city had offered which services to which families (Joyce, 1943). Long (1973, 3) noted that the majority of I&R centers existing at the time could be traced directly back to social service exchanges - exchanges, he argued, that were "organized to prevent rather than facilitate access to human services."

However, the I&R centers proliferating in the 1970s were primarily organized to help people in need find social services. At the time, I&R centers were being established in public libraries, hospitals, welfare offices, schools, local United Way offices, and other community centers. Within these centers, paid staff members maintained lists of organizations and services to which they would refer inquiring community members (either via office visits or via phone) (Long et al., 1971). AIRS was founded in response to the growth of I&R centers and published its first set of standards organizing I&R service delivery in 1973.

While numerous 24/7 I&R hotlines existed before 211, the United Way of Atlanta was the first to designate 211 as an abbreviated dialing code for I&R services in May 1997 (Chepesiuk, 2001). One impetus for designating this dialing code was to remove the burden of folks calling 911 with non-emergency related inquiries. In 1998, a lobbying group comprised of the United Way, AIRS, and local organizations in Connecticut, Texas, Florida, and Georgia filed a petition with the Federal Communications Commission (FCC) asking for 211 to be established as a universal dialing code for I&R services (Telecommunications and Information Policy Institute, 2002). In July 2000, the FCC granted this request. Now when a citizen calls 211, they are connected to local organizations offering referral information.

I&R providers maintain databases of programs and services offered in their local communities and provide mechanisms for supporting help-seekers in finding these services. For instance, I&R providers may keep office hours so that help-seekers can visit and meet with trained specialists to track down a service that can meet their needs. I&R providers may also create public-facing online tools, where

help-seekers can browse for appropriate services. Most I&R providers (including 211s) offer a phone number that help-seekers can call or text to be connected with an I&R specialist.

I&R specialists are trained to be good listeners, problem solvers, information navigators, and communicators. In training, I&R specialists are informed that often help-seekers will not be able to articulate their exact situation or what services they need. Further, help-seekers may withhold certain information out of distrust for authorities or out of concern of being stigmatized in asking for help. Help-seekers may be emotional or angry, and they may not speak the specialist's language fluently. Further, I&R specialists are constantly presented with complex and sensitive situations. A person may call with an issue such as: I need to find a domestic abuse shelter that will accept myself and my children, but I need it to be open within a very specific time range so that my husband does not see me leaving. Or: I need to find a health clinic, but I'm an undocumented immigrant and do not have health insurance. Or: I need to get to a food pantry to feed my family, but I do not have access to transportation.

I&R specialists need to be skilled in collecting relevant information from help-seekers, sorting through help-seekers' complex situations, and finding appropriate resources in spite of knowledge gaps. Notably, offering the wrong information to help-seekers can potentially cause them a great deal of burden or harm. Because of this, many in the human service domain understand the "human" component of I&R to be essential to quality service delivery. I&R specialists are empathetic search engines that can navigate the complexity of social problems: "Real People, Real Help."

Still, finding the "right" service within a database populated with often thousands of entries is challenging even for a trained specialist. To support I&R specialists in tracking down the "right" services for help-seekers, AIRS has produced a number of standards and quality indicators designed to ensure that the information stored within referral databases is accurate, consistent, and timely and that data is indexed in such a way that specialists and help-seekers can easily find it (Alliance of

Information and Referral Systems, 2016).<sup>2</sup> For example, AIRS outlines which data elements are mandatory within referral databases (such as an agency's name, legal status, and phone number), and which are recommended (such as whether a site has a web address or access for people with disabilities). AIRS has also developed an open source data schema, which suggests a structure for organizing referral entries and their associated properties into databases.<sup>3</sup> For instance, the AIRS data schema suggests that "sites" have attributes such as "addresses," "times open," and "bus service access" and that "application processes" have attributes like "steps" and "descriptions." AIRS accredits I&R programs (including many 211s) based on their adherence to these standards and offers training and professional certificates to I&R specialists.

AIRS has been publishing standards like these since 1973, and as the primary professional association for I&R, most I&R organizations seek accreditation and certification through AIRS. Yet, if most organizations are building their resource directories according to AIRS standards, why does data sharing in I&R remain so difficult? One of the reasons is that AIRS began producing data standards before the concept of "open government data sharing" had proliferated in the United States and internationally. The standards were thus designed, not so much to ensure that the data was structured in such a way that diverse communities could use and interpret it but instead to guarantee "quality" within closed 211 systems. There is a great deal of flexibility for adapting the standards for the needs of a specific community, but there is less consideration for structuring the data to be "interoperable"<sup>4</sup> with diverse systems. Thus, I&R organizations still structure their data differently and describe their data according to different vocabularies. In the following section, I outline several reasons why there is such semantic heterogeneity in the human services and discuss why it makes designing semiotic infrastructures in this domain especially challenging.

---

<sup>2</sup>AIRS also publishes standards that suggest how I&R programs should be structured, governed, and staffed and how programs should prepare for disasters.

<sup>3</sup>AIRS has also invested in designing a linked open data vocabulary so that referral data may be shared and linked on the World Wide Web. However, through interviews I've conducted with the designers of this vocabulary, I've learned that there has been little uptake of the vocabulary.

<sup>4</sup>In this domain, data interoperability means having at least the minimum common characteristics needed for diverse data models to be mapped onto each other.

## 5.2 Semiotic Heterogeneity in the Human Services

In his blog "Human Service Informatics: Notes in Support of an Emerging Wholeness," Derek Coursen (2013) writes, "This blog starts from a simple premise: managing human service information is difficult because there are many voices at the table and they speak different languages." Coursen goes on to outline some of the voices at the human service table. There are service providers, program evaluators, social scientists, funders, software developers, and each of these voices can have different substantive focuses - "child welfare, homelessness, substance abuse, domestic violence, job training," for example. Coursen writes:

Each group of stakeholders has its own worldview and its own vocabulary. Each has its own agenda for collecting data and producing information. Interests overlap in some respects but clash in others. With perspectives so fragmented, building information systems and formulating information policy is fraught with complexity and risk.

Language in the human services has consistently been described to me as "overloaded" with meaning. Words in the human services tend to be polysemic - having different meanings depending on the context in which they are used. In what follows, I describe some of the reasons why language in the human services tends to be especially diverse and suggest why this makes sharing referral data especially challenging.

### 5.2.1 Semiotic Regulations, Bureaucracies, and Networks

Human services can only operate because philanthropists, government agencies, service providers, I&R organizations, and program evaluators agree to collaborate. The human services operate as a network of independent stakeholders, each with their own bureaucracies, incentives, and vocabularies. There is not a single body governing human services, nor is there a standard definition determining what is considered a part of the human services (Zins, 2001). In the United States, several different governmental bodies contribute to what is often classified under the heading "human services." The Department of Health and Human Services administers Medicare, Medicaid, mental health services, and clinics. The Department

of Housing and Urban Development administers shelters and emergency housing. The Social Security Administration administers disability and retirement benefits. The Department of Agriculture administers supplemental nutrition programs. The Department of Education administers after-school programs and special education services. The Department of Homeland Security administers programs responding to domestic disasters. The Department of Labor administers unemployment services and benefits. The list goes on. These departments have several sub-agencies and conduct separate operations at federal, state, and regional levels.

Since services are managed and administered in different federal, state, and regional departments - each with their own aims, perspectives, leadership, and regulatory contexts - there are often different definitions for concepts that discern eligibility for human services. Take the word "homelessness" for instance. For the Department of Housing and Urban Development, the Homeless Emergency Assistance and Transition to Housing Act of 2009 (P.L. 111-22, Section 1003) defines "homelessness." This definition lists four categories of homelessness: 1) people "living in a place not meant for human habitation," 2) people losing their primary nighttime residence within 14 days, 3) families with children in unstable housing situations, and 4) people fleeing domestic violence. Homeless assistance programs can be distributed based on whether an individual's situation aligns with one of these categories. However, for the Department of Health and Human Services, the Public Health Service Act (42 U.S.C., 254b) defines "homelessness." This definition is a bit broader than the HUD definition; an individual is considered homeless if (s)he is living in any sort of unstable housing situation (including hotels, motels, shelters, or situations where they are "doubled up" - living with a series of friends and family members). Health centers use this definition to determine eligibility for social services.

Different agencies and departments also use different vocabularies to describe the structure of human services. Words like "organization," "program," "project," and "service" all mean different things to different departments. Greg Bloom told me during an April 2017 interview that since service providers receive their funding from these different agencies, they often have to mold their definitions towards whatever

definitions their primary funders use. This means that, amongst service providers, there is also a great deal of semantic heterogeneity for describing eligibility criteria and for structuring programs.

Most I&R databases list services offered by a variety of service providers, funded by a variety of governmental programs and donors. I&R databases are sites of federation, bringing together human services that would otherwise be accessible only through silos. Thus, I&R databases must accommodate heterogeneous perspectives and vocabularies. As Bowker (2000) has argued, databases designed to integrate diverse data tend to gloss over semantic heterogeneity. They "converge" on particular representations of the world as they order data from multiple sources and representing multiple perspectives. Yet, different I&R organizations converge on different representations of the world within their databases. There is no central body controlling how information gets added to, deleted from, or structured within I&R databases. While the 211 dialing code is federally standardized across the U.S. and Canada, 211 programs across both countries operate independently. The structure of 211 programs can thus vary drastically from region to region. In most cities across the United States, dialing 211 connects citizens to specialists at their local United Way. However, in Memphis, dialing 211 connects citizens to the Memphis public library. In Miami, dialing 211 connects citizens to the Jewish Community Services of South Florida. And in perhaps the most unique situation, dialing 211 in New York City reroutes the call to 311 - a call center established through the city's mayor's office to field calls about non-emergency related municipal issues (such as noise complaints, potholes, etc.).

And, of course, there are many I&R initiatives that operate independently from 211 entirely. Today, there are I&R centers in hospitals, schools, libraries, and other community centers. Within these different contexts, I&R organizations have different degrees of capacity to implement technical solutions. Some organizations use sophisticated software (such as iCarol and Aunt Bertha) that has been designed specifically to manage referral data. Others will list their referrals in Microsoft spreadsheets. Thus, there is no one standard way to federate semantically heterogeneous referral data. I&R databases look different and represent different

perspectives and bureaucratic genealogies in different contexts.

### 5.2.2 Semiotic Styles, Habits, and Commitments

During an October 2016 interview with me, Teresa Pardo, Director of the Center for Technology in Government, described why she believes criminal justice departments have been so far out ahead of other governmental departments in adopting data sharing protocols and infrastructures. One reason is that they have to be; when an officer arrests an individual in one state, the officer needs to be able to access fingerprint databases and arrest histories in other states. Criminal justice, Pardo argues, cannot be siloed by geographic boundaries. "The other reason that they may be out in front," she argues, is that:

It's a real command and control environment. It's that kind of para-military [environment], and there's a rule and a standard and a book about everything. So it was a comfortable space to be saying, "What are the rules? What are we all doing together? We have to follow the rules." So the kind of para-military command and control structure facilitates that kind of creation of a shared resource. It doesn't always create consensus, but it does motivate and incentivize people to participate in ways that, if they had more choices, they may not otherwise do.

There are cultural grounds for the ways different data sharing communities think about, value, and use language. Different data sharing communities have different degrees and styles of commitment to aligning their vocabularies. As Pardo notes, individuals in criminal justice fields tend to be especially committed to using the same words in particular contexts because, within criminal justice fields, there's already a culture of privileging of rules over liberties and order over accommodation. However, in the human services, individuals tend to think about, value, and use language in much less formal ways. Derek Coursen, an information systems designer with an expertise in data modeling in the justice, human service, and public health sectors, described this to me in a June 2017 interview:

... in the hard sciences people are very, very, very careful to, or they attempt to do work sharing words to make sure that everyone is on the

same page about what the word means. I know this because I grew up in a family with a lot of hard scientists around, and ... the human service sector ... is sort of in this funny place between kind of administrative sciences like public administration or business administration and soft sciences where there just isn't the same kind of respect for language. And people will make up ... [or] use language for essentially rhetorical or political purposes and not think that there is something counterproductive about that.

This is in part because the human services itself is made up of such diverse stakeholders, each with their own language habits, assumptions, and commitments. The cultures of language represented in the human services are quite diverse. There is not a command and control culture motivating, incentivizing, and, frankly, demanding the use of shared standards in the human services, and there is not a para-military culture of conformity.

Using common vocabularies in common contexts also tends to be less of a formal commitment in the human services because words in this domain tend to be highly politicized both within the domain and within broader public discourse. Politicians, government administrators, service providers, and citizens leverage words like welfare, health, poverty, and abortion towards specific, divergent, and often conflicting political ends. There are often political reasons why certain communities (including services providers and donors) do not want to commit to specific definitions of these words. Further, formal definitions of words like welfare, health, poverty, and abortion change often in new political and regulatory climates. For instance, since the early 2010s, standard definitions of marijuana as an "illegal" substance have changed as certain states have legalized recreational and medical marijuana. Standardizing semiotic infrastructure for ordering and representing referral data tends to gloss over the diversity of language ideologies that characterize the human services. As a result, different I&R databases tend to not only be technically non-interoperable, but also ideologically non-interoperable.

### 5.3 "Sorting" Semantic Differences with Semiotic Infrastructure

Most semiotic technologists in the human services, and particularly in I&R, view the field's semantic heterogeneity as a problem that needs to be overcome. For one, they see it as getting in the way of directing help-seekers to the right service. They also see the field's diverse language as making it almost impossible for researchers and program evaluators to assess how well service providers are meeting the needs of people in their communities.

The Health Communication Research Laboratory at Washington University in St. Louis created and launched the website [211counts.org](http://211counts.org) in 2014. For most states in the U.S., the website displays data about what sorts of issues citizens are reporting when calling 211. Navigating to [211counts.org](http://211counts.org) in November 2017, the first thing I saw was a map of the United States. States reporting their 211 data were shaded in green. When I clicked on the state of Massachusetts, I was directed to a page that listed the top request categories of calls the Massachusetts 211 had received for the past year. 26.8% of calls related to child care and parenting; 17.6% of calls related to mental health and addictions, and 15.5% of calls related to housing and shelter. I could click on any of these categories to get a more detailed breakdown of the issues relating to these calls. For instance, when I clicked on Mental Health and Addictions, I found that 83.7% of those calls related to crisis intervention and suicide and 10.8% of those calls related to mental health services. I could also examine a map that displayed where the majority of calls in the state came from. Hovering over my hometown - Blackstone, MA - I could see that ten 211 calls were made from this zip code in the past year. I could also compare this to census data about my hometown, including its poverty rate, unemployment rate, high school diploma rate, and rental housing rate. On the About Us page, the site's creators note that the primary audiences for the platform are government representatives, service agencies, and philanthropies.

However, different 211 call centers use different schemes for coding their

calls into the top request categories. So while 211 counts offers a detailed snapshot of what issues people are calling about in a specific community, it would be difficult for the system to compare the issues of communities serviced by different call centers.

Semantic differences in I&R databases also make it difficult for I&Rs to collaborate during national disasters - a time when human service needs are particularly complicated and telecommunications networks are particularly vulnerable. For instance, during Hurricane Katrina, all of the phone lines in and around New Orleans were down, and nearby United Way offices had to step in to answer phone calls (Strom, 2005). During Hurricane Irma and in the wake of the October 2017 Las Vegas shootings, the call centers in Florida and in Las Vegas were so inundated, that 211s across the country were asked to step in to answer calls directing people to services. Further, many services are interrupted during a disaster: shelters are underwater; transportation routes stop operating, and the employees running non-profit counseling services cannot make it to work. I&R specialists in diverse organizations and locations need to be able to understand how entries are organized and described in I&R databases in order to be able to help people find services in these particularly complex situations.

I have found that semiotic technologists in the human services often approach their work with an assumption that the diversity of meanings and definitions can be unified - that polysemy can be sorted out. Typically they position semiotic infrastructures as tools for federating semantic differences in the domain - first, because they believe that the design of semiotic infrastructures requires modeling how things "actually are" in the domain, and second, because they argue that semiotic infrastructures institute common languages in the field.

In a June 2017 interview with me, Derek Coursen described how designing semiotic infrastructures could help stakeholders in a domain figure out the "hidden structure" of the domain and "what's really going on underneath the surface":

...there are plenty of domains where the language itself is so flawed and the semantics are so muddled that what has to happen is somebody has to walk in and sit down with all of the stakeholders and help them to

pick up their language that they are using and look at it from different angles and figure out what they're actually trying to say and then reformulate the language. And it may involve actually developing entirely new concepts that were hidden in the previous muddled language. And you have to look at what's inside - what's hiding in that language - what's the hidden structure of it. And pull it out and name it to the stakeholders and say this looks like what's really going on under the surface, and if they can then validate that, then you have the basis for building something. But by the time you've exited, you haven't merely built an information system, you've changed everyone's language.

In other words, designing semiotic infrastructures is often understood to homogenize semantic heterogeneity because it models a higher order of meaning. Many semiotic technologists in the human services understand there to be a unifying logic hidden within the domain's muddled language. Their work, then, involves disentangling muddled semantics so that meanings, concepts, and relationships are represented "logically." In what follows, I narrate the design of the AIRS/211 LA County taxonomy - a project that attempted to establish a common or standard language in the human services by ordering and defining concepts according to their "logical niche."

#### **5.4 Finding a Term's Logic Niche: The Design of the AIRS/211 LA County Taxonomy**

The AIRS/211 LA County taxonomy is a classification system for I&R data. The taxonomy acts like an encyclopedia for I&R concepts; within the taxonomy, terms and concepts that can be used to classify and index referral data are defined. Specialists creating I&R databases can use the terms defined in the taxonomy to describe database entries, so that it is easier to locate and retrieve the data at a later date. The AIRS/211 LA County taxonomy includes terms that describe types of service organizations (e.g. libraries and clinics), types of services offered (e.g. counseling and housing), or types of people served (e.g. veterans and children). Today, the taxonomy has more than 10,000 terms defined within it.

Georgia Sales has been managing the AIRS/211 LA County taxonomy since the early 1980s - two decades before most 211s existed. At the time, Sales worked for Info Line - one of the primary I&R non-profits in Los Angeles. She also held a Master's degree in Analytic Philosophy - an education that she described to me as giving her a unique expertise in solving puzzles. Sales described to me during an April 2017 interview how she went about creating the taxonomy from scratch, how the taxonomy became a standard in the field, and how she has worked to maintain it over the past several decades. For Sales, building a hierarchical structure for terms in the human services ensured that every concept would have a "logical niche" - that hierarchies would have increasingly specific sets and subsets and that people would know where to find concepts with them. In this section, I outline how Sales went about dividing, defining, and organizing I&R concepts in order to demonstrate what it means to her to find a term's "logical niche."

#### **5.4.1 Bracketing, Ordering, Scoping, and Re-ordering**

Prior to the development of the AIRS/211 LA County taxonomy, I&R groups across California had all been using different variations of the category scheme in the California Resource Information Bank (CRIB) for classifying human service concepts. At the time, the disorganized structure of CRIB had been the source of a great deal of frustration for I&R groups. Sales noted that CRIB had too many basic (or highest level) categories, so people didn't know where to look for concepts. In addition, the basic categories "represented different perspectives." Amongst the basic categories, some terms described a type of service, while other terms described the type of group that the term served. As a result, often concepts nested within these basic categories would overlap. For instance, the concept "charter school" would fall under both "youth serving programs" and "educational programs." For Sales and others, this meant that the concepts within CRIB did not have a "logical niche." At one point, representatives from several I&R programs convened discussions to "fix" the structure of CRIB. However, they eventually decided that, rather than trying to fix the state's category structure, they needed to develop their own - what would become the AIRS/211 LA County taxonomy.

The early work in designing the taxonomy involved pulling definitions from existing semiotic infrastructures and associating differing definitions for the same concept. Sales described:

What I did as the first step was a lot of research. What are the existing human service classifications systems and what were their similarities and what were their differences? I began making lists of concepts that came out of those explorations and also notes on the side about how some of the other systems were structured. I looked at references like the United Way of America Services Identification System (UWASIS) and the CRIB (California Resource Information Bank) system that we had actually been using. But I also reviewed resources like Black's Law Dictionary, the most current Diagnostic and Statistical Manual of Mental Disorders (DSM) that was available, and Taber's Cyclopedic Medical Dictionary. I described the process that we went through as understanding the structural goals of the system we were aiming to create and "vacuuming" existing resources for potential term concepts. As my research progressed I noticed that people were sometimes using different words to describe the same concepts. As I was writing up my notes, I would add an equals sign between words that looked to me like synonyms and place them on the same line.

Based on her research, Sales went on to make decisions about the structure of the taxonomy. She designated 10 terms to serve as the very basic categories organizing the rest of the taxonomy. Sales told me that, aside from having to make a few adjustments to align with minor changes to language in the DSM, they had gotten these 10 general categories "right from the beginning."<sup>5</sup>

After distinguishing the 10 basic categories, she began assigning terms to these "buckets," organizing all terms into a hierarchy with 5 levels - each subsumed level

---

<sup>5</sup>Today, these 10 basic categories include 1) Basic Needs, 2) Consumer Services, 3) Criminal Justice and Legal Services, 4) Education, 5) Environment and Public Health/Safety, 6) Health Care, 7) Income Support and Employment, 8) Individual and Family Life, 9) Mental Health and Substance Abuse Disorder Services, and 10) Organizations/Community/International Services. One additional category has been added – 11) Target Populations, a taxonomy describing the populations that certain services may serve.

signifying increasing specificity. For instance, under "Basic Needs" (level 1) you would find "Food" (level 2), and under that you would find "Emergency Food" (level 3), and under that you would find "Food Pantries" (level 4), and under that you would find "Occasional Emergency Food Programs" (level 5). Sales convened a review group of I&R specialists throughout California (the same individuals that had participated in discussions to fix CRIB) to assess how well they had divided concepts into hierarchies. After settling on a hierarchy, Sales began defining the words in the taxonomy. She noted that, in the process of defining the words, they discovered that they needed to move about 30% of them to other sections:

When we read the definitions we wrote for the concepts, we found that some weren't really a subset of the higher level category to which they had been assigned and needed to be moved. For me, that shows how important it is not only to have recognizable term names that most people would understand, but also definitions to identify the scope and define very specifically what a particular service entails.

For Sales, by logically dividing hierarchies and categories, in addition to logically scoping term definitions, the spot where terms belonged in the hierarchy would logically be revealed. Sales and her collaborators believed that getting this structure "right" would help align how different I&R groups in California classified their data.

#### **5.4.2 Discerning Authoritative Language**

The first edition of the taxonomy was published in 1987 as print pamphlet distributed to I&R programs throughout California. Soon after this, Sales attended an AIRS annual conference where the attendance was small enough to convene a discussion-based plenary session with all attendees and AIRS Board Members. She recalled to me:

I remember vividly, one woman standing up and saying, "We have a huge problem in our field; everybody is using a different classification structure for services. We can't talk to one another. We can't compare information about what people are requesting, whether communities have services in

place to meet the needs of what people are calling about and where there are gaps in services." And she said, "Can't AIRS do anything about it?"

For the second and third editions of the printed taxonomy (published in 1991 and 1994 respectively), Sales and her team at Info Line LA partnered with AIRS, transitioning the taxonomy towards a national standard for indexing I&R data. This, of course, required that the scope of the taxonomy be expanded to include programs that may not be as prevalent in California (such as snow removal). The first Web version of the taxonomy was published in 2004, and the first Canadian version of the taxonomy was published in 2005.

With these expansions, the taxonomy not only had to sort diverse departmental perspectives about how human service concepts should be defined and ordered; it also had to do so across state and even national borders. To help maintain a common taxonomy across this international community, AIRS developed an online forum where AIRS members could make suggestions about additions or changes to the taxonomy. Sales noted that this community has been vital to the taxonomy's upkeep - that specialists "on the ground" have a better sense of what new services are emerging and serve as an "early alarm system" for how the taxonomy will evolve.<sup>6</sup>

However, the online forum also brought several more voices into the mix with diverse perspectives about how human service terms and concepts should be defined and ordered. In order to continue scoping and organizing terms according to their "logical niche" in the face of this increased diversity, Sales described to me having to discern when language that various stakeholders wanted to see in the taxonomy was "permanent" versus when it represented a political fad:

...for me the biggest challenge is making sure that a language change is permanent, not a fad, and has been accepted by the various different authorities in the field. If you're looking at medical terminology, you want to see how the Centers for Disease Control and Prevention are defining the concept and you want to make sure that you are avoiding sources that want to change the vocabulary in order to push a particular

---

<sup>6</sup>For instance, over the past year, there have been many requests for taxonomy terms that classify services responding to the opioid crisis, which has become prevalent since the mid-2010s.

perspective. I once had somebody who objected to the concept of obesity because she felt that it was insulting to people who were overweight. And I had to think about that. You can never reject a question that you get from someone out of hand. Because there's reason they're asking. But what I had to tell her is I'm really sorry, but obesity is a public health problem. It's not a comment on somebody's appearance.

For Sales, changes to language in the human services can be considered permanent when experts have vetted them and when a variety of individuals and organizations are using them. Sales conducts a great deal of research in order to discern a term's "permanence." She described to me that she spends a great deal of time in libraries, researching how various "authorities" define human service terms. Sales also maintains a list of "expert" sources, which she relies on to make authoritative decisions about changes to the language:

I still have informants - "go to" people with expertise that I contact when I have questions as well as authoritative documents like the DSM that are a "court of last resort" when deciding whether we need to make a change. Do I have it right? What is the updated language?

Sales maintains a library of pamphlet files for several human service organizations, which she references to get an idea of how others are defining and using concepts in the human services, and at times, Sales calls practitioners to find out more about what they do. Finally, Sales doesn't make final decisions about additions and changes to the taxonomy alone. Through AIRS, Sales chairs a Taxonomy Committee, comprised of resource specialists. This group meets 10 months out of the year to review suggested updates to the taxonomy and to advise on how to respond to contentious suggestions. Sales has maintained an "Acknowledgements" document where she outlines in bibliographic form, all of the sources she has referenced in building the taxonomy, as well as individuals that have contributed to the review of the taxonomy. By October 2017, this document had grown to over 120 pages long.

In characterizing certain language in the human services as more "permanent" than other language (and in characterizing certain groups and individuals to have more "expertise" on language than others), Sales signifies her assumption that there are certain modes of defining and ordering terms in the human services that are more correct and objective than other modes. Designing the taxonomy is understood to help federate language in the human services both by making definitions considered to be "authoritative" more visible and by encouraging everyone in the community - across both departmental and geographic divides - to align their language with these authoritative definitions.

### **5.4.3 Translating For Diverse Communities**

While certain signs are privileged as the authoritative terms for signifying a human service concept in the AIRS/211 LA County Taxonomy, Sales also acknowledges that different communities will use different words to define the same thing. To address this, the taxonomy also lists over 30,000 synonyms, or "use references" for words defined within the taxonomy. In designating certain words to be synonymous with indexed terms, a user or I&R specialist does not need to know the exact term defined within the taxonomy in order to find an appropriate database entry. For instance, if a user were to search an I&R database for "food shelf program" or "grocery pantries," the user would be directed to database entries that had been classified as "food pantries" because "food shelf program" and "grocery pantries" have been defined as synonymous with "food pantries." All of these synonyms share the same definition in the taxonomy.

In this sense, through the taxonomy, Sales and her collaborators have attempted to sort out the field's polysemy by linking diverse signs to their common meaning. The taxonomy has served not only as an authority on how terms are defined and ordered in the human services, but also as a translating device, policing sameness and difference. Use references are linked to words that have already been organized into their "logical niche." So for Sales and her collaborators, even if different individuals and organizations are using different words to describe the human services, as long as they link their synonyms through use references, the

logical structure of this common language will remain in tact.

#### 5.4.4 A Standard Language for the Human Services?

In training material Sales has developed for I&R resource specialists, she begins:

When studying the Taxonomy keep an open and accepting mind. It might make sense after 2 hours or 2 days but rarely straight away. Don't fight it by expecting it to fit into an image that you are carrying in your own head about how you think it should work.

In learning to leverage the taxonomy to index data in I&R databases, resource specialists are often not only learning a new technology; they are also learning a new language and a new way of understanding the field in which they work. The semiotic infrastructure has been "scripted" to suggest to its users a particular way of organizing their databases, their language, and their work.<sup>7</sup> It has been scripted to enforce a common language for the human services.

In October 2017, the AIRS/211 LA County taxonomy website listed a number of testimonials endorsing the taxonomy, and perhaps the most prominent acclamation for the taxonomy is that it created a "common language" in the human services, allowing I&R specialists in different locations to talk to each other and share their data. However, today, not everyone in the I&R community is using the AIRS taxonomy to index their data. One of the reasons for this is that the taxonomy is proprietary. Users have to pay to subscribe to the taxonomy - a fee that helps defray the cost of having a full time staff member maintaining it. Further, when using a controlled vocabulary to index services in a database, people querying that database need to be able to anticipate an exact term or phrase (or its "use references") in order to locate a relevant service. While I&R specialists can be trained to

---

<sup>7</sup>Madeleine Akrich (1992, 207-208) has argued that designers, when defining the affordances of an innovation, form an imagination for their technology's future users - the users' "specific tastes, competences, motives, aspirations, [and] political prejudice," and designers "assume that morality, technology, science, and economy will evolve in particular ways." Throughout the design process, they "inscribe" the technology in accordance with this vision. The product thus comes to represent an instruction manual or recipe for a user that eventually adopts the technology. In "de-scribing" a technology, a user can either adhere to or veer from its script.

find these concepts within the hierarchy, a help-seeker searching a database online may not know the exact word or phrase that they need to enter in order to track down the correct service. For instance, a help seeker may type the phrase "reduced lunch" into a search box and not find anything because the official term for student lunch programs in the AIRS/211 LA County taxonomy is "School Lunches/Snacks," and the phrase "reduced lunch" has not been designated as a use reference for the term.

To address these challenges, Aunt Bertha, a corporation that develops software for referral organizations, launched the Open Eligibility taxonomy in 2010 - a completely separate taxonomy from AIRS/211 LA County's that is open access and attempts to define the human services with simpler terms. Schema.org<sup>8</sup> has also developed a separate taxonomy for describing data on "Civic Service" web pages. So while the aim for developing taxonomies for the human services was to create a "common language" so that everyone would index their data consistently, as more standards have emerged, the common language continues to splinter. In what follows, I show how the Open Referral project responded to this challenge, at first attempting to logically model the diversity of meanings in the human services but eventually taking an approach that used the term "standard" in air quotes.

## 5.5 Creating an Interlingua for the Human Services: The Design of the Human Services Data Specification

The Open Referral website describes the Human Services Data Specification (HSDS) as "essentially an interlingua – in other words, it's a common language that can be used by anyone to enable community resource directories to 'talk' to each other" (Bloom 2015). The Open Referral project published the first version of HSDS in March 2015 and published an updated version in April 2017. The data specification was designed to enable making referral data open, accessible, and exchangeable without requiring that everyone use the same vocabulary.

Bloom launched the Open Referral project in 2014 - seeking to design the information infrastructure needed to support common pool community resource di-

---

<sup>8</sup>See Chapter 4 for more on Schema.org.

rectories.<sup>9</sup> Inspired by Elinor Ostrom's work on common pool resource management, Bloom believed that with proper governance and a distribution of responsibilities amongst stakeholders, referral data could become a common pool resource for I&R organizations.<sup>10</sup> For Bloom, making referral data a community resource would cut considerable costs for I&R organizations, make it easier for help-seekers to find the services they are looking for, and make it possible for evaluators to assess how well I&R organizations are meeting diverse needs across communities. In other words, it was a win-win-win situation. However, Bloom was skeptical of systems that attempted to curate all referral data into one central clearinghouse. He instead imagined a situation, much like the World Wide Web, where resource specialists could interlink distributed resources into a community resource directory by structuring their data according to a simple set of protocols:

...talking with folks in the D.C. area about open data and civic technology, I developed a vision about what would it look like for this system to be accessible through a platform or through an ecosystem rather than in the old approach, which never worked, where people tried to build the central clearinghouse one stop shop. ... So rather than to be in between the status quo of all fragments, redundant, ineffective silos and the impossible utopia of the centralized clearinghouse one-stop shop, we sort of envisioned the sort of practical, potentially realizable outcome where this data just flows among many systems, like through a platform or a network.

---

<sup>9</sup>Bloom's official title with the Open Referral project is "Chief Organizing Officer," a role that involves facilitating negotiation and consensus amongst the stakeholders that have become involved in the Open Referral project.

<sup>10</sup>The main precursor to Open Referral was a project called Open211. Open211 was launched in 2011 as part of the first round of Code for America fellowships. Noting that their Bay-Area 211 program did not have the infrastructure (more specifically an Application Programming Interface or API) for sharing its data, the designers of Open211 scraped data from a number of referral sources and loaded them into a brand new application for querying referral data. The designers of Open211 expected that community members and service organizations would add to and update the data in the application. However, this did not happen. Writing on the Open211 project, Greg Bloom (2013) noted that many 211 administrators predicted this crowd-sourced approach to the Open211 application would not work. Collecting, organizing, and updating referral data is considerably labor-intensive. This is why 211s hire research teams to do the work. With no one having the designated responsibility for contributing and maintaining data in the Open211 application, the project was largely abandoned after a year.

The Human Services Data Specification (HSDS) defines these protocols. The Specification outlines how I&R data should be formatted so that it can be exchanged between human service directory information systems and third parties. It defines types of data objects (e.g. programs, organizations, services, and locations), their fields (e.g. name, email, legal status, and latitude), and their relationships (e.g. services are linked to organizations and phone numbers can be linked to locations). When resource specialists format their data according to the Specification, they can more easily share their data with each other and with third parties because diverse database environments can anticipate how resource specialists are structuring I&R data. Formatting data according to the Specification also streamlines the development of platforms that can curate data from many different I&R programs. In September 2017, Socrata, the most popular software for storing, managing, and sharing open government data, announced that they planned to adopt the Human Service Data Specification in their platforms (Bloom, 2017).

### 5.5.1 Modeling Semantics Relationally and Contextually

In early discussions, the Open Referral team considered whether they should design HSDS into an ontology for the human services. Many on the Open Referral team acknowledged that many different stakeholders were using many different vocabularies in their work and that it was unlikely that this was going to change. Ontologies are often understood to tolerate semantic heterogeneity, however. Ontologies standardize the logic of a domain, not the language of a domain. Ontologists understand "things" modeled within an ontology to have meaning, not because they have written-out formal definitions attached to them, but instead because their relationships to other "things" are formally modeled. This means that it does not matter if five different groups are using five different words to describe the same career readiness service, as long as they all know its address, its legal status, the organization that it operates within, and by what criteria an individual is eligible for it. These characteristics make a "career readiness service" meaningful, and ontologies model the relationships between these types of characteristics. The rationale for designing HSDS as an ontology was that people could come to understand what the

career readiness service *is*, not according to the definition that had been attached to it, but instead according to all of the data surrounding it.

Furthermore, using the Web Ontology Language (OWL), when designers need to use words to label certain relationships between concepts (e.g. that an "organization" offers "services" which have a "location"), they can separate out how these structural words are defined differently in different contexts. For instance, one issue that came up often in the design of HSDS was that different stakeholders in the human services would use the words "program" or "service" interchangeably to refer to the same concept. This meant that different stakeholders had different understandings of what constituted a "program" and what constituted a "service." When using one of these words as a label for data in an ontology, the ontology could specify the context in which the word should be interpreted by labeling the word with a particular *namespace*.

A namespace designates a vocabulary used within an ontology. When building an ontology, an ontologist will often begin by associating links to Web vocabularies with particular variables. So maybe there is a vocabulary on the Web for how the U.S. Department of Housing and Urban Development defines their concepts, and a separate vocabulary for how the U.S. Department of Health and Human Services defines their concepts. An ontology designer may associate the links to those vocabularies with the variables "hud" and "hhs" respectively, creating a namespace for HUD and a namespace for HHS. Then, whenever a word is used in the ontology, the ontology designer can prefix the word with the variable "hud:" or the variable "hhs:" designating the context in which the word should be interpreted. If the ontology was modeled to suggest that "services" are part of "organizations," the word "service" could be prefixed with "hud:" to designate that anyone using the ontology should assume that the word "service" is being used here according to how the U.S. Department of Housing and Urban Development defines "service." In this sense, using an ontology to model the human services, the Open Referral team would not need to get everyone using even structural words consistently. Instead they could logically model the differences between how different users defined different words.<sup>11</sup>

---

<sup>11</sup>Ribes and Bowker (2009, 12) have described how knowledge representation experts designing ontologies for the earth science community purported that they could "represent the uncertainty"

Almost paradoxically, they would endure the ambiguity of words and definitions in the human services by modeling the context of words and definitions in the domain with a lot more specificity. An ontology presumed that there was a unified hidden logic to how the human services are structured and that polysemy could be parsed by revealing it, deciphering, and encoding it.

### 5.5.2 Using the Term "Standard" in Air Quotes

Greg Bloom told me in an interview that they decided not to develop HSDS into an ontology once he asked around and learned how difficult it was to adjust ontologies once multiple communities began using them:

I dropped the whole thing when I realized, oh, we could put all of this work into developing an ontology and then once it's out there we can't change it. Like once people start using it, we can't go back and adjust stuff. And I'm just like wow that is an existential nightmare because fundamentally these domains change.

Ontologies are difficult to revise *because* concepts within an ontology have meaning based on their formally defined relationships to other concepts. Changing one relationship in the ontology can impact how all other concepts in the ontology relate to each other; it can create cascading semantic inconsistencies. Instead, the Open Referral team decided to develop HSDS into a schema that "logically modeled" the relationships between I&R data. In some ways, they kept the spirit of an ontology in the architecture of the Specification. While there were not formal semantics describing the relationships between different data objects, such as "organizations" and "services," the databases were set up to show that there were links between these objects. For instance, in the database for "services," every entry was required to include an ID to an entry in the "organization" database, indicating that the organization associated with that ID *provided* the service.

They also kept the spirit of a namespace within the architecture of the Specification. HSDS enabled resource specialists to indicate which taxonomy they use to with semantic modeling, "encoding and representing disagreement, uncertainties, ambiguities, or ambivalences." They note that this was a "promised capability" but that they never saw it carried out in practice.

index their services (such as AIRS/211 LA County, Open Eligibility, or Schema.org) in a taxonomy database. Each service could be linked to different terms in the taxonomy database, enabling users to trace back to where the term had been defined. As Greg Bloom told me, this "sidestepped" the taxonomy problem in the human services:

We've developed a spec that's just describing the facts about organizations - the facts about the services, the contact information, but the subjective description about what is this service and the much more variable description about who is eligible to be served - those are things where...our spec is like "insert your model for that here."

Instead of holistically modeling all the different components, contexts, and perspectives of the human services into one common model, the Open Referral team decided that they should make their model "deliberately minimal" - that the Specification should model a minimum set of data objects, relationships, and fields common across the human services. They suggest that Open Referral only encodes "factual data" about organizations and their services (such as their address and their hours of operation) and that the Specification does not prescribe a specific taxonomy because the categories within such taxonomies are "inherently subjective." And while I find this distinction between "factual" and "inherently subjective" data to be arbitrary, I am hopeful that this lightly structured semiotic infrastructure can advance work in the human services - because, while it does prescribe a common vocabulary, it purposefully works to limit that prescription to just what's needed to bring about coordination within a certain knowledge domain. Sticking to encoding just this "factual data" means that not everyone has to approach their work with a common data model in order to share data. The light structure brings data together in spite of its semantic heterogeneity - potentially doing, for the human services, what HTTP did for the Web and what structured analytics did for PECE. Bloom and the team at Open Referral are also committed to ensuring that the model can change dynamically as the human services (inevitably) evolve.

This, I believe, is why the Open Referral team refers to HSDS as an "interlingua" for I&R data. When I asked Bloom the process by which he saw Open Referral

becoming a data standard, he told me that he tends to use the term "standard" in air quotes. He told me that while often people believe that designing a standard meant designing the "killer app" - the thing that everyone will use - he instead aimed for Open Referral to "build a practical consent." More than just building technical infrastructure, designing HSDS was also about starting conversations - "getting them to recognize their common interests and building their capacity to work together, which is very different from either the killer app, or the industrial legislature of old school standards." In other words, designing Open Referral was not about finding a logical hidden structure to the human services, sorting out everyone's polysemy, or designing the one "killer app" data model. Instead it was about enabling exchange and cooperation amongst diverse stakeholders with diverse languages.

This conception of the term "standard" as in air quotes marks its use in this context to be catachrestic - a forced imposition of a sign on a meaning that is never fully closed because it is deliberately "inappropriate." Standards, understood as common ways of doing things, never become permanently and exhaustively standard. Standards iterate and disseminate just as language does. However, to support individuals in need in finding appropriate human services, the I&R field really does need a common language - they really need precise definitions and shared semantics because directing individuals to the wrong services can cost help-seekers their livelihoods and their lives. Bloom uses the word "standard" in air quotes because developing language standards is necessary in the human services, but it is also impossible. HSDS, as an "interlingua" for the human services, creatively endures this double bind by focusing on creating the conditions for local exchange and cooperation rather than designing the infrastructure into a holistic translation device.

## 5.6 Strategically Encoding Semiotic Infrastructure

In the wake of growing international inequalities and increasingly common man-made natural disasters that devastate communities, there is an urgent need to better help people find help. To respond to the complexities of contemporary problems, I&R needs to be able to better coordinate across its silos and to precisely and consistently exchange information. In this domain, the stakes are incredibly

high to design semiotic infrastructure that enables stakeholders to understand each other and to understand the information that they share. And just like the reader of this dissertation can only understand what I'm writing on this page because, at some point in time, semiotic infrastructure has impacted the way we similarly interpret the meaning of the words I use, semiotic infrastructure will be necessary to advance communication in the human services. What should this semiotic infrastructure look like? What approach to semiotic infrastructure design is most appropriate for advancing the human services?

Recognizing that the vast and diverse set of stakeholders constituting the human services are unlikely to reach consensus on words, definitions, and ways of ordering I&R data (and that the semiotics of the human services are likely to remain messy, politicized, and decentralized), some semiotic technologists in this community have attempted to engineer semiotic infrastructures as translation devices. Often, this has involved attempting to reveal and encode the logic behind the domain's semantic diversity. These semiotic technologists presume that there is a hidden structure to the human service's language and that encoding that structure can align semantic inconsistencies. They presume that in modeling this hidden structure, the engineered infrastructure alone can do the translation work.

Translation, in this sense, is understood to operate according to a structural-functional semiotics. In other words, the language ideology guiding this approach assumes that, within social systems, meaning has internal cohesion - even if the signs used to signify that meaning are not cohesive. It assumes that the relationships between signs and what they signify are purposeful, logical, and designed to support the functioning of their respective social systems. Semiotic technologists designing according to this language ideology believe that if they can discern the logic to these sign/signified pairs, then they can model the internally cohesive meaning that holds the system together and enables the system to function. From here, semiotic technologists operating according to this language ideology have a few design options. First, they can design a semiotic infrastructure that forces everyone to change the signs that they are currently using to a shared set of signs that align with the field's common meaning; this typically manifests as a taxonomy. Alternatively, they can

design a semiotic infrastructure that models the relationships between signs and what they signify in that domain; this typically manifests as an ontology. And while these two semiotic infrastructures look different, demand different types of engagement from stakeholders, and operate in different ways, they both have a similar language ideology interwoven through their design - one that presumes meaning to be logical and fixed.

Yet, feminist scholars have argued that structural-functionalist approaches to semiotics ignore the *power relationships* that produce meaning within semiotic systems (de Lauretis, 1984). In assuming the relationship between signs and what they signify to be purposeful and logical, structural-functionalist approaches eclipse the role that entrenched conservative politics and oppression play in the design of sign systems. These approaches ignore the role that sign systems play in enacting a "real world" where those that do not fit into any of the formally specified categories are excluded from accessing help.

In the human services there are notable consequences to eclipsing these power relationships. I&R specialists are constantly responding to help-seekers that describe complex problems for which there are not yet formal definitions or even meaning in the human services. They respond to individuals that do not identify with the formal meaning encoded into the human services' sign systems - either because the systems have historically ignored their difference, or because, as Virginia Eubanks (2018) eloquently describes, many digital systems in the human services are purposefully designed to "profile, police, and punish the poor."<sup>12</sup> This is the value of maintaining the human dimension ("real people, real help") in I&R. It ensures that help-seeking does not get relegated to automated decision-making tools designed according to formal sign systems that cannot account for proliferating differences.

In this sense, we need semiotic infrastructures in the human services that are at once more durable and resilient - to ensure that specialists can efficiently

---

<sup>12</sup>Eubanks (2018) powerfully demonstrates how, since the 1970s, algorithms and other forms of automated decision-making have been used to determine a person's eligibility for human services. Through extensive historic and ethnographic research, she demonstrates how, hidden beneath layers of acclaim for reforming welfare fraud and removing the human bias from eligibility decision-making, these systems are in fact designed to surveil the poor, to criminalize the poor, and to dismantle the systems that ensure their rights.

communicate, collaborate, and direct help-seekers to the right service - while also more flexible and dynamic - so as not to preclude representing the complexity and variability of contemporary problems. The human services is a prime example of a domain where semiotic infrastructures need to be both neat and scruffy, where they need to be both restrictive and complex. Like on the Semantic Web (but with a recognition of the heightened stakes), the semiotic technologists in this community are grappling with how to work in messy middles - how to push the limits posed by these tradeoffs.

This, I believe, is where light structure, as a design strategy, becomes powerful. Light structures are like "standards" in air quotes. In a lightly structured infrastructure, certain structuring components of the infrastructure are temporarily hardened - not because there is a presumption that the hardened parts of the structure are logical or fixed - but instead to strategically and momentarily enable coordination

amongst collaborators.<sup>13</sup> Designing lightly structured infrastructures is strategically catachrestic - temporarily fixing components of an infrastructure that the semiotic technologist knows are not actually "fixed." Notably, temporarily and strategically hardening components of an infrastructure not only enables diverse communities to come together; as diverse communities enrich the components of the infrastructure that remain open with their diverse interpretations, it also opens up the possibility to compare and contrast where meaning is diverges, where it is inconsistent, and where it is incomplete. This type of juxtaposition is not possible without some structure.

The Human Services Data Specification is an example of a light structure; it temporarily and strategically hardens certain "standards" while leaving other areas of the infrastructure open. To take part in the collaboration, I&R databases do

---

<sup>13</sup>Some may call a lightly structured semiotic infrastructure a "boundary object." Boundary objects, according to Star and Griesemer (1989), are situated between and shared amongst diverse social worlds, helping communities work and collaborate in spite of their semantic differences. Even though communities may ascribe different meaning to them, boundary objects share a vague, ill-structured common identity across communities. I spent a great deal of time trying to figure out whether I considered the Human Service Data Specification to be a boundary object. It was difficult because the term boundary object itself has held different meanings as it is been applied in different contexts - something that Susan Leigh Star (2010, 601) makes abundantly clear in her later article "This is Not a Boundary Object: Reflections on the Origin of a Concept." In her article, she notes that aspects of the concept have been used mistakenly:

Much of the use of the concept has concentrated on the aspect of interpretive flexibility and has often mistaken or conflated this flexibility with the process of tacking back-and-forth between the ill-structured and well-structured aspects of the arrangements.

She uses the article to clarify what she and Griesemer *meant* when they introduced the concept in 1989. Yet, even in this clarification, certain aspects of the concept described in the original article are downplayed. For instance, in the original article, the concept of "translation" - a concept that I do not find particularly compelling - is a definitional component of boundary objects: "They have different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation" (emphasis mine; 393). On the other hand, the word "translation" appears only once in the 2010 article and not in a definitional capacity. This is an ironic demonstration of a primary thesis of this dissertation - that even as we try rigorously to say what we mean, that meaning is constantly escaping its center. For me (and for (Galison, 1997)), whether or not the definition for a boundary object includes a translational component is a difference that makes a difference.

In this process of going back and forth trying to decide if the Specification was a boundary object, I realized that I was operating a mode of *reproduction* - asking myself "does my case reproduce what Star and Griesemer found?" I realized that the activity was futile because the interesting question was not whether we found the same thing - whether the "edges" of our terms were the same - but instead how my situated reading of the domain could iterate meaning in the study of information infrastructures; this would mean operating in a mode of *differential reproduction*.

have to describe their data as "services" that are part of "organizations" that have a "phone number" and a "physical location;" however, the encoding and interpretation of what type of services are being offered and who is eligible for them is left to individual I&R centers. I&R centers can thus exchange their data even when they use different vocabularies to describe it.

The Specification alone will not solve all of I&R's problems. For instance, as one I&R specialist noted during an open meeting discussing the design of the Specification, a minimal approach to encoding I&R information can put a burden on help-seekers in terrible catch-22s such as "I need housing for my kids but I can't get my kids out of foster care because I don't have housing." In such cases, I&R specialists will need language "standards" beyond (what Bloom describes as) the "factual information" that the Specification encodes. Bloom's putting "standards" in air quotes suggests an alternative design paradigm for such language "standards" - one that does not seek to hard-code a formal semantics but instead to lightly, temporarily, and strategically (in other words, catachrestically) fix certain terms. When such a design paradigm is applied, the language ideology interwoven through the resulting semiotic infrastructure is more reflexive, flexible, and dynamic than the language ideologies that have traditionally guided semiotic infrastructures in the human services. It is a language ideology that presumes that while such strategic essentializing is often violent, it is also necessary for advancing social justice and a "real world" where individuals can find the help they need.

## 6. CONCLUSION

As the final tallies of the 2016 U.S. election results were being reported, and the prospect of a Trump presidency was looking increasingly probable, news anchors and political strategists alike seemed "rudderless" (Poniewozik, 2016). An overwhelming majority of news stations and political pundits had predicted a Clinton presidency. The night before the election fivethirtyeight - a website that uses statistical analysis to make sense of election polls - suggested that Clinton was ahead in national polling by about 4 points (based on 1,106 national polls). At no point in the election year had the site put Trump ahead in average national polling. As of the day before the election, Nate Silver (2016), statistician, founder of fivethirtyeight, and the epitome of the new "data scientist" election figure following the Obama elections, had put the chances of a Clinton Presidency at 70%.

Around 2AM Wednesday, November 10, 2016, Mike Murphy, a GOP political strategist, pronounced the death of data - well at least a sort of death:

My crystal ball's been shattered into atoms here, since I predicted the exact opposite. I'm a typical campaign consultant type. We've been living and dying by data for a long time. Tonight data kind of died. The exit polls were originally off, the most credible polling was off. (MSNBC, 2016)

He tweeted out to his followers:

I've believed in data for 30 years in politics and data died tonight. I could not have been more wrong about this election.

Steve Lohr and Natasha Singer (2016), big data bloggers for the New York Times, reported:

It was a rough night for number crunchers. And for the faith that people in every field – business, politics, sports and academia – have increasingly placed in the power of data.

Donald J. Trump's victory ran counter to almost every major forecast – undercutting the belief that analyzing reams of data can accurately predict events. Voters demonstrated how much predictive analytics, and election forecasting in particular, remains a young science: Some people may have been misled into thinking Hillary Clinton's win was assured because some of the forecasts lacked context explaining potentially wide margins of error.

They noted that while data science has enabled us to "see things as never before," it also tends to miss "context and nuance." They suggested that, in the excitement over advances in data science during this election, forecasters failed to grapple with limits of data science as well as the "potentially flawed assumptions of the people who build predictive models."

This jump from celebrating the promise of data to lamenting the death of data mirrors much larger discourse trends around data in the late 2010s. Discourse around 'data' has never been more central to scientific research and governing. Data science is burgeoning as a discipline. A new adjective, 'computational', is preceding many more traditional academic disciplines (computational biology, computational linguistics, computational physics, and computational social science, for example) and marking an expectation that data analysis will be the core tool to advance research in these areas. Data science graduate and undergraduate programs are being established in universities across the country, and data science jobs are some of the fastest-growing in the United States (Burning Glass Technologies et al., 2017). Many federal and local government agencies have created a new position - Chief Data Officer - an individual responsible for managing data assets and advancing data analytics. Algorithms are used in courts to determine prison sentences (Barry-Jester, 2015) and replacing caseworker decisions about whether families will be awarded welfare benefits (Eubanks, 2018). And, since the late-2000s, there has also been new civic engagement around data. Civic technology groups focused on advocating for and hacking with open government data have been established in cities across the globe (Schrock, 2016), and in April 2017, at Marches for Science held across the globe, protesters held signs advocating that policy be enacted based on scientific

data rather than ideological leanings. Often, the assumptions that undergird these celebrations of data suggest that truth is inside the data, and that it is the job of a data scientist to uncover that truth.

Yet, in the wake of an election where fake news seemed to be the *modus operandi* and where dishonesty became a political strategy, discourse has promoted the phrase 'post-truth' to describe a new form of governing - a phrase that denotes "circumstances in which objects and facts are less influential in shaping public opinion than appeals to emotion and personal belief." This definition comes from the Oxford Dictionaries, which named 'post-truth' the word of the year for 2016, noting that its international usage spiked 2000% from 2015 to 2016 (Dictionaries, 2016). A post-truth society has a conflicted relationship with data. In describing the age of post-truth politics in a New York Times column, William Davies (2016) suggested that it marked "a transition from a society of facts to a society of data." He wrote, "It is possible to live in a world of data and no facts," arguing that while data is being produced at unprecedented rates, this does not translate to citizens and governments putting faith in the data as representative of reality. For Davies, in a post-truth era, dominant discourse suggests that numbers can be manipulated to tell whatever story the analyst wants, and thus scientific claims that threaten capitalism and patriarchy, based on data, can be written off as "fake science."

Thus, since the 2016 United States presidential election, discourse around data has been rather polarizing. Some suggest that data died with the election, dismissing data as biased and ideological, while others are placing more faith in data than ever, arguing that the dismissal of data is biased and ideological. Taken to their extremes, both framings place an unfair onus on numbers to purport truth - the former rejecting numbers for not offering truth, and the latter tending to fetishize numbers as truth. At both extremes, the framing is dangerous - the former enabling very powerful actors to write off evidence that their decisions and actions will increase inequalities and cause harm to the environment, and the latter blinding us to the marginalities that are always inherent within data. But the framing does not have to be this way; it is possible to use data to inform decisions without assuming that the data is complete, consistent, or objective. We need more flexible

language for talking about how knowledge can and should emerge from data, and we need more flexible language for talking about much-needed data expertise.

## 6.1 Studying Data and Data Infrastructures Ethnographically

Scholarship in Science and Technology Studies and Information Studies can help generate more nuanced ways of talking about data and the role that data can and should play in representing problems and concerns. Within these fields, there have certainly been promising strides in this direction. For instance, the academic journal *Big Data and Society* publishes scholarship that explores the major debates around Big Data and how the field is changing the way that governments, businesses, scholars, and activists operate. The conference *Data Power*, held every two years since 2015, brings together interdisciplinary researchers to theorize the ways data can both empower and disempower communities. Further, many Information Studies programs are building out separate areas of expertise in data ethics and socio-technical data studies.

Yet, there's still a long way to go, particularly when it comes to producing scholarship that helps data practitioners figure out how to work in the face of data's limits. I have been to several conferences, meetings, and workshops that have attempted to bring data practitioners together with scholars studying the politics and social impacts of data. Typically the aims of such meetings are to discuss challenges that arise in a society where big data is playing such a prominent role in decision-making and to chart pathways forward. I almost always leave these gatherings feeling as though there was a missed opportunity to imagine new types of collaboration, new research trajectories, and new data design paradigms. Practitioners will come with "real world" problems, such as: I need to represent the voices of my community in decision-making, but the data that I have available to me doesn't represent the voices of my community. Or my organization has all of this data that we would like to make a public resource, but we do not know how to sort through the privacy implications of opening the data. Folks in academia, on the other hand, will present their research - on data bias, data privacy, data

surveillance. It's incredibly important work, but it's rarely responding to the issues practitioners describe. Instead, it reiterates the issues they've already hashed out - that data is not necessarily representative of the things it is supposed to represent, or that there are significant issues around privacy when it comes to data. This is one of the points I've attempted to draw out in this dissertation. Increasingly, data practitioners know their data does not exhaustively represent an issue; the challenge is that they need to figure how to work and make decisions in spite of this.

I hope that the research methodology that I introduce in this dissertation can help bridge this gap. While interviewing and archival work was a significant component of my study design, perhaps the most important insight I had into the data communities I studied was as a data practitioner myself. In working on the design of PECE, I had to learn how to implement metadata standards and to examine whether it made sense to "semantify" the platform. I had to figure out how to structure PECE to enable collaboration and to make it possible to discover our data, without over-determining the meaning that diverse researchers would bring to the data. Having had to work through this challenge, when I eventually heard the practitioners I interviewed describe similar challenges, I was able to empathetically (and technically) understand the conflicting injunctions at play.

Throughout my research, I spent just as much time learning how the data frameworks, code, and platforms I studied actually worked as I did interviewing, observing participants, and sifting through archives. I learned about the architecture of the Web and the inner-workings of many applications that sit on top of it by taking a graduate course on Web Science, and I watched many YouTube video lectures breaking down the design and implementation of Semantic Web frameworks. I had a core group of computer scientist collaborators that I could turn to when I needed them to explain how something worked, and over time, I learned how to read through and understand code manuals, data documentation, and standards specifications on my own.<sup>1</sup> Knowing how the systems worked allowed me to ask those I interviewed

---

<sup>1</sup>Notably, this research was only possible because the digital systems and standards I studied were, for the most part, open systems and standards. Their documentation was freely available, and in many cases, their design was extremely well-documented because much of it was carried out on public Web forums that are now all archived. These methods are much more difficult for researchers examining proprietary software.

why they made particular design choices, and it made it easier for me to understand the stakes when they described the tradeoffs that ordered their work.

Studying data infrastructures this way can give researchers insight, not only into data politics, but also into potential design interventions. It can help data researchers better understand the forces that drive data infrastructures to represent certain things and not others - perhaps because maligned politics are seeking to suppress certain voices, because entrenched ways of representing the world have ignored difference, and/or because the data design was considered the best option amongst competing alternatives. Notably, the aim of understanding these forces should not always and necessarily be to fix the data design - because datasets, no matter how hard we try to scrape them of biases will never completely and consistently represent an issue. Instead, ethnographically studying the inner-workings of data and data infrastructures can help data researchers and practitioners better understand when and how it is appropriate to suggest that data "represents" something versus when it is important to highlight data's limits. Studying this way can help researchers and practitioners better assess when it is appropriate to inform their decisions based on data, and can help identify when more information is needed. Practitioners can help with this; more often than not, they know about the forces that shape their datasets. The time is ripe for new data research engagements and collaborations.

## 6.2 Shifts in Expert Sign Systems

This dissertation furthers understanding of the cultural forces at play shaping how data becomes meaningful. It narrates how the experts designing infrastructure for organizing data and encoding the meaning of data often have different ideas about what constitutes meaning and how meaning emerges, circulates, evolves, and corrodes. These ideas inform the way they approach their work. For instance, those assuming that meaning is purposeful, logical, and internally cohesive have worked to find and encode the "hidden structure of meaning" - determining where concepts start and end and how they logically link together. Others have presume meaning to be more situated - emerging in different ways depending on time, context, and speaker; these folks have attempted to build semiotic infrastructures that "delay

semantic commitment," encoding how to say something rather than encoding what can be said. Because of this, the infrastructures representing our data are animated by diverse semiotics.

However, getting their infrastructures to advance knowledge representation in the "real world" has challenged this semiotics. Semiotic technologists may design an infrastructure according to particular assumptions about how meaning is made and find that the resulting infrastructure does not actually meet the needs of the domain for which they are designing - perhaps because their design approach produces a mess, making it difficult to collaborate; because the approach over-codifies meaning in a domain in which practitioners need to be able to define their data flexibly; or because it can't account for the way that meanings inevitably change. While diverse ideas about semiotics may animate semiotic technologists' work, the political and practical dynamics of diverse data domains challenge these ideas, pushing them to assess which tradeoffs they are willing to make.

Getting their semiotic infrastructures to work in the "real world" - a real world where sign systems change - has thus provoked considerable changes in semiotic technologists' sign systems. They have described having to learn to become more "flexible over time," having to learn to "swim in troubled waters," or to be "consistent in the face of pragmatic realities." It is interesting that these changes in sign systems are happening in a community that has been categorized according to their shared efforts to encode the meaning of signs. It means that they are constantly redefining their expertise. Semiotic infrastructure design is becoming more experimental and exploratory than formulaic and restrictive. Semiotic technologists have had to learn how to approach diverse data domains not assuming that they'll know the right way to encode the meaning of data ahead of time, but instead that they will consistently get better at encoding data through trial and error - learning about the dynamics of the domain along the way.

In this sense, this dissertation offers a new way of thinking about how expertise takes shape, operates, and shapes socio-cultural order. Just like the meaning of the signs semiotic technologists encode are perpetually iterating, what it means to be an expert in this community is perpetually iterating - perhaps stabilizing in

certain times and contexts, but taking on new meaning as they approach different data domains. It takes on different meaning as they are exposed to and learn to experimentally work in the face of different "real world" challenges. And because their diverse approaches to data design shape the infrastructures that organize data and encode its meaning, the way this expertise evolves impacts how knowledge is produced and represented. To prepare semiotic technologists to expertly move through these worlds, we need to prepare them with skill in reading diverse data domains and enduring double bind.

### **6.3 The Need for New Kinds of Data Expertise**

This dissertation, in part, reiterates the refrain (repeated by several cultural analysts of data) that "raw data is an oxymoron" (Bowker, 2000; boyd and Crawford, 2012; Kitchin, 2013; Gitelman, 2013). Data cannot exist divorced from the context of its production and the technologies and discourses that frame it; instead, all data is made meaningful by those that collect, describe, order, and analyze it. This dissertation adds to this conversation that the tools that data scientists have for describing and ordering their data - the semiotic infrastructures that encode the meaning of their data - also cannot be divorced from the contexts of their production. These infrastructures too have particular ideologies interwoven in their design, helping to shape how data becomes meaningful for diverse people. Yet, while this would suggest that biases influence data at every level, I do not intend to move to the extreme of dismissing data as "dead." I do, however, intend to mark all knowledge representation work as political and interpretive; data scientists enact "real worlds" in choosing to represent the world in the way that they do. The dissertation thus suggests the need for more robustly developing new kinds of data expertise - expertise that can effectively acknowledge and unpack the politics of practical representation and expertise that can creatively think, work, and write in the face of double bind.

Since the 1970s, feminist cultural theorists such as Gayatri Spivak (2012)), Drucilla Cornell (1992), and Teresa de Lauretis (1984) have characterized how representing difference is a double bind. To acknowledge difference, diversity, or what

typically is "Other" to dominant discourse, an analyst needs to identify and define it. However, no sign can completely represent difference; no sign can offer a coherent identity to the referents it intends to signify. For instance, as Gayatri Spivak and Harasym (1990, 104) argues, there are no "literal referents" to words like "worker" or "woman;" in other words, "there are no 'true examples' of the 'true worker' [or] the 'true woman'." And while it is often very important to mark "worker" and "woman" in order to advance social justice, doing so eclipses the diversity of what it can mean to be a worker and a woman.<sup>2</sup> As soon as a semiotic technologist signifies difference - as soon as the analyst encodes difference with a sign - the analyst also eclipses difference. Representing difference is catachresis - both necessary and violent - forcing a signifier on a meaning that is not and never will be fully closed. This is the double bind with which semiotic technologists must grapple.

Throughout the dissertation, I've documented communities that are learning to represent messy and complex data. In some cases, this work has involved cleaning data representations up and restricting the meaning that can be drawn from them with language standards. At other times, modes of knowledge representation have been looser, more flexible, and scruffier - not necessarily mandating a common way of describing, structuring, ordering, or interpreting something, but instead allowing diverse communities to "say anything about anything." Sometimes, the work has been situated somewhere in the "middle," sacrificing cleanliness for expressivity, or, alternatively, the ability to "say anything about anything" for the ability to produce a single answer.

The most creative approaches to knowledge representation, however, have emerged as semiotic technologists acknowledge the need to work at all ends of these spectrums - as they acknowledge that, in order to advance scientific research, good

---

<sup>2</sup>With this in mind, Spivak and Harasym (1990, 46) describes how she likes the word "subaltern" because of its scruffiness:

I like the word 'subaltern' for one reason. It is truly situational. 'Subaltern' began as a description of a certain rank in the military. The word was used under censorship by Gramsci: he called Marxism 'monism,' and was obliged to call the proletarian 'subaltern.' That word, used under duress, has been transformed into the description of everything that doesn't fall under class analysis. I like that, because it has no theoretical rigor.

governance, and social justice, polysemy often really does need to be sorted out, but also that doing so forecloses the possibility for representing complex differences. In other words, the most creative approaches to knowledge representation have emerged as they acknowledge the double bind that troubles their work - that they need standards to be able to name something (or to mark difference), but the act of standardizing arrests the marking of difference. Greg Bloom acknowledged this when he described putting the word "standard" in air quotes. Putting "standard" in air quotes acknowledges that it is impossible to point to a "true example" of a "true standard;" it acknowledges that the word "standard" is catachrestic. While the human services desperately need some common language or a "true standard," there are always more differences that will disrupt the full closure of what it means for something to be "standard. It is also this paradoxical positioning of "standard" as both necessary and impossible to fully achieve that provoked Ora Lassila to describe a need for designing semiotic infrastructures that "delay semantic commitment," or, in other words, "standard" infrastructures that do not demand "standard" semantics. Neither approach resolves the double bind of representing difference, but both offer a different, more situational (and notably scruffier) way of thinking about and designing standards. They both acknowledge that designing standards is always a pursuit - never quite reaching the one right way to represent something - never quite pinpointing the sweet spot along the neat-middle-scruffy spectrum (in part because that spot cannot exist) - but always striving for it.

To be better at pursuing un-standardized standards, the skill of a semiotic technologist needs to be about more than disambiguating terms and logically modeling meaning. We need data experts that can represent and draw meaningful insights from data in the "real world" - a world where meaning is scruffy, logic is plural, and data can only ever represent a limited point of view. As algorithms are increasingly being leveraged to inform policy and to make decisions about prison sentences, the distribution of government services, and whether or not an individual gets hired for a job, we need to train data scientists to recognize when designing semiotic infrastructures that clean up or eclipse complexity and difference can literally cost people their livelihoods and their lives. However, we also need to train data scientists to

recognize when they must (at least temporarily) clean up or eclipse complexity and difference - because not doing so can disarm people's ability to find information, interpret information, and communicate, also potentially costing livelihoods and lives. And because it is impossible for data scientists to do both - to design infrastructures that both gloss over difference and do not gloss over difference - data scientists need to be better at discerning when representing differences makes a political, ethical, and practical difference. They need to be better at discerning when they should approach their work with neat modalities, when they should approach their work with scruffy ones, and when they should make compromises between the two. To do so, they need to be skilled at reading the domains in which they work so that they can better assess the consequences of reducing friction, sorting out polysemy, and trading off the ability to expressively represent complex information for the ability to produce a single answer.

## 6.4 Learning to Critically Represent Data

This type of skill cannot be taught with textbooks, code manuals, or formulas alone - in part, because resolving a double bind is paradoxical, and thus cannot be "neatly" represented. Data science students need practice in pursuing impossible aims, and this demands a rethinking of how data science is taught in university courses. Data science students should be exposed to more "real world" data and data problems - not just so that they get exposure to the messiness, inconsistency, and incompleteness that riddles "real world" datasets, but also so that they get practice in examining the information pragmatics and politics of diverse data domains.

I've begun developing this type of curriculum for information technology and data science students. In one course I designed called *Critical Data Mapping*, students spent one entire 2-hour class period designing a census for the Rensselaer Polytechnic Institute campus. To begin the class, we discussed the history of the U.S. census - outlining how and why the division of categories like race, ethnicity, and nationality have been important for directing government resources to diverse groups, troubling in how they identify certain groups, and always incomplete in capturing the country's diversity. I then asked the students to imagine that RPI

had hired them as a consulting company to design and conduct a census that would eventually be used to help the university identify the types of diversity programs that should be developed. We spent the first half hour simply discussing how they'd ask census-takers about their gender. Some students in the class suggested that the field should be free-text - that census-takers should be able to fill in whatever gender they'd like. Other students in the class pushed back, noting that if census-takers could write-in whatever they wanted (if they could "say anything about anything") then the responses may be so diverse that the census would not actually help RPI in planning for diversity programming. Students in the class moved on to suggest that the census include a long list of specific genders, from which census-takers could select one, but that it also include a list of more general genders, allowing census-takers to mark off if they identify with any more general categories. But this required students in the class to begin outlining a sort of ontology of gender - disambiguating more general gender categories from more specific ones, and they quickly began feeling uncomfortable with the activity - finding it difficult to discern when gender categories should be considered specific or broad. They ultimately opted for a "middle" approach, where census-takers could select one or more genders from a list of about ten, but also had an option to fill in a field labeled "I identify as..." with free-text.

In debriefing at the end of class, the students expressed how frustrating the activity was for them. This was in part because, there was no "right" way to divide the categories to ensure that diversity was completely and consistently represented. The task posed a no-win situation, and this required a mode of thinking and work that many information technology, computer, and data science students never have to encounter in their coursework but will inevitably encounter in the "real world." It presented a case where students could not rely on formulas, proofs, or best practices to structure their thinking and work. Instead, it demanded them to consider the politics along with the pragmatics of their data work, and to make good choices when there were no correct choices. Later in the semester, as we began to use GIS technologies to map census data, students were noticeably more thoughtful in characterizing what their maps "represented." Consider what one student wrote

after building a map that compared census data about income levels in NYC with complaints the city had received about water quality:

The above map intertwines locations of Water Quality Complaints (based on 311 Service Requests) with U.S. Census Income data. The 5 boroughs of New York City are represented by their census block groups, where the darker block groups indicate a greater number of households with an income of 25,000to29,999. While the Water Quality complaints appear spread out overall, Manhattan distinctly has more instances of Water Quality Complaints compared to areas in the Bronx, Brooklyn, and Queens. This could be for a variety of reasons, including Manhattan's higher population density, accessibility to technological infrastructure to contact 311, and knowledge to call 311 in such instances; however, these reasons cannot be determined from this map alone. Furthermore, the census data on income is highly restrictive and potentially misleading. This data is divided by different income ranges containing the number of households that fall in this income range. As a result, this data cannot be represented as income as a percentage of the population in that block group, which is misleading.

One of my goals in assigning this activity is to encourage students to see census data as incredibly biased, while not seeing it as useless. I want them to thoughtfully consider how they can work with biased data and how they can represent problems and solutions with biased data. Suggesting to data science students that subjective data or biased data is *bad* does an injustice to the promise of data science and fuels the polarization of discourse around data. It sets an impossible bar for datasets that can never, as this dissertation has shown, achieve "rawness." It also passes over an opportunity to teach students how to contextualize their findings, how to look out for what's been ignored, and how to represent missing information. Instead, I like to have students work with datasets that are of critical importance to holding industries and governments accountable to the public but that are also glaringly non-representative of the full extent of the public problems they intend to represent.

Take, for example the Environmental Protection Agency's Toxic Release Inventory (TRI) - a dataset that documents the amounts of toxic chemicals industrial facilities release into the environment each year. Since 1986, the mandated publishing of the TRI has been effective in prodding industries to cut down on their release of toxic emissions, and the dataset has also been invaluable to community members and activists investigating environmental pollutants in their neighborhoods (Fung and O'rourke, 2000; Fortun, 2004; Hamilton, 2005). However, the data is also industry-reported with little auditing, and facilities have developed creative strategies to under-report emissions to the EPA (Marchi and Hamilton, 2006; Hiar, 2012). Or consider the Home Mortgage Disclosure Act, which mandates financial institutions to disclose to the public data about the demographics of their mortgage applicants and whether or not they are approved for loans. The publishing of this data has helped government officials, researchers, and activists monitor against banks' discriminatory lending practices (Hamilton, 2005; Mendez et al., 2011). However, the dataset is far from representative of discrimination as research has shown that information about the race and ethnicity of applicants is habitually missing in the datasets (Berkovec and Zorn, 1996; Dietrich, 2002; Avery et al., 2007).

These datasets are important. They've advanced the public's understanding of civic problems, and they've provided scaffolding for multiple stakeholders to respond to the problems. The availability and format of the data is also profoundly shaped by political ideologies and business strategies bent towards ensuring that the data never fully represents the extent of industrial toxic pollution or discriminatory lending practices. Students need to learn to work with these data sets reflexively - carefully considering the limits of the data along with the value. The exercise I described above, where students are tasked with redesigning a dataset - is one method for doing this. It provokes them to acknowledge how datasets do not emerge from nothing, but instead are crafted through a process of discernment. The exercise encourages them to consider how individuals and institutions decide what gets included in a dataset and how categories get divided, and it allows them to experience how these decisions are often far from straightforward.<sup>3</sup> It does not suggest that there are

---

<sup>3</sup>Another strategy for encouraging more reflexive data work is to assign students to ethnographically research datasets, prompting them to unpack a dataset's historical and political genealogy.

"best practices" for representing data, but instead that there are more appropriate approaches to representing data in particular contexts. Finally, it gives them an opportunity to practice enduring double bind. To cultivate more thoughtful, critical, creative, and even sometimes devious data practitioners, university courses should not encourage students to shy away from or attempt to sort out the paradoxes of data work; instead, they should challenge students to think, work, and represent information in the face of them.

## 6.5 Conclusion

Today's "real world" is a world where the climate is warming, discrimination is amplifying, and support for government programs that help the poor is under attack. Understanding and responding to these complex problems is going to require data, and it is going to require collaboration across groups that will likely speak different languages. Data scientists need to be prepared not just for the technical challenges of ordering, integrating, and interpreting data across these groups, but also for the semiotic, cultural, and political challenges of doing so. We need semiotic infrastructures to be designed in ways that are appropriate to the conditions of the diverse data domains they will support. And while language practices may paralyze the capacity to encode such infrastructures, semiotic technologists need to figure out ways to think, work, and write through the paralysis - because right now the stakes are far too high to arrest the acknowledgement and identification of sameness and difference. Now more than ever, semiotic technologists need to be prepared to represent complex information in scruffy worlds.

---

Not only does this assignment give students exposure to the infrastructure that makes it possible for data to exist, but it also positions data as something deserving of cultural study, helping to combat a discourse that suggests that numbers can speak for themselves. Since the concept of studying data ethnographically is often very new for data science students, I've assigned it in a very structured way - as a series of questions students can fairly easily find answers to with a quick Web search. The point of the assignment is to provide students with tools for investigating diverse data domains - so that when they are put in situations where they have to decide how to structure, order, and clean up data, they know how to first examine the domain and better assess, in context, the consequences of the decisions that they make.

## REFERENCES

- Abbate, Janet. 2000. *Inventing the Internet*. Cambridge, MA: MIT Press.
- Adam, Alison. 1998. *Artificial Knowing: Gender and the Thinking Machine*. London; New York: Routledge.
- Akrich, Madeleine. 1992. "The De-scription of Technical Objects." In *Shaping Technology / Building Society: Studies in Sociotechnical Change*, edited by Wiebe Bijker, 205–224. Cambridge, MA: MIT Press.
- Allemang, Dean and James Hendler. 2011. *Semantic Web for the Working Ontologist, Second Edition: Effective Modeling in RDFS and OWL* (2nd ed.). Waltham, MA: Morgan Kaufmann.
- Alliance of Information and Referral Systems. 2016. "AIRS Standards and Quality Indicators for Professional Information and Referral Version 8.0." [https://www.airs.org/files/public/AIRS\\_Standards\\_8\\_0.pdf](https://www.airs.org/files/public/AIRS_Standards_8_0.pdf) (Date Last Accessed, November, 11, 2017).
- Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." [http://archive.wired.com/science/discoveries/magazine/16-07/pb\\_theory](http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory) (Date Last Accessed, April 17, 2018).
- Asad, Talal. 1986. "The Concept of Cultural Translation in British Social Anthropology." In *Writing Culture: The Poetics and Politics of Ethnography : a School of American Research Advanced Seminar*, edited by James Clifford and George E. Marcus, 141–164. Berkeley, CA: University of California Press.
- Ashmore, Malcolm, Robin Wooffitt, and Stella Harding. 1994. "Humans and Others: The Concept of " Agency" and Its Attribution." *American Behavioral Scientist* 37, no. 6 (May): 733–740.
- Austin, John Langshaw. 1975. *How to Do Things with Words*. Oxford: Clarendon Press.
- Avery, Robert, Kenneth Brevoort, and Glenn Canner. 2007. "Opportunities and Issues in Using HMDA Data." *Journal of Real Estate Research* 29, no. 4 (January): 351–380.
- Barad, Karen. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham, NC: Duke University Press.
- Barry-Jester, Anna Maria. 2015. "Should Prison Sentences Be Based On Crimes That Haven't Been Committed Yet?" <https://fivethirtyeight.com/features/prison-reform-risk-assessment/> (Date Last Accessed, January, 19, 2018).

- Bateson, Gregory. 1972. *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. Chicago: University of Chicago Press.
- Berg, Maggie. 1991. "Luce Irigaray's "Contradictions": Poststructuralism and Feminism." *Signs: Journal of Women in Culture and Society* 17, no. 1 (October): 50–70.
- Berkovec, Jim and Peter Zorn. 1996. "How Complete is HMDA?: HMDA Coverage of Freddie Mac Purchases." *Journal of Real Estate Research* 11, no. 1 (January): 39–55.
- Berners-Lee, Tim. 1999. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. San Francisco: Harper Business.
- Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284, no. 5 (May): 29–37.
- Bijker, Wiebe E., Thomas Parke Hughes, and Trevor J. Pinch. 1987. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, MA: MIT Press.
- Bloom, Greg. 2013. "Towards a Community Data Commons." In *Beyond Transparency: Open Data and the Future of Civic Innovation* (1st ed.), edited by Brett Goldstein, 255–270. San Francisco: Code for America Press.
- Bloom, Greg. 2017. "Introducing Open Referral: A Better Way to Manage Health, Human, and Social Services Data." <https://socrata.com/blog/introducing-open-referral-better-way-manage-health-human-social-services-data/> (Date Last Accessed, November, 14, 2017).
- Bobrow, Daniel G. and Terry Winograd. 1977. "An Overview of KRL, a Knowledge Representation Language." *Cognitive Science* 1, no. 1 (January): 3–46.
- Boellstorff, Tom. 2016. "For Whom the Ontology Turns: Theorizing the Digital Real." *Current Anthropology* 57, no. 4 (June): 387–407.
- Bogost, Ian and Nick Montfort. 2009. "Platform Studies: Frequently Questioned Answers." *Digital Arts and Culture*,. <https://escholarship.org/uc/item/01r0k9br> (Date Last Accessed, April 12, 2018).
- Borgida, Alexander, Ronald J. Brachman, Deborah L. McGuinness, and Lori Alperin Resnick. 1989. "CLASSIC: A Structural Data Model for Objects." In *Proceedings of the 1989 ACM SIGMOD International Conference on Management of Data* 58–67. New York, NY, USA: ACM.

- Borgman, Christine L. 2012. "The Conundrum of Sharing Research Data." *Journal of the American Society for Information Science and Technology* 63, no. 6 (June): 1059–1078.
- Bowker, Geoffrey. 1994. "Information Mythology and Infrastructure." In *Information Acumen: The Understanding and Use of Knowledge in Modern Business*, edited by Lisa Bud-Frierman, 231–247. London, New York: Routledge.
- Bowker, Geoffrey, Karen Baker, Florence Millerand, and David Ribes. 2009. "Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment." In *International Handbook of Internet Research*, edited by Jeremy Hunsinger, Lisbeth Klastrup, and Matthew Allen, 97–117. Netherlands: Springer.
- Bowker, Geoffrey C. 2000. "Biodiversity Datadiversity." *Social Studies of Science* 30, no. 5 (October): 643–683.
- Bowker, Geoffrey C. and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.
- boyd, danah and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15, no. 5 (June): 662–679.
- Brachman, Ronald, Deborah McGuinness, Peter Patel-Schneider, Lori Alperin Resnick, and Alex Borgida. 1991. "Living with CLASSIC: When and How to Use a KL-ONE-like Language." In *Principles of Semantic Networks: Explorations in the Representation of Knowledge*, edited by John Sowa, 401–456. San Mateo, CA: Morgan Kaufmann.
- Brachman, Ronald J. 1978. "A Structural Paradigm for Representing Knowledge." Ph. D. thesis, Harvard University.
- Brachman, Ronald J. 1987. "The Myth of the One True Logic." *Computational Intelligence* 3, no. 1 (February): 168–172.
- Brachman, Ronald J., Richard E. Fikes, and Hector J. Levesque. 1983. "Krypton: A Functional Approach to Knowledge Representation." *Computer* 16, no. 10 (October): 67–73.
- Brachman, Ronald J and Hector J. Levesque. 1982. "Competence in Knowledge Representation." In *AAAI-82* 189–192. Philadelphia: Morgan Kaufmann Publishers Inc.
- Brachman, Ronald J. and James G. Schmolze. 1985. "An Overview of the KL-ONE Knowledge Representation System\*." *Cognitive Science* 9, no. 2 (April): 171–216.
- Burning Glass Technologies, Business-Higher Education Forum, and IBM. 2017. *The Quant Crunch: How the Demand for Data Science Skills is Disrupting the Job Market*. Technical report, Boston, MA: Burning Glass Technologies.

- Bush, Vannevar. 1945. "As We May Think." *Atlantic Monthly* 176, (July): 101–108.
- Butler, Judith. 1988. "Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory." *Theatre Journal* 40, no. 4 (December): 519–531.
- Callon, Michel. 1986. "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of Saint Brieuc Bay." In *Power, Action and Belief: a New Sociology of Knowledge?*, edited by John Law, 196–233. London: Routledge.
- Carnap, Rudolf. 1937. *Logical Syntax of Language*. New York: Psychology Press.
- Carr, E. Summerson. 2010. "Enactments of Expertise." *Annual Review of Anthropology* 39, no. 1 (October): 17–32.
- Castells, Manuel. 2009. *The Rise of the Network Society: The Information Age: Economy, Society, and Culture Volume I* (2nd ed.). Chichester, West Sussex; Malden, MA: Wiley-Blackwell.
- Charniak, Eugene. 1981. "A Common Representation for Problem-Solving and Language-Comprehension Information." *Artificial Intelligence* 16, no. 3 (July): 225–255.
- Charniak, Eugene. 1986. "A Neat Theory of Marker Passing." In *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence* 584–588. Philadelphia, Pennsylvania: AAAI Press.
- Chepesiuk, Ron. 2001. "Dial 211: Libraries Get Involved with a New Social Service Initiative." *American Libraries* 32, no. 11 (December): 44–46.
- Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton.
- Cilliers, Paul. 1998. *Complexity and Postmodernism: Understanding Complex Systems* (1st ed.). London; New York: Routledge.
- Clifford, James and George E. Marcus. 1986. *Writing Culture: The Poetics and Politics of Ethnography: a School of American Research Advanced Seminar*. Berkeley, CA: University of California Press.
- Collins, Allan M. and M. Ross Quillian. 1969. "Retrieval Time from Semantic Memory." *Journal of Verbal Learning and Verbal Behavior* 8, no. 2 (April): 240–247.
- Collins, Harry. 1992. *Artificial Experts: Social Knowledge and Intelligent Machines*. Cambridge, MA: MIT Press.
- Collins, Harry. 2004. "Interactional Expertise as a Third Kind of Knowledge." *Phenomenology and the Cognitive Sciences* 3, no. 2 (June): 125–143.

- Collins, Harry and Robert Evans. 2002. "The Third Wave of Science Studies: Studies of Expertise and Experience." *Social Studies of Science* 32, no. 2 (April): 235–296.
- Cornell, Drucilla. 1992. *The Philosophy of the Limit*. New York; London: Routledge, Chapman & Hall, Incorporated.
- Course, Magnus. 2010. "Of Words and Fog: Linguistic Relativity and Amerindian Ontology." *Anthropological Theory* 10, no. 3 (September): 247–263.
- Coursen, Derek. 2013. "Why This Blog?" <https://humanserviceinformatics.wordpress.com/about/> (Date Last Accessed, November, 14, 2017).
- Crevier, Daniel. 1994. *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York: BasicBooks.
- Davies, William. 2016. "Opinion | The Age of Post-Truth Politics." <https://www.nytimes.com/2016/08/24/opinion/campaign-stops/the-age-of-post-truth-politics.html> (Date Last Accessed, January, 19, 2018).
- DBpedia. 2017. "Facts and Figures." <http://wiki.dbpedia.org/about/facts-figures> (Date Last Accessed, July, 16, 2017).
- de Lauretis, Teresa. 1984. *Alice Doesn't: Feminism, Semiotics, Cinema*. Bloomington: Indiana University Press.
- Derrida, Jacques. 1970. "Structure, Sign, and Play in the Discourse of the Human Sciences." In *The Languages of Criticism and the Sciences of Man: The Structuralist Controversy*, edited by Richard Macksey and Eugenio Donato, 247–265. Baltimore, MD: Johns Hopkins University Press.
- Derrida, Jacques. 1982a. *Margins of Philosophy*. Brighton: Harvester Press.
- Derrida, Jacques. 1982b. *Positions* (1st ed.). Chicago: University of Chicago Press.
- Derrida, Jacques. 1983. *Dissemination* (1st ed.). Chicago: University of Chicago Press.
- Derrida, Jacques. 1988. *Limited Inc* (1st ed.). Evanston, IL: Northwestern University Press.
- Derrida, Jacques. 1994. *Specters of Marx: The State of the Debt, the Work of Mourning, and the New International*. New York; London: Routledge.
- Derrida, Jacques. 1996. *Archive Fever: A Freudian Impression*. Chicago, IL: University of Chicago Press.
- Descola, Philippe. 2010. "Cognition, Perception and Worlding." *Interdisciplinary Science Reviews* 35, no. 3-4 (December): 334–340.

- Dictionaries, Oxford. 2016. "Oxford Dictionaries Word of the Year 2016 is..." <https://www.oxforddictionaries.com/press/news/2016/11/17/WOTY-16> (Date Last Accessed, January, 19, 2018).
- Dietrich, Jason. 2002. "Mortgage Applications with Missing Race Data and the Implications for Monitoring Fair Lending Compliance." *Journal of Housing Research* (1): 51–84.
- Doctrow, Cory. 2001. "Metacrap: Putting the torch to seven straw-men of the meta-utopia." <http://www.well.com/doctorow/metacrap.htm#0> (Date Last Accessed, October, 20, 2016).
- Dreyfus, Hubert L. 1972. *What Computers Can't Do : A Critique of Artificial Reason* (1st ed.). New York: Harper & Row.
- Drucker, Johanna. 2011. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5, no. 1 (March): 1–21.
- Drupal. 2017. "Guidelines for taxonomy design." <https://www.drupal.org/docs/7/organizing-content-with-taxonomies/guidelines-for-taxonomy-design> (Date Last Accessed, December, 15, 2017).
- Edwards, Paul, Matthew S. Mayernik, Archer Batcheller, Geoffrey Bowker, and Christine Borgman. 2011. "Science Friction: Data, Metadata, and Collaboration." *Social Studies of Science* 41, no. 5 (August): 667–690.
- Edwards, Paul N.. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.
- Ensmenger, Nathan. 2011. "Is Chess the Drosophila of AI? A Social History of an Algorithm." *Social Studies of Science* 42, no. 1 (October): 5–30.
- Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.
- Fischer, Michael M. J. 2003. *Emergent Forms of Life and the Anthropological Voice*. Durham; London: Duke University Press.
- Fischer, Michael M. J. 2012. "Culture and Cultural Analysis as Experimental Systems." *Cultural Anthropology* 22, no. 1 (November): 1–65.
- Fleck, Ludwik. 1981. *Genesis and Development of a Scientific Fact*. Chicago, IL: University of Chicago Press.
- Forsythe, Diana E. 1993. "Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence." *Social Studies of Science* 23, no. 3 (August): 445–477.
- Fortun, Kim. 2001. *Advocacy after Bhopal Environmentalism, Disaster, New Global Orders*. Chicago, IL: University of Chicago Press.

- Fortun, Kim. 2004. "From Bhopal to the Informing of Environmentalism: Risk Communication in Historical Perspective." *Osiris* 19, 283–296.
- Fortun, Kim. 2012. "Ethnography in Late Industrialism." *Cultural Anthropology* 27, no. 3 (August): 446–464.
- Fortun, Kim. 2014. "From Latour to late industrialism." *HAU: Journal of Ethnographic Theory* 4, no. 1 (June): 309–329.
- Fortun, Kim and Mike Fortun. 2015. "An Infrastructural Moment in the Human Sciences." *Cultural Anthropology* 30, no. 3 (August): 359–367.
- Fortun, Kim, Mike Fortun, Erik Bigras, Tahereh Saheb, Brandon Costelloe-Kuehn, Jerome Crowder, Daniel Price, and Alison Kenner. 2014. "Experimental Ethnography Online." *Cultural Studies* 28, no. 4 (February): 632–642.
- Fortun, Michael and Herbert J. Bernstein. 1998. *Muddling Through: Pursuing Science and Truths in the 21st Century*. Washington D.C.: Counterpoint.
- Foucault, Michel. 1982. *The Archaeology of Knowledge*. New York: Vintage.
- Fung, Archon and Dara O'rourke. 2000. "Reinventing Environmental Regulation from the Grassroots Up: Explaining and Expanding the Success of the Toxics Release Inventory." *Environmental Management* 25, no. 2 (February): 115–127.
- Gad, Christopher, Casper Bruun Jensen, and Brit Ross Winthereik. 2015. "Practical Ontology: Worlds in STS and Anthropology." *Natureculture* no. 3 67–86.
- Gal, Susan and Judith Irvine. 1995. "The Boundaries of Languages and Disciplines: How Ideologies Construct Difference." *Social Research* 62, no. 4 (Winter): 967–1001.
- Galison, Peter. 1997. *Image and Logic: A Material Culture of Microphysics*. Chicago, IL: University of Chicago Press.
- Galloway, Alexander R. 2014. "The Cybernetic Hypothesis." *differences* 25, no. 1 (January): 107–131.
- Gerlitz, Carolin and Anne Helmond. 2011. "Hit, Link, Like, and Share: Organizing the Social and the Fabric of the Web in a Like Economy." <https://pdfs.semanticscholar.org/6419/1a8751d6ed4986bf67f92b3391cccbd154fa.pdf> (Date Last Accessed, April 12, 2018).
- Gillespie, Tarleton. 2010. "The Politics of 'Platforms'." *New Media & Society* 12, no. 3 (May): 347–364.
- Gillespie, Tarleton. 2014. "The Relevance of Algorithms." In *Media Technologies: Essays on Communication, Materiality, and Society* 167–194. Cambridge, MA: MIT Press.

- Gitelman, Lisa. 2013. *Raw Data Is an Oxymoron*. Cambridge, MA: MIT Press.
- Golumbia, David. 2009. *The Cultural Logic of Computation*. Cambridge, MA: Harvard University Press.
- Grint, Keith and Steve Woolgar. 1997. *The Machine at Work: Technology, Work and Organization*. Cambridge, U.K.: Polity.
- Gruber, Thomas R. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 5, no. 2 (June): 199–220.
- Guha, R. V., Dan Brickley, and Steve Macbeth. 2016. "Schema.Org: Evolution of Structured Data on the Web." *Commun. ACM* 59, no. 2 (January): 44–51.
- Gödel, Kurt. 1962. *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*. New York: Basic Books, Inc.
- Halford, Susan, Catherine Pope, and Mark Weal. 2013. "Digital Futures? Sociological Challenges and Opportunities in the Emergent Semantic Web." *Sociology* 47, no. 1 (February): 173–189.
- Hall, Stuart. 1980. "Encoding/Decoding." In *Culture, Media, Language: Working Papers in Cultural Studies, 1972-79*, edited by Centre for Contemporary Cultural Studies, 123–138. London: Hutchinson.
- Halpin, Harry, Patrick J. Hayes, James P. McCusker, Deborah L. McGuinness, and Henry S. Thompson. 2010. "When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data." In *The Semantic Web – ISWC 2010*, edited by Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, 305–320. Berlin Heidelberg: Springer.
- Hamilton, James. 2005. *Regulation Through Revelation: The Origin, Politics, and Impacts of the Toxics Release Inventory Program*. New York: Cambridge University Press.
- Haraway, Donna. 1988. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." *Feminist Studies* 14, no. 3 (Autumn): 575.
- Haraway, Donna. 1997. *Modest-Witness@Second-Millennium.FemaleMan-Meets-OncoMouse: Feminism and Technoscience*. New York: Routledge.
- Hardy, Quentin. 2016. "The Web's Creator Looks to Reinvent It." *The New York Times*,. <http://www.nytimes.com/2016/06/08/technology/the-webs-creator-looks-to-reinvent-it.html> (Date Last Accessed, October, 26, 2016).

- Harmelen, Frank van, Ian Horrocks, Peter Clark, Peter F. Patel-Schneider, Michael Uschold, Marie-Christine Rousset, James Hendler, and Guus Schreiber. 2002. "Ontologies' KISSES in Standardization." *IEEE Intelligent Systems* 17, no. 2 (March): 70–79.
- Hayes, Patrick J. 1977. "In Defense of Logic." In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 1* 559–565. San Francisco, CA: Morgan Kaufmann Publishers Inc.
- Hayes, Patrick J. 1979. "The Naïve Physics Manifesto." In *Expert systems in the micro-electronic age*, edited by Donald Michie, 242–270. Edinburgh: University Press.
- Hayes, Patrick J. 1987. "A Critique of Pure Treason." *Computational Intelligence* 3, no. 1 (February): 179–185.
- Hayles, N. Katherine. 1993. "The Materiality of Informatics." *Configurations* 1, no. 1 (January): 147–170.
- Hayles, N. Katherine. 2004. "Print is Flat, Code is Deep: The Importance of Media-Specific Analysis." *Poetics Today* 25, no. 1 (March): 67–90.
- Heflin, Jeff, James Hendler, and Sean Luke. 1999. "SHOE: A Knowledge Representation Language for Internet Applications." ., <http://drum.lib.umd.edu/handle/1903/1044> (Date Last Accessed, February, 17, 2017).
- Hendler, James, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. 2008. "Web Science: An Interdisciplinary Approach to Understanding the Web." *Commun. ACM* 51, no. 7 (July): 60–69.
- Hiar, Corbin. 2012. "EPA's Toxics Release Inventory doesn't offer full picture of pollution." <https://www.publicintegrity.org/2012/01/09/7836/epas-toxics-release-inventory-doesnt-offer-full-picture-pollution> (Date Last Accessed, July, 29, 2016).
- Hodes, Harold T. 1984. "Logicism and the Ontological Commitments of Arithmetic." *The Journal of Philosophy* 81, no. 3 (March): 123–149.
- Hui, Yuk. 2016. *On the Existence of Digital Objects*. Minneapolis, MN: University of Minnesota Press.
- Irigaray, Luce. 1980. "When Our Lips Speak Together." *Signs* 6, no. 1 (October): 69–79.
- Jackson, Sarah J. and Brooke Foucault Welles. 2015. "Hijacking #myNYPD: Social Media Dissent and Networked Counterpublics." *Journal of Communication* 65, no. 6 (December): 932–952.

- Joyce, John W. 1943. "The Social Service Exchange and Probation." *Federal Probation* 7, 34.
- Keller, Evelyn Fox. 2003. *Making Sense of Life: Explaining Biological Development with Models, Metaphors, and Machines*. Cambridge, MA: Harvard University Press.
- Kelly, John D. 2014. "Introduction: The Ontological Turn in French Philosophical Anthropology." *HAU: Journal of Ethnographic Theory* 4, no. 1 (June): 259–269.
- Kitchin, Rob. 2013. "The Real-Time City? Big Data and Smart Urbanism." *Geo-Journal* 79, no. 1 (November): 1–14.
- Knox, Hannah and Antonia Walford. 2016. "Digital Ontology." <https://culanth.org/fieldsights/820-digital-ontology> (Date Last Accessed, August, 4, 2017).
- Landow, George P. 2006. *Hypertext 3.0: Critical Theory and New Media in an Era of Globalization*. Baltimore, MD: Johns Hopkins University Press.
- Larkin, Brian. 2013. "The Politics and Poetics of Infrastructure." *Annual Review of Anthropology* 42, no. 1 (October): 327–343.
- Lassila, Ora and Deborah McGuinness. 2001. *The Role of Frame-Based Representation on the Semantic Web*. Technical Report Knowledge Systems Laboratory Report KSL-01-02: Stanford University.
- Latour, Bruno. 1986. "Visualization and Cognition: Drawing Things Together." In *Knowledge and Society: Studies in the Sociology of Culture Past and Present*, edited by H. Kuklick, 1–40. Stamford, CT: Jai Press.
- Latour, Bruno. 1991. *We Have Never Been Modern*. Cambridge, MA: Harvard University Press.
- Latour, Bruno. 2013. *An Inquiry Into Modes of Existence*. Cambridge, MA: Harvard University Press.
- Lenat, Douglas B., Mayank Prakash, and Mary Shepherd. 1985. "CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks." *AI Magazine* 6, no. 4 (March): 65–85.
- Lenzerini, Maurizio. 2002. "Data Integration: A Theoretical Perspective." In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* 233–246. New York, NY: ACM.
- Leonelli, Sabina. 2010. "Documenting the Emergence of Bio-Ontologies: Or, Why Researching Bioinformatics Requires HPSSB." *History and Philosophy of the Life Sciences* (1): 105–125.

- Leonelli, Sabina. 2012. "When Humans are the Exception: Cross-Species Databases at the Interface of Biological and Clinical Research." *Social Studies of Science* 42, no. 2 (February): 214–236.
- Levesque, Hector J. 1984. "Foundations of a Functional Approach to Knowledge Representation." *Artificial Intelligence* 23, no. 2 (July): 155–212.
- Levesque, Hector J. and Ronald J. Brachman. 1987. "Expressiveness and Tractability in Knowledge Representation and Reasoning." *Computational Intelligence* 3, no. 1 (February): 78–93.
- Lohr, Steve and Natasha Singer. 2016. "How Data Failed Us in Calling an Election." *The New York Times*,. <https://www.nytimes.com/2016/11/10/technology/the-data-said-clinton-would-win-why-you-shouldnt-have-believed-it.html> (Date Last Accessed, January, 19, 2018).
- Long, Nicholas. 1973. "Information and Referral Services: A Short History and Some Recommendations." *Social Service Review* 47, no. 1 (March): 49–62.
- Long, Nicholas, Jacqueline Anderson, Reginald Burd, Mary Elizabeth Mathis, and Sheldon P. Todd. 1971. *Information and Referral Centers: A Functional Analysis*. Technical report, Minneapolis, MN: American Rehabilitation Foundation. Institute for Interdisciplinary Studies.
- MacGregor, Robert. 1991. "Inside the LOOM Description Classifier." *SIGART Bulletin* 2, no. 3 (June): 88–92.
- Marchi, Scott de and James T. Hamilton. 2006. "Assessing the Accuracy of Self-Reported Data: an Evaluation of the Toxics Release Inventory." *Journal of Risk and Uncertainty* 32, no. 1 (January): 57–76.
- Marcus, George E. 1995. "Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography." *Annual Review of Anthropology* 24, (January): 95–117.
- Marcus, George E. and Michael M. J. Fischer. 1986. *Anthropology as Cultural Critique: An Experimental Moment in the Human Sciences*. Chicago, IL: University of Chicago Press.
- Marino, Mark. 2006. "Critical Code Studies." *electronic book review*,. <http://www.electronicbookreview.com/thread/electropoetics/codology> (Date Last Accessed, April, 12, 2018).
- McCarthy, John and Patrick J. Hayes. 1969. "Some Philosophical Problems from the Standpoint of Artificial Intelligence." In *Machine Intelligence*, edited by B. Meltzer and Donald Michie, 463–502. Edinburgh: Edinburgh University Press.

- McCarthy, John, Marvin Minsky, Nathaniel Rochester, and Claude E. Shannon. 1955. "A Proposal fro the Dartmouth Summer Resarch Project on Artificial Intelligence."
- McCorduck, Pamela. 2004. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* (2nd ed.). Natick, MA: A K Peters/CRC Press.
- McDermott, Drew. 1987a. "AI, Logic, and the Frame Problem." In *The Frame Problem in Artificial Intelligence* 105–118. New York: Elsevier.
- McDermott, Drew. 1987b. "A Critique of Pure Reason." *Computational Intelligence* 3, no. 1 (February): 151–160.
- McGuinness, Deborah L. 2001. "Description Logics Emerge from Ivory Towers." In *Proceedings of the International Workshop on Description Logics* 201–203. Stanford, CA.
- McGuinness, Deborah L. and J. R. Wright. 1998. "An Industrial-Strength Description Logic-Based Configurator Platform." *IEEE Intelligent Systems and their Applications* 13, no. 4 (July): 69–77.
- Mendez, Dara D., Vijaya K. Hogan, and Jennifer Culhane. 2011. "Institutional Racism and Pregnancy Health: Using Home Mortgage Disclosure Act Data to Develop an Index for Mortgage Discrimination at the Community Level." *Public Health Reports* 126, no. Suppl 3 (September): 102–114.
- Miller, Paul. 2011. "The Semantic Link." <http://www.dataversity.net/the-semantic-link-episode-11-october-2011/> (Date Last Accessed, June, 3, 2015).
- Minsky, Marvin. 1967. *Computation: Finite and Infinite Machines*. Upper Saddle River, NJ: Prentice-Hall, Inc.
- Minsky, Marvin. 1974. "A Framework for Representing Knowledge." .. <http://hdl.handle.net/1721.1/6089> (Date Last Accessed, April, 18, 2018).
- Montfort, Nick, Patsy Baudoin, John Bell, Ian Bogost, Jeremy Douglass, Mark C. Marino, Michael Mateas, Casey Reas, Mark Sample, and Noah Vawter. 2012. *10 PRINT CHR\$(205.5+RND(1)); : GOTO 10*. Bellingham, WA: MIT Press.
- Morville, Peter. 2011. "Information Architect." [http://semanticstudios.com/information\\_architect/](http://semanticstudios.com/information_architect/) (Date Last Accessed, January 5, 2018).
- Morville, Peter and Louis Rosenfeld. 2006. *Information Architecture for the World Wide Web: Designing Large-Scale Web Sites*. Sebastopol, CA: O'Reilly Media, Inc.

- MSNBC. 2016. “The Election Night ‘No One Saw Coming.’” <https://www.youtube.com/watch?v=MhT5qT116wo> (Date Last Accessed, April, 18, 2018).
- Newell, A. and H. A Simon. 1961. *GPS, A Program that Simulates Human Thought*. Technical Report P-2257: Rand Corp. Santa Monica, CA.
- Norvig, Peter. 2016. “The Semantic Web and the Semantics of the Web: Where Does Meaning Come From?” Keynote Delivered at the WWW2016 Conference.
- Nye, Andrea. 1990. *Words of Power: A Feminist Reading of the History of Logic*. New York: Routledge.
- Patel-Schneider, Peter F. 1985. “A Decidable First-order Logic for Knowledge Representation.” In *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 1* 455–458. San Francisco, CA: Morgan Kaufmann Publishers Inc.
- Pickering, Andrew. 2010. *The Mangle of Practice: Time, Agency, and Science*. Chicago, IL: University of Chicago Press.
- Plumwood, Val. 1993. “The Politics of Reason: Towards a Feminist Logic.” *Australasian Journal of Philosophy* 71, no. 4 (December): 436–462.
- Poirier, Lindsay. 2017. “Devious Design: Digital Infrastructure Challenges for Experimental Ethnography.” *Design Issues* 33, no. 2 (April): 70–83.
- Poirier, Lindsay, Dominic DiFranzo, and Marie Joan Kristine Gloria. 2014. “Light Structure in the Platform for Experimental Collaborative Ethnography.” Workshop paper delivered at WebSci14.
- Poniewozik, James. 2016. “A Rudderless Night, as News Networks Struggle With a Surprise Victory.” *The New York Times*, <http://www.nytimes.com/2016/11/10/arts/television/a-rudderless-night-as-news-networks-struggle-with-a-surprise-victory.html> (Date Last Accessed, November, 9, 2016).
- Porter, Theodore M. 1996. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.
- Poster, Mark. 2001. *What’s the Matter with the Internet?* Minneapolis, MN: University of Minnesota Press.
- Quillian, M. Ross. 1968. “Semantic Memory.” In *Semantic Information Processing*, edited by Marvin Minsky, 27–70. Cambridge, MA: MIT Press.
- Quine, W. V. 1975. “On Empirically Equivalent Systems of the World.” *Erkenntnis* (1975-) 9, no. 3 (November): 313–328.

- Rheinberger, Hans-Jorg. 1998. "Experimental Systems, Graphematic Spaces." In *Inscribing Science: Scientific Texts and the Materiality of Communication*, edited by Timothy Lenoir, 285–303. Palo Alto, CA: Stanford University Press.
- Ribes, David and Geoffrey C. Bowker. 2009. "Between Meaning and Machine: Learning to Represent the Knowledge of Communities." *Information and Organization* 19, no. 4 (October): 199–217.
- Ronallo, Jason. 2012. "HTML5 Microdata and Schema.org." *The Code4Lib Journal* no. 16. <http://journal.code4lib.org/articles/6400> (Date Last Accessed, October 16, 2014).
- Russell, Stuart and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach* (1st ed.). Upper Saddle River, NJ: Prentice Hall.
- Said, Edward W. 1979. *Orientalism* (1st ed.). New York: Vintage.
- Schank, Roger C. and Robert P. Abelson. 1975. "Scripts, Plans, and Knowledge." In *Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1* 151–157. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Schopper, Herwig. 2014. "Viewpoint: The 1980s: Spurring Collaboration - CERN Courier." <http://cerncourier.com/cws/article/cern/56613> (Date Last Accessed, April 12, 2018).
- Schrock, Andrew R. 2016. "Civic Hacking as Data Activism and Advocacy: A History from Publicity to Open Government Data." *New Media & Society* 18, no. 4 (April): 581–599.
- Searle, John. 1977. "Reiterating the Differences: A Reply to Derrida." *Glyph* 1, 198–208.
- Sedgwick, Eve Kosofsky and Adam Frank. 2003. *Touching Feeling: Affect, Pedagogy, Performativity*. Duke University Press.
- Shadbolt, Nigel, Tim Berners-Lee, and Wendy Hall. 2006. "The Semantic Web Revisited." *IEEE Intelligent Systems* 21, no. 3 (May): 96–101.
- Silver, Nate. 2016. "2016 Election Forecast." <http://projects.fivethirtyeight.com/2016-election-forecast/> (Date Last Accessed, January, 19, 2018).
- Simon, Herbert A.. 1965. *The Shape of Automation for Men and Management*. New York: Harper & Row.
- Singhal, Amit. 2012. "Introducing the Knowledge Graph: things, not strings." <https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html> (Date Last Accessed, April 13, 2018).

- Spivak, Gayatri Chakravorty. 1980. "Revolutions That as Yet Have No Model: Derrida's Limited Inc." *Diacritics* 10, no. 4 (Winter): 29–49.
- Spivak, Gayatri Chakravorty. 1988. "Can the Subaltern Speak?" In *Marrxism and the Interpretation of Culture* 271–315. Champaign, IL: University of Illinois Press.
- Spivak, Gayatri Chakravorty. 1993. *Outside in the Teaching Machine*. New York: Routledge.
- Spivak, Gayatri Chakravorty. 2012. *An Aesthetic Education in the Era of Globalization*. Cambridge, MA: Harvard University Press.
- Spivak, Gayatri Chakravorty and Sarah Harasym. 1990. *The Post-Colonial Critic: Interviews, Strategies, Dialogues*. New York; London: Routledge.
- Star, Susan Leigh. 1990. "Power, Technology and the Phenomenology of Conventions: On Being Allergic to Onions." *The Sociological Review* 38, no. S1 (May): 26–56.
- Star, Susan Leigh. 1991. "Invisible Work and Silenced Dialogues in Knowledge Representation." In *Women, Work and Computerization: Understanding and Overcoming Bias in Work and Education*, edited by Inger Eriksson, Barbara Kitchenham, and Kea Tijdens, 81–92. Amsterdam, Oxford: North Holland.
- Star, Susan Leigh. 1995. "The Politics of Formal Representations: Wizards, Gurus, and Organizational Complexity." In *Ecologies of Knowledge: Work and Politics in Science and Technology* 88. Albany, NY: SUNY Press.
- Star, Susan Leigh. 1999. "The Ethnography of Infrastructure." *American Behavioral Scientist* 43, no. 3 (November): 377–391.
- Star, Susan Leigh. 2010. "This is Not a Boundary Object: Reflections on the Origin of a Concept." *Science, Technology, & Human Values* 35, no. 5 (September): 601–617.
- Star, Susan Leigh and James R. Griesemer. 1989. "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." *Social Studies of Science* 19, no. 3 (August): 387–420.
- Star, Susan Leigh and Karen Ruhleder. 1996. "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces." *Information Systems Research* 7, no. 1 (March): 111–134.
- Strom, Stephanie. 2005. "After Hurricanes, Growing Support for 211 Call Service." *The New York Times*, <https://www.nytimes.com/2005/11/20/us/nationalspecial/after-hurricanes-growing-support-for-211-call-service.html> (Date Last Accessed, October 18, 2017).

- Suchman, Lucy. 2006. *Human-Machine Reconfigurations: Plans and Situated Actions* (2nd ed.). Cambridge; New York: Cambridge University Press.
- Tarski, Alfred. 1944. "The Semantic Conception of Truth: and the Foundations of Semantics." *Philosophy and Phenomenological Research* 4, no. 3 (March): 341–376.
- Telecommunications and Information Policy Institute. 2002. *Telecommunications and 211*. Primer, Austin, TX: University of Texas at Austin.
- 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, no. 1 (August): 69–89.
- Towle, Charlotte. 1949. "The Client's Rights and the Use of the Social Service Exchange." *Social Service Review* 23, no. 1 (March): 15–20.
- Traweek, Sharon. 1992. *Beamtimes and Lifetimes: The World of High Energy Physicists*. Cambridge, MA: Harvard University Press.
- Traweek, Sharon. 2000. "Faultlines." In *Doing Science + Culture*, edited by Roddey Reid and Sharon Traweek, 21–48. New York; London: Routledge.
- United Way Worldwide. 2017. "Real People, Real Help." <http://www.211.org/pages/about> (Date Last Accessed, November 14, 2017).
- Van Heijenoort, Jean. 1967. "Logic as Calculus and Logic as Language." *Synthese* 17, no. 3 (September): 324–330.
- Veltman, Kim H. 2006. "Towards a Semantic Web for Culture." *Journal of Digital Information* 4, no. 4 (February): 1–87.
- Viveiros de Castro, Eduardo. 2002. "O Nativo Relativo." *Mana* 8, no. 1 (April): 113–148.
- Viégas, Fernanda B., Martin Wattenberg, and Kushal Dave. 2004. "Studying Cooperation and Conflict Between Authors with History Flow Visualizations." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 575–582. New York, NY, USA: ACM.
- W3C. 2012. "WebSchemas/StyleGuide - W3C Wiki." <https://www.w3.org/wiki/WebSchemas/StyleGuide> (Date Last Accessed, April 20, 2018).
- W3C Public Vocabs Group. 2011. "Re: Need for a new type Activity." <https://lists.w3.org/Archives/Public/public-vocabs/2011Nov/0003.html> (Date Last Accessed, April, 18, 2018).

- W3C Public Vocabs Group. 2013a. “Re: schema.org growth what are the limits?” <https://lists.w3.org/Archives/Public/public-vocabs/2013Jul/0116.html> (Date Last Accessed, April, 18, 2018).
- W3C Public Vocabs Group. 2013b. “Re: strange identifiers in schema.org.” <https://lists.w3.org/Archives/Public/public-vocabs/2013Nov/0033.html> (Date Last Accessed, April, 18, 2018).
- W3C RDF Logic Group. 2000. “Re: random thoughts on web logic.” <https://lists.w3.org/Archives/Public/www-rdf-logic/2000Sep/0009.html> (Date Last Accessed, April, 18, 2018).
- W3C WebOnt Working Group. 2001. “more on the relationship between RDF and DAML+OIL.” <https://lists.w3.org/Archives/Public/www-webont-wg/2001Dec/0081.html> (Date Last Accessed, April, 18, 2018).
- W3C WebOnt Working Group. 2002a. “Fwd: Re: CHAIR-NOTE: Defaults and etc.” <https://lists.w3.org/Archives/Public/www-webont-wg/2002Jan/0170.html> (Date Last Accessed, April, 18, 2018).
- W3C WebOnt Working Group. 2002b. “new names for OWL lite/fast/large.” <https://lists.w3.org/Archives/Public/www-webont-wg/2002Oct/0129.html> (Date Last Accessed, April, 18, 2018).
- Waller, Vivienne. 2016. “Making Knowledge Machine-Processable: Some Implications of General Semantic Search.” *Behaviour & Information Technology* 35, no. 10 (May): 1–12.
- White, Hayden. 1975. *Metahistory: The Historical Imagination in Nineteenth-Century Europe*. Baltimore, MD: Johns Hopkins University Press.
- Whitehead, Alan N. and Bertrand B. Russell. 1925. *Principia Mathematica*. Cambridge, England: The University Press.
- Wilson, Elizabeth A. 2010. *Affect and Artificial Intelligence*. Seattle, WA: University of Washington Press.
- Winner, Langdon. 1986. “Do Artifacts Have Politics?” In *The Whale and the Re-actor: A Search for Limits in an Age of High Technology* 19–39. Chicago, IL: University of Chicago Press.
- Winner, Langdon. 1993. “Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology.” *Science, Technology, & Human Values* 18, no. 3 (Summer): 362–378.
- Winograd, Terry. 1975. “Frame Representations and the Declarative/Procedural Controversy.” In *Representation and Understanding: Studies in Cognitive Science*, edited by Daniel Bobrow and Allan Collins, 185–210. New York: Academic Press, Inc.

- Winograd, Terry. 1980. "What Does it Mean to Understand Language?" *Cognitive Science* 4, no. 3 (July): 209–241.
- Wittgenstein, Ludwig. 1922. *Tractatus Logico-philosophicus*. San Diego, CA: Harcourt, Brace, Incorporated.
- Wodtke, Christina. 2001. "Defining the Damn Thing." <http://eleganthack.com/defining-the-damn-thing/> (Date Last Accessed, April 18, 2018).
- Zins, Chaim. 2001. "Defining Human Services." *Journal of Sociology & Social Welfare* 28, (March): 3–21.