Powerful Structure: Inspecting Infrastructures of Information Organization in Wikimedia Foundation Projects

Katherine Thornton

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Allyson Carlyle, Chair

David McDonald

Jin Ha Lee

Barbara H. Kwaśnik

Program Authorized to Offer Degree:
Information Science

University of Washington

Abstract

Powerful Structure: Inspecting Infrastructures of Information Organization in Wikimedia Foundation Projects

Katherine Thornton

Chair of the Supervisory Committee:
Associate Professor Allyson Carlyle
Information School

This dissertation investigates the social and technological factors of collaboratively organizing information in commons-based peer production systems. To do so, it analyzes the diverse strategies that members of Wikimedia Foundation (WMF) project communities use to organize information.

Key findings from this dissertation show that conceptual structures of information organization are encoded into the infrastructure of WMF projects. The fact that WMF projects are commons-based peer production systems means that we can inspect the code that enables these systems, but a specific type of technical literacy is required to do so.

I use three methods in this dissertation. I conduct a qualitative content analysis of the discussions surrounding the design, implementation and evaluation of the category system; a quantitative analysis using descriptive statistics of patterns of editing among editors who contributed to the code of templates for information boxes; and a close reading of the infrastructure used to create the category system, the infobox templates, and the knowledge base of structured data.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

STRUCTURE OF INFORMATION ORGANIZATION: I use the phrase 'structure of information organization' to refer to a specific piece of infrastructure from a specific computational system. For example, the category system of English Wikipedia is a structure of information organization.

KNOWLEDGE ORGANIZATION SYSTEM (KOS): Knowledge organization systems include classification schemes, category schemes, subject headings, authority files, highly-structured vocabularies, thesauri, semantic networks and ontologies [Hodge et al., 2003].

KNOWLEDGE GRAPH: I follow Sowa who defines graphs saying: "Formally a *graph* G consists of a nonempty set N, whose elements are called *nodes*, and a set A, whose elements are called *arcs*" [Sowa, 1983, 375].

KNOWLEDGE REPRESENTATION RESOURCES(KRR): I follow [Wright, 2007] who defines KRR as an umbrella term for all KOS as well as the mark-up languages through which KOS are made part of computational systems.

CONCEPTUAL STRUCTURE OF INFORMATION ORGANIZATION: A conceptual structure of information organization refers to the entities and the relationships between entities in the abstract, before they have been instantiated in a specific computational system. This definition was informed by [Sowa, 1983].

ONTOLOGY: I follow the definition of Noy and McGuinness:

> "An ontology is a formal explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions))" [Noy and McGuinness, 2001, 3].

Put simply, an ontology is a structure that defines and makes explicit entities (named and operationalized) and the relationships (named and operationalized) between entities.

DATA MODEL: In a generic sense, a data model is an abstract specification that describes data and how they are related. In a specific sense, the data model I refer to most frequently in this text is the data model of Wikidata [Vrandečić and Krötzsch, 2014].

INFRASTRUCTURE: Infrastructure encompasses all of the code that enables the software and technologies of a computational system.

WIKITEXT: Wikitext is markup that instructs the browser how to render the content, similar to HTML [MediaWiki, 2016].

ACKNOWLEDGMENTS

Of the three research projects that I describe in this dissertation, two are the result of collaborative work. The project described in Chapter Four is the result of collaborative work with David W. McDonald. The project described in Chapter Five is the result of collaborative work with David W. McDonald and Martez Mott.

I was supported for two academic years by the National Science Foundation Award IIS-1162114 led by PI Mark Zachry and David W. McDonald. The research we performed under this grant funding involved many discussions which had fruitful overlap with the theoretical concepts discussed in the research I describe in Chapter 6.

DEDICATION

for Ally

Chapter 1

WHERE THE CROWD STRUCTURES DATA

1.1 Overview of the Research Territory

Recently we have seen a flurry of publications related to the topic of artificial intelligence (AI). From the book *Superintelligence* [Bostrom, 2014], to *The Intelligent Web* [Shroff, 2013], to *Decoding Reality* [Vedral, 2010], to *Who Owns the Future* [Lanier, 2014], to *The Most Human Human* [Christian, 2012], these books trace the multiple histories of our technologies of, research of, ethics of, and responses to, artificial intelligence in the 20th and 21st centuries. The authors of these texts make arguments such as that of Schroff:

> "Applying web-scale computing power on the vast volume of 'big data' now available because of the internet, offers the potential to create far more intelligent systems than ever before: this defines the new science of web intelligence" [Shroff, 2013, xix].

What might these intelligent systems look like? Bostrom, a member of the faculty of philosophy at the University of Oxford, who believes that humans are currently creating the seeds of superintelligence, states:

> "Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct. Superintelligence is a challenge for which we are not yet ready now and will not be ready for a long time. We have little idea when the detonation will occur, though if we hold the device to our ear we can hear a faint ticking sound" [Bostrom, 2014, 259].

The emphasis some of these authors place on the unknowability of the power of the technologies of AI is intended to persuade readers that this is a topic worthy of immediate collective human attention and action. Each of these accounts discuss how the amount of data (both structured and unstructured) that is available on the web has impacted the recent developments in artificial intelligence. While each of these texts certainly discuss the value of structured data[1], none discuss the fact that much of the data, the existence of which is a prerequisite for machine intelligence, is currently human-sourced and human-structured. The details of this process are obscured by the distributed, behind-the-scenes nature of how sets of structured data are currently being reused across the web between the locations where they are being created, and the locations where companies create revenue streams built to monetize slices of structured data. Similar to the importance of immediate collective human intelligence to questions around our development AI, I describe how decisions about how to design knowledge bases of structured data—crucial components of AI systems—also demand consideration.

One type of repository for structured data is the knowledge base. A knowledge base is a genre of information system, composed of stores of structured data and ontologies that describe the data [Färber et al., 2015]. Many of the largest companies in the world are currently investing in the creation of knowledge bases. For example, Google has announced creation of a Knowledge Vault [Dong et al., 2014a]. Intelligent agents such as Apple's Siri, Amazon's Echo, or Microsoft's Cortana use knowledge bases to provide answers to their users' queries [Bissig, 2015]. The knowledge bases being created by for-profit companies are closed to our inspection due to issues of intellectual property. The largest openly available knowledge base of structured data is Wikidata, a project of the Wikimedia Foundation. This knowledge base of structured data is a product of the Wikimedia community. It is collaboratively-built and enabled by free software. I investigate three structures

[1]Structured data is data that is expressed in a syntax that communicates how data is interconnected. It is both machine-readable and human-readable. Structured data about the world is a necessary elements of all systems of artificial intelligence.

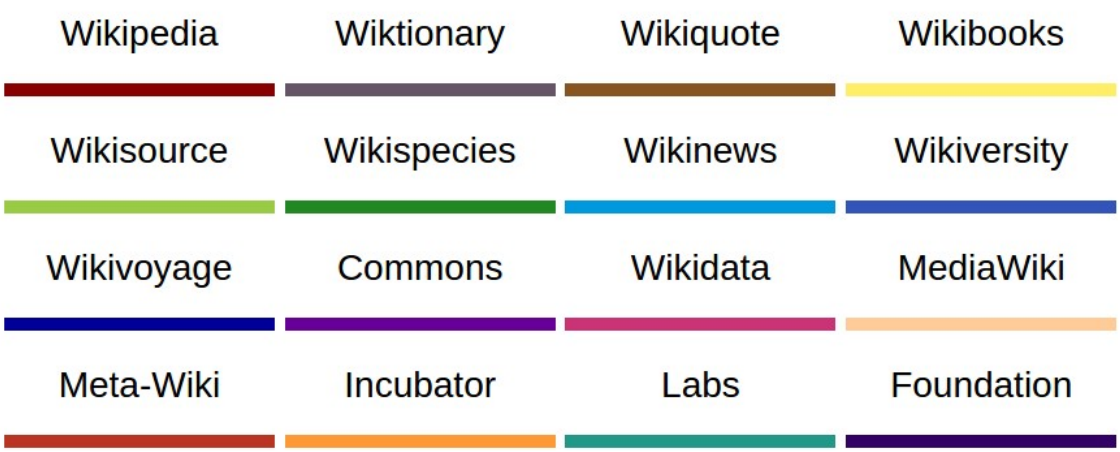

Figure 1.1: Projects of the Wikimedia Foundation

used by the Wikimedia community to organize information: the category system, the templates for information boxes and the knowledge base of structured data.

Each of the structures that I select to analyze are part of the infrastructure of WMF projects. As illustrated in Figure 1.1, there are currently sixteen projects within the Wikimedia Foundation, however Wikipedia is actually two hundred and ninety-four different language Wikipedias [Wikipedia, 2016]. Infrastructure is defined as the "pervasive enabling resources in network form" that make computational systems possible [Bowker et al., 2010, 98]. Many of us are familiar with the infrastructure of the built environment. For example, transportation infrastructure includes roads, ports, railways, airports, etc. The infrastructure of a building might be the plumbing system, the ductwork, the electrical wiring, hookups to utilities, etc. The architecture of the Centre Georges-Pompidou in Paris, France, see Figure 1.2, expresses a conceptualization of making infrastructure visible [Proto, 2005]. The building infrastructure that would normally be enclosed between the building exterior and the interior walls was placed outside of the building and color coded to indicate the function of the system [Georges-Pompidou, 2016].

Figure 1.2: Inside Out Infrastructure Expressed Architecturally

I pose the question of how we might make infrastructure of a commons-based peer production system legible. The infrastructure of peer production systems poses challenges for those who might want to read it or even lay eyes on it. This is due to the fact that infrastructure is often purposely hidden from view, just as the majority of architects conceal infrastructure within opaque interior walls. In order to read the infrastructure of a peer production system one must have the ability to read the code used to program the technologies that make up the system. When I say that I am in search of ways to make infrastructure *legible*, I mean that I would like to make it, first, visible and, second, understandable in terms of what conceptual systems it encodes. In this research I describe the details of the infrastructure across three sites related to how peer production communities undertake information organization: the category system, the infobox templates, and the knowledge base of structured data. Each of these subsets of infrastruc-

ture also encode conceptual systems related to information organization, I argue that the design of these conceptual systems has implications for information organization.

In the remainder of this chapter, I review the concepts of information organization and data structuring in the context of Wikimedia Foundation projects in order to introduce the site of the research activities. I provide an overview of the role the Wikimedia community plays in the provision of infrastructure and workflows for the creation of free-to-reuse structured data on the web. I introduce the research questions for each of the three projects that comprise this dissertation.

1.2 Wikimedia Projects as the Site of Research Activities

The Wikimedia Foundation (WMF) is an umbrella organization that owns the domains for WMF projects, provides hosting for WMF projects and raises funds through its status as a non-profit and charitable organization [Wikipedia, 2016]. The most well-known of the WMF projects is Wikipedia, an online encyclopedia created entirely of user-generated content [Morell, 2011]. WMF projects are an especially promising venue for the analysis of structures of information organization for several reasons. First, the large majority of the work in WMF projects is recorded in the publicly-available edit history which provides the opportunity to uncover the structures themselves in the context of the discussions and edits of the community members who created the structures [Benkler et al., 2013]. Second, because WMF project content is freely licensed and widely reused, it is likely that what we learn about how structures of information organization are created and used across WMF projects, will also help us understand how these structures play a role in the larger contexts of the web and the semantic web. I introduce three of these structures of information organization in detail: the category system, the infoboxes, and the knowledge base.

1.2.1 Category System

The category system is a structure of information organization that allows editors to group content conceptually. In 2003, roughly two years after Wikipedia began, the community decided to create a category system to organize and tag the content of the site [Voss, 2006]. The category system has changed over time, as have conceptualizations of the role it should serve in Wikipedia. The category system was proposed, designed, and implemented entirely by members of the Wikipedia community, those who edit Wikipedia. Anyone with access to the internet can edit Wikipedia. Any Wikipedia editor has the ability to apply category labels to pages, to remove category labels from pages, to create new categories, and to suggest categories to be considered for deletion. The incremental additions to, revisions of, and deletions of category labels to pages make up the category system. Discussions of this work take place on the talk pages of the wiki, such as the talk page for `Categorization` [Wikipedia, 2016b].

I investigate the category system in terms of how the community collaborated to create it. If decisions about the organization of information are subjective, how does a peer production community coordinate these activities?

1.2.2 Infoboxes

Infoboxes are structures of information organization that provide quick facts about a topic. Infoboxes are presented graphically on the right-hand side of articles (see Figure 1.3). This infobox has an image of a nautilus and information related to the classification of the organism. The Wikimedia community describes the purpose of information boxes to be tools to support readers [Wikipedia, 2015].

As the community of editors of Wikipedia embraced the infobox structure, they created numerous infoboxes on article pages. Infoboxes are created on pages by applying markup for a template into the wikitext [2] of an article. Due to the way at-

[2]Wikitext is the markup language of the MediaWiki software. Wikitext provides instructions to web browsers about how to render the content.

Figure 1.3: Nautilus Infobox

tributes and their values are described in markup, each information box also represents a chunk of structured data presented in a machine readable format. As the number of information boxes grew, these chunks of machine readable structured data accumulated, and people recognized that alongside the wealth of unstructured information and data in Wikipedia, there exists a growing store of structured data. This structured data is highly-desirable as an increasing number of organizations build applications (such as automated agents to answer questions) that require large corpora of basic facts in machine-readable format. Thus WMF projects are the site of information organization and data structuring activities that produce the structured data that powers many applications we use today.

I investigate infoboxes due to the fact that they were the first targets for data mining efforts to extract structured data from Wikipedia [Erxleben et al., 2014, Sen et al., 2014, Bao et al., 2012]. Through this investigation I was able to gain contextual knowledge of the mechanism of infobox creation in the MediaWiki software known as transclusion. I was able to understand how edits can be made that affect more than one unit of data at time.

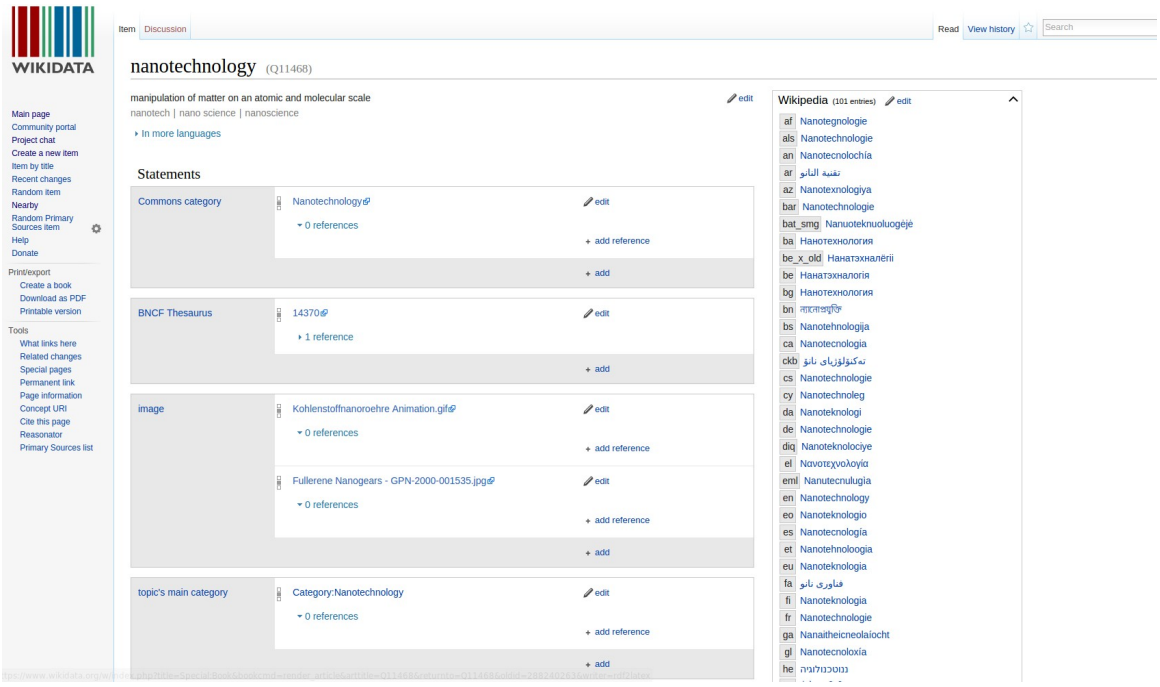

Figure 1.4: Page for Wikidata entity Q114678 'nanotechnology'

1.2.3 Wikidata: a Knowledge Base of Structured Data

The knowledge base of structured data, Wikidata, is a structure of information organization that combines a data model with structured data. Editors curate and provide source information for structured data [Ismayilov et al.,] in this project. Wikidata contains data about entities structured in a way that is machine-readable as well as human-readable [Erxleben et al., 2014]. As the data management platform of all Wikimedia Foundation projects [Erxleben et al., 2014], the data is free and open for reuse within all WMF projects, and also is freely available for reuse outside of WMF projects.

In Figure 1.4, we see a screenshot of the Wikidata page for entity Q114678 'nanotechnology'. Each item is allotted a page in Wikidata and has a unique identifier, with prefix *Q* plus a string of numbers, ex. Q114678, which is assigned to the item *nanotechnology*. In Figure 1.4 we see four statements about *nanotech-*

nology. Each of these statements expressed a property of the item Q114678[3]. There is a statement that the *commons category* for this entity has a value of 'nanotechnology'. This statement is a link to the category that this entity belongs to the sister project Wikimedia Commons, a repository of image and media files for WMF projects. This is structured data because not only do we know the value for this claim about the entity *nanotechnology*, we also know the relationship between them is that of *commons category*[4].

Wikidata is a knowledge base containing many millions of such items, each described by a set of statements. These pages are open for editing by anonymous IP, or by registering for an account as an editor. In addition to being a data management platform for Wikimedia projects, Wikidata itself is a commons-based peer production system [Benkler, 2002, Benkler, 2006, Müller-Birn et al., 2015]. The Wikidata community collectively structures data. The fact that the Wikidata project is an example commons-based peer production is salient for those interested in the inspectability of the infrastructure of the semantic web, because the Wikimedia community claims to hold itself accountable to the values of freedom and transparency [Jemielniak, 2014]. Inspectability is a crucial dimension of infrastructure because infrastructure is often hidden [Star and Bowker, 2006]. Those who seek to understand how organizing systems are instantiated in infrastructure must be able to inspect the infrastructure to discover how it operates.

I investigate Wikidata because of the fact that the edit histories are available for review, and the code for the enabling software is also inspectable. Thus we can see how the software enables and constrains the expression of structures of information organization. I speculate that Wikidata will become the most-widely reused source of structured data on the web. Thus, what we learn about the

---

[3]Each of these properties has a identifier assigned to it. The property identifiers all begin with prefix *P* plus a string of numbers.

[4]The screenshot in Figure 1.4 is from the graphical user interface intended for human editors. A program called a bot would interact with Wikidata via the Application Programming Interface of the MediaWiki software. In 2014, Steiner found that eighty-eight percent of the activity in Wikidata is performed by bots [Steiner, 2014]

inspectability of the structures of information organization in Wikidata may serve as a key to understanding how this data is reused in proprietary systems that we may never have the opportunity to inspect. The work of human decision making about how to structure data is taking place in many knowledge bases. Open, inspectable knowledge bases, like Wikidata, allow us to understand this work of information organization and data structuring.

1.3 Structures of Information Organization

Knowledge organization systems (KOS) are conceptual structures used to organize information. I follow [Gödert et al., 2014] in their conceptualization of how KOS are leveraged by the programmers of computational systems.

Knowledge organization systems (KOS) include classification schemes, category schemes, subject headings, authority files, highly-structured vocabularies, thesauri, semantic networks and ontologies [Hodge et al., 2003]. Each of these systems is designed to represent conceptual information, and to document the relationships between pieces of information. Hodge states that this definition, created by participants at the 'Networked Knowledge Organization Systems Working Group' meeting, dates from 1998. In this dissertation I will refer to knowledge organization systems as 'conceptual infrastructure' for information organization.

In this way, through reference to Figure 1.3, I will disambiguate the conceptual and the implementation level. I refer to infrastructure on the conceptual level as 'conceptual infrastructure' and I refer to infrastructure on the implementation level as 'technical infrastructure'. I do not intend to imply that conceptual infrastructure can be cleanly separated out from infrastructure, but rather to emphasize that conceptual structures such as classification schemes, authority files, and thesauri, may be encoded into technical infrastructure in a variety of ways. These decisions are recorded in the edit histories of commons-based peer production systems, but in proprietary systems these decisions are not inspectable.

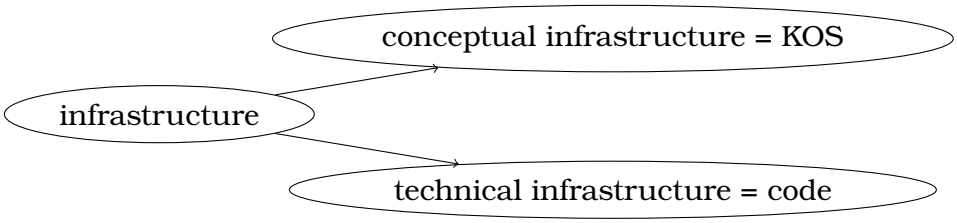

Figure 1.5: Separation of conceptual infrastructure from infrastructure

In this project I describe a unit of analysis of 'structure of information organization'. A 'structure of information organization' is the technical structure in which a knowledge organization system (or parts of one or more system(s)) are instantiated. When I use the term 'structure of information organization,' I am using this term to indicate a specific implementation within a specific computational system. Structures of information organization exist at the implementation level. For this project, all of the structures of information organization are specific to WMF projects.

The technical infrastructure of a computational system is the code that enables the system. While there may be multiple ways to represent a conceptual structure in a computational system, the code contains evidence of human decisions about the details of the choices related to representation. Lessig theorizes code as the architecture of control in computational contexts [Lessig, 2006]. He goes on to state that technologies are optimized for governmental regulation to greater or lesser degrees depending on how they are implemented through code. Similarly, within socio-technical systems, code is the architecture of control for the structures of information organization. Because code is the architecture of control that governs how structures of information organization are represented in computational systems, we must be able to inspect code in order to discover what conceptual relationships exist in the system and how they operate. We know that KOS are never neutral, and that it is the human decisions about how to organize information that may privilege certain claims or ways of knowing over others [Olson, 2013, Feinberg, 2007]. Thus the ability to inspect the places where these decisions are encoded is

the key to auditing structured data for issues of bias.

1.4 The Problem: How Do We Inspect Infrastructure to Discover Structures of Information Organization

Structures of information organization are specific subsets of infrastructure. For example, in Wikipedia we see the category system and information boxes as some of the most commonly-used structures of information organization. Each of these structures is created in Wikitext. Wikitext is markup that instructs the browser how to render the content, similar to HTML [MediaWiki, 2016]. What is encoded in these structures is the conceptual infrastructure of information organization that describe how content is related in Wikipedia.

In the domain of computational systems, the concept of infrastructure is used to describe technologies that support information systems. Information infrastructure thus has an important relationship with the structures and functions of the information systems it supports by making certain relationships expressible and others inexpressible [Ribes et al., 2013, 3].

Star and Ruhleder note that infrastructure is often invisible, and because of this, many people take it for granted [Star and Ruhleder, 1994]. By referring to infrastructure as 'invisible' these authors highlight the fact that infrastructure is often purposely designed to be available only to those who are building or repairing it. For example, the infrastructure of the search algorithms used by Google are not made visible to users of the search engine. The database structure of Amazon.com is not made visible to visitors of the website. Due to the fact that many systems make it difficult to observe infrastructure, it can be difficult to study it. The fact that many people take infrastructure for granted is one reason that researchers infrequently make infrastructure an object of study.

Although it is challenging, Bowker and collaborators have demonstrated that infrastructure can be studied [Bowker et al., 2010, 98]. I use the dimensions of infrastructure articulated in [Star and Ruhleder, 1996, Star, 1999, Star and Ruhleder, 1994, Star, 2010] to analyze the infrastructure of information organization

in Wikimedia projects. Star outlines her recommendations for how to undertake an ethnography of infrastructure because of the need she saw for "...new methods to understand this imbrication of infrastructure and human organization" [Star, 1999, 379]. I base my analysis of the infrastructure of Wikimedia projects in Star's recommendations. I use these techniques to allow me to perform a close reading of the infrastructural components of these systems in order to discover how the infrastructure makes the expression of certain relationships and claims possible to express while making others impossible to express. Star discusses how ethnography can be applied to information systems saying:

> "Information systems encode and embed work in several ways. They may directly attempt to represent that work. They may sit in the middle of a work process like a rock in a stream, and require workarounds in order that interaction proceed around them. They also may leave gaps in work processes that require real-time adjustments, or articulation work, to complete the processes. Finding the invisible work in information systems requires looking for these processes in the traces left behind by coders, designers, and users of systems" [Star, 1999, 385].

Due to the records of editing history that are preserved in the wiki architecture of WMF projects I have full access to the traces left behind by coders designer and users of these projects. Star argues that an ethnographic sensibility allows one to pay attention to: "an idea that people make meanings based on their circumstances, and that these meanings would be inscribed into their judgments about the built information environment" [Star, 1999, 383]. It is precisely a subset of these judgements, those that pertain to how information is organized, and the structures that are used to accomplish this work, and the ways in which these structures are instantiated into the infrastructure of WMF projects that I observe.

Star (1999), and Star in conjunction with collaborators [Star and Ruhleder, 1996, Star and Bowker, 2006] advanced and refined a theory of how to analyze infrastructure. They state that the following dimensions can be found in infras-

tructure: embeddedness, transparency, reach or scope, learned as part of membership, links with conventions of practice, embodiment of standards, built on an installed base, becomes visible upon breakdown, is fixed in modular increments [Star and Ruhleder, 1996, Star and Bowker, 2006, Star, 2010]. I use these dimensions to identify how infrastructure plays an important role in constraining the expressivity of knowledge organization systems, which I understand to be the conceptual infrastructure. I apply these dimensions across several sites where information organization work is carried out in WMF projects. I also use these dimensions to identify how these sites are also examples of infrastructure when regarded from the position of applications external to the context of WMF projects.

1.5 The Research

Speaking to recent work in the our shared understandings of the role that structured data plays in commons-based peer production systems, artificial intelligence and the semantic web, it is relevant to understand the context in which people performed the work of data structuration. I chose to undertake analysis of the infrastructure of WMF projects. I made this choice because of the frequent reuse of structured data that had been curated in the context of WMF projects [Benkler et al., 2013, Müller-Birn et al., 2015, Bizer et al., 2009, Bao et al., 2012, Sen et al., 2014].

The infrastructure used to create sets of structured data, and ultimately repositories of heterogeneous sets of structured data, involves making design decisions which allow for some types of knowledge to be expressed and makes some types of knowledge inexpressible.

1.5.1 Research Questions

I formulate a set of motivating questions for this dissertation:

1. How does collaboration occur in relation to the creation of the category system?

2. What do editors see as the purpose of the category system?

3. How are infoboxes created?

4. How can we measure and distinguish different styles of infobox template editing?

5. In what ways does the knowledge base manifest as infrastructure?

6. How are KOS instantiated in infrastructure?

I address these research questions through three related research projects which I will describe in detail.

1.5.2 Overview of Remaining Chapters

The following chapters of my dissertation consist of: Literature Review, Research Activities, Category System, Templates, Close Readings of Infrastructure, Implications, and Conclusion.

Chapter 2

LITERATURE REVIEW

The three research projects that comprise this dissertation concern information organization within Wikimedia Foundation (WMF) projects. WMF projects are examples of commons-based peer production systems, in which community members engage in social computing. My research questions focus on the infrastructure of these projects and the conceptual structures of information organization encoded therein. In order to advance arguments about the implications of how these structures of information organization are encoded, I will build upon work in social computing, knowledge organization and semantic web technologies.

2.1 Social Computing

Social computing involves interaction with others while making use of networked computational systems. Parameswaran and Whinston define social computing saying that it consists of:

> "...applications and services that facilitate collective action and social interaction online with rich exchange of multimedia information and evolution of aggregate knowledge" [Parameswaran and Whinston, 2007, 762].

WMF projects are examples of systems that enable social computing. Members of WMF projects engage in collective action by working together to contribute to the elaboration of these projects. This interaction is mediated via members' web-based participation around text, images, code, etc. WMF communities collaborate via social computing to create content and infrastructure and to encode conceptual structures of information organization into these projects.

When people interact with one another in a way that is mediated by computer software they are engaging in social computing [Schuler, 1994]. Examples of platforms for social computing are wikis, blogs, podcasts, and social networking systems [Roush, 2005]. Systems that support social computing are enabled by software, some of which is proprietary, some of which is open-source, and some of it is free software. The software that is used to create WMF projects is free software.

2.1.1 Free-Libre Open Source Software

A key to understanding the Free-Libre Open Source Software (FLOSS) movement is represented in the movement's conceptualization of "free-libre" to refer to their beliefs about freedom [Stallman, 2009]. The word "free" here indicates that users can inspect, modify, and redistribute the source code of the software [Stallman and Lessig, 2010]. In this section, I will discuss the values and philosophy of the free software movement and explain how they pertain to our ability to inspect the infrastructure of the WMF projects.

Free software projects are envisioned and enacted as alternatives to proprietary software. The software development processes are distinct from those of companies creating proprietary software.

> "This knowledge and culture of how to shape collaboration allows us to build outstanding technology in small, distributed teams across language, country and cultural barriers around the world, outperforming much larger development teams in some of the world's largest corporations" [Greve, 2012, ix].

The Free Software Foundation (FSF) is a organizational, theoretical, and logistical hub of activity for the Free Software movement. The FSF publishes information pertaining to the values that drive their movement both on their website and through the distribution of books. The FSF established and refined evaluative criteria for software projects. The FSF articulates four essential freedoms that must be ensured in order for a piece of software to qualify under their definition of free:

- The freedom to run the program as you wish, for any purpose.
- The freedom to study how the program works, and change it so it does your computing as you wish. Access to the source code is a precondition for this.
- The freedom to redistribute copies so you can help your neighbor.
- The freedom to distribute copies of your modified versions to others. By doing this you can give the whole community a chance to benefit from your changes. Access to the source code is a precondition for this [Foundation, 1996, 1].

The software that are used to run the wikis that enable WMF projects is free software. These are collaboratively-built software applications maintained by communities and many of the members of these communities are volunteers. The free software context in which this software is created is one of the primary reasons that the infrastructure of WMF projects is inspectable. The freedom to study how a program works is the freedom that ensures that it will be possible for users to determine how information is structured in a program. The freedom to distribute copies of this software is a primary reason that it is widely used. While researchers may never have the opportunity to inspect the code that enables proprietary knowledge bases, we do have the ability to inspect the code created in free software projects.

2.1.2 Wiki Technology

Each of the WMF projects is created through the use of wiki technology. In this section I explain the basics of wiki technology. Wikis are a type of software used to allow networked collaboration and creation. Wiki technology was first implemented in the early 1990s by Ward Cunningham [Koren, 2012]. Wikis are open to editing by users, and allow content to be connected to other parts of the wiki programmatically. In his review of the history of MediaWiki, a specific suite of software used to create wikis, Koren states:

> "There was no great distinction for the first five years or so between the code used to run a wiki and the content on it, partly because there was nearly a 1:1 correspondence between the two: many of the original wiki administrators were programmers, and they tended to create their own new, or modified, version of the software to run their own wikis" [Koren, 2012, 2].

This quote brings up an interesting feature of the early days of wiki creation when technologists writing code to enable these systems were also the primary content creators. Today there are many content contributors who may never edit the infrastructure of the wiki.

Priedhorsky discusses the need for what they call "transparent changes" saying "It must be easy for all readers to see how the wiki is being changed, and by whom" [Priedhorsky and Terveen, 2011]. This makes the community very valuable as a site where the creation of infrastructure is more inspectable precisely because the architecture of the wiki makes transparency a priority.

Inspectability of the infrastructure of WMF projects allows for the possibility that users can either change the infrastructure or make requests that infrastructure be changed. Jemielniak states that Wikipedia exists as an alternative to the capitalist orientation of corporate involvement in information online systems [Jemielniak, 2014]. Jemielniak also notes that we must also pay attention to the degree to which Wikipedia may reenact established systems of knowledge production. Only through holding ourselves accountable to inspecting the decisions our communities have made about how information is represented and data is structured can we ensure that we are indeed an alternative infrastructure of the semantic web.

A wiki is a particular type of information system:

> "A *wiki* is a website that lets people freely create, edit, and link a collection of articles. Now, every website can be considered a bunch of

> interlinked pages, but wikis allow the the content and the structure to be changed by a community" [Barrett, 2008, 2].

The wiki software used in WMF projects is called MediaWiki. The MediaWiki software itself is a FLOSS project [med, 2016]. The fact that MediaWiki software is a FLOSS project means that MediaWiki code is open to inspection. In short, the content created by WMF projects is inspectable, and the infrastructure of these projects is also inspectable. The communities that create these projects have shared values of transparency and openness.

2.1.3 Commons-based Peer Production

In this section I explain how FLOSS and wikis come together in WMF projects, and how the domain of commons-based peer production is theorized. Benkler introduced the term commons-based peer production to capture the characteristics that make certain collaborations productive in a digitally networked environment [Benkler, 2002, 374]. Benkler differentiates this type of production noting that the labor is voluntary and contributors have diverse motivations for participation. He calls these systems commons-based because the products of these communities are designated under licenses that make them reusable by all, designating them to be part of humanity's shared intellectual commons. The examples Benkler provides in his analysis are drawn from the community that develops the GNU/Linux operating system as situated in the Free Software movement [Benkler, 2002].

Hill and Shaw apply theories of commons-based peer production to a range of communities that create wikis, and find that oligarchic organizational structures and behavior typify these communities. This finding stands in contrast to the many claims that these communities employ democratic organizational forms [Shaw and Hill, 2014]. Based on their findings, I conclude that it is important to examine the infrastructure of commons-based peer production systems to discover the conceptual structures of information organization, as there is a risk that these structures may represent the interests or beliefs of powerful leaders within

the community as opposed to the views and beliefs of members who do not occupy leadership roles.

In her analysis of a successful FLOSS community, Coleman describes the tension between the leadership of the community and the community as a whole saying:

> "If democratic rule is sometimes treated with overt suspicion and dislike, there is a far more subtle fear concerning the importance of meritocracy and the meritocrats it produces–namely, the fear of corruption. Specifically, there is discomfort with the idea that the technical guardians could (as they are vested to do) exercise their authority without consulting the project as a whole, thereby foreclosing precisely the neutral, technical debate that allowed them to gain their authority in the first place" [Coleman, 2013, 127].

Scholars researching these communities note how much power the 'technical guardians' (those who have the knowledge to create or edit infrastructure) have through their understanding of the technology. Of course, the fact that the infrastructure is, by definition, open for inspection in commons-based peer production contexts and FLOSS projects provides the raw material for a user to educate themselves and grow into technical expertise, and perhaps join the ranks of the 'technical guardians' themselves.

Schmidt and Bannon called for such an open approach to infrastructure in the early 1990's saying:

> "Therefore, instead of pursuing the elusive aim of devising organizational models that are not limited abstractions and thus in principle brittle when confronted with the inexhaustible multiplicity of reality, organizational models in CSCW applications should be conceived of as resources for competent and responsible workers. That is, the system should make the underlying model accessible to users and, indeed, support users in

> interpreting the procedure, evaluate its rationale and implications. It should support users in applying and adapting the model to the situation at hand. It should allow users to tamper with the way it is instantiated in the current situation, execute it or circumvent it, etc. The system should even support users in modifying the underlying model and creating new models in accordance with the changing organizational realities and needs. The system should support the documentation and communication of decisions to adapt, circumvent, execute, modify etc. the underlying model. In all this, the system should support the process of negotiating the interpretation of the underlying model, annotate the model or aspects of it, etc." [Schmidt and Bannon, 1992, 26].

The vision that Schmidt and Bannon describe in 1992 is fulfilled by WMF projects because the infrastructure for these projects is made up of free software and entirely inspectable and up for modification. It is through inspecting this infrastructure that we can uncover the conceptual structures of information organization. From this inspection, we determine the expressivity of these structures (Are the structures limited to expressing statements that are valid within the framework of first-order logic? How do the structures handle temporality?). This information then allows us to determine if the infrastructure needs to be revised or extended to represent additional types of knowledge.

2.2 Knowledge Organization and Representation

Knowledge organization is a discipline with roots in library science. The origins of the field are found in human attempts to provide access into the bibliographic universe [Wilson, 1968]. People working in the domain of knowledge organization have developed conceptual structures of information organization often referred to as knowledge organization systems (KOS).

Figure 2.1 demonstrates how discrete KOS can be made interoperable by ontologies, and how ontologies are described through logical markup to then be in-

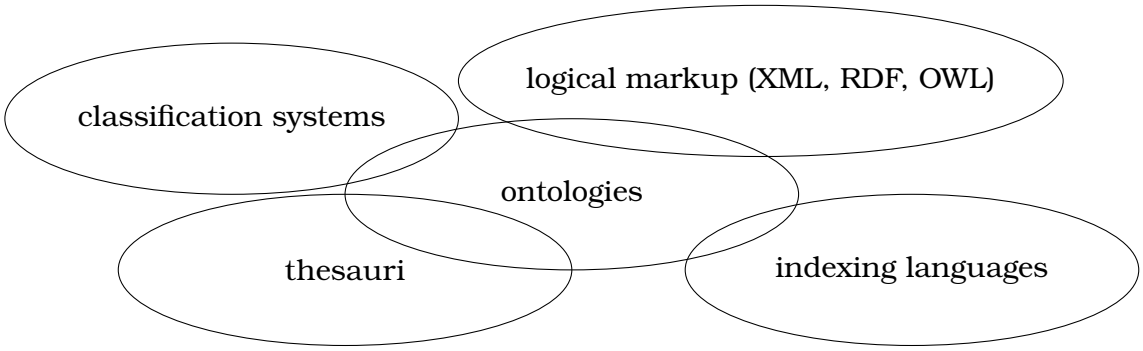

Figure 2.1: Knowledge Representation Resources (inspired by [Wright, 2007]

stantiated in technical infrastructure.

2.2.1 Ontology

Ontologies are difficult to discuss because it is possible to talk about an ontology and be referring to a conceptual structure of information organization, and it is also possible to talk about an ontology and be referring to how that ontology is instantiated in technical infrastructure.

The history of the word 'ontology' is one reason for some of the confusion about the meaning. Consulting the Oxford English Dictionary (OED), we are presented with two definitions of the word 'ontology', the philosophical and the logical. While the philosophical definition marks ontology as, "the science or study of being," the logical definition restricts ontology to, "a system similar in scope to modern predicate logic, which attempts to interpret quantifiers without assuming that anything exists beyond written expressions" [OED, 2004]. I focus on definitions of ontology used by computer scientists, information scientists, linked data professionals and technologists over the last two decades. Still inchoate and indeterminate, the definitions of ontologies we describe here have not yet been codified in the OED.

For the more recent use of 'ontology' there are conflicting opinions and definitions. This is due, in part, to the fact that many different communities have developed working definitions of 'ontology' for a range of purposes [Luczak-Rösch

et al., 2014]. Several excellent summaries of these different opinions and definition have been undertaken [Noy and McGuinness, 2001, Wright, 2014]. A commonly accepted definition is that of Noy and McGuinness:

> "An ontology is a formal explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions))" [Noy and McGuinness, 2001, 3].

Put simply, an ontology is a structure that defines and makes explicit entities (named and operationalized) and the relationships (named and operationalized) between entities. In computational systems, ontologies must be usable by both human users and machines. Ribes and Bowker discuss machine readability saying:

> "Building ontologies involves gathering domain knowledge, formalizing this knowledge into a machine computable format, and encoding it into machine language" [Ribes and Bowker, 2009, 216].

Ontologies in this sense play a key role in establishing common terminology among their users, and thus promise to facilitate the alignment of greater quantities of data [Mitraka et al., 2015], which can then be analyzed across physical space, across academic disciplines, and across methodologies.

> "Ontologies are an information technology for representing specialized knowledge in order to facilitate communication across disciplines, share data or enable collaboration. In a nutshell, they describe the sets of entities that make up the world-in-a-computer, and circumscribe the sets of relationships they can have with each other" [Ribes and Bowker, 2009, 199].

Luczak-Rosch et al. note that ontologies are one of the layers of the semantic-technologies stack [Luczak-Rösch et al., 2014]. Ontologies can be viewed as a boundary infrastructure [Star, 2010] that allows for the interoperability of multiple conceptual structures of organization. Multiple conceptual structures of organization can be implemented in a single ontology, allowing for these conceptual structures to be combined.

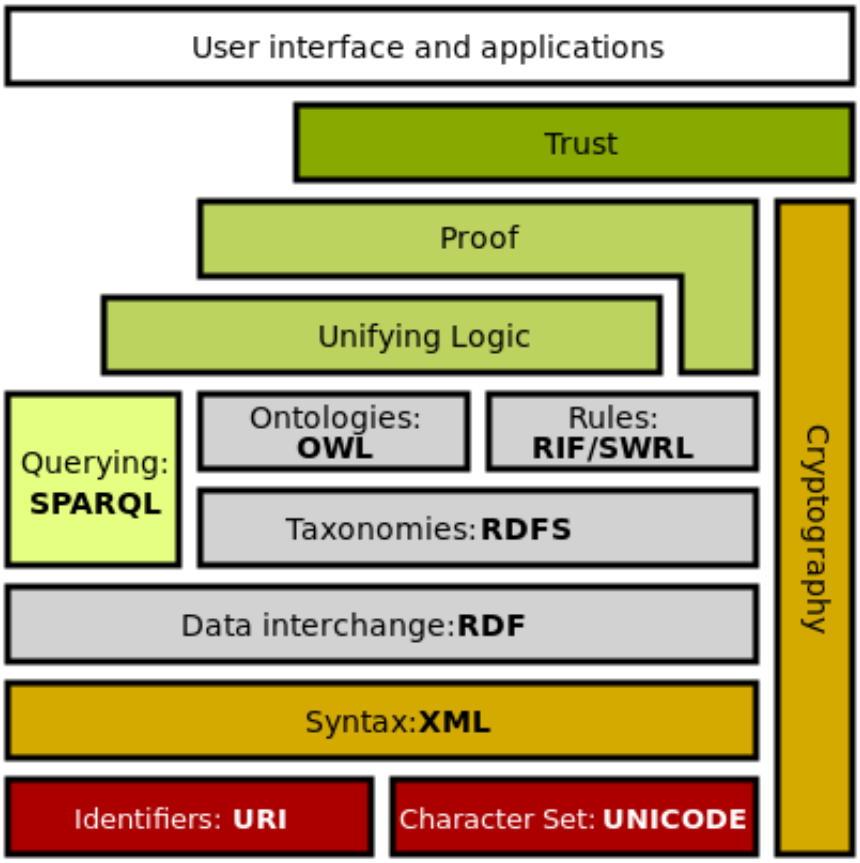

Figure 2.2: Semantic Web Stack [Bratt, 2007]

An ontology can be designed to indicate how many different conceptual structures relate to one another.

2.2.2 KO Approaches to Understanding Ontologies

Ontologies are created to describe a domain, representing entities in the domain and the relationships between entities. Ontologies can also describe other ontologies, allowing for description and meta-description. This means that we can take an ontology, for example the Library of Congress Name Authority File (LCNAF), and describe it within the ontology of the Wikidata data model. This allows the expressive power of the LCNAF to be harnessed by Wikidata.

The way that these conceptual systems are formalized when they are implemented as technical infrastructure is a key complicating factor of understanding ontologies [Luczak-Rösch et al., 2014]. Conceptual mismatch and ambiguity have been identified as challenges to system performance when multiple systems interoperate [Beek et al., 2014]. Features of the various systems of conceptual infrastructure of information organization can and do fade in the context of their implementation in ontologies as they become technical infrastructure. This is because human judgement is involved at all decision points of how to describe a conceptual system in a technical system. The wiki architecture of WMF projects allows us the opportunity to inspect the changes in the system over time and see when and where decisions about how a conceptual system is encoded have been made.

In Figure 2.2 we see one visualization of the semantic web stack. This diagram attempts to illustrate how different technologies of the semantic web are built upon one another. The rectangles labeled: 'Ontologies', 'Rules', 'Taxonomies', 'Data interchange', and 'Syntax' are all layers where conceptual systems are encoded.

Mai points out that we are at a juncture when the design criteria of knowledge organization systems need to be considered for their applicability to information organization concerns in the digital environment [Mai, 2004, 94]. How will we analyze the conceptual infrastructure of our digital information systems? When a conceptual system is formalized it can then be instantiated as infrastructure and the structure used to express it will encode the conceptual relationships that we

need to be able to document so that they are inspectable.

2.3 Semantic Web

The vision of the semantic web is to formalize a layer of semantic metadata to be used across information systems and resources on the web. Tim Berners-Lee describes it saying:

> "The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [Berners-Lee et al., 2001].

The semantic web enables people and computers to work in cooperation, but people must design and implement the technologies that allow computers to interact with the semantic web in an effective way. Architects of the semantic web have created a set of design principles that they hope will allow others to contribute to the fulfillment of the semantic web vision. Groth et al. note that the design principles of semantic web technologies instruct that we:

> "...make structured and semistructured data available in standardized formats on the web; make not just the datasets, but also the individual data-elements and their relations accessible on the web; describe the intended semantics of such data in a formalism, so that this intended semantics can be processed by machines" [Groth et al., 2012, 2].

2.3.1 Semantic Web Technologies

There are several core technologies that make up the semantic web. The most relevant are (1) Resource Description Framework (RDF), (2) Web Ontology Language (OWL), and (3) SPARQL Protocol and RDF Query Language (SPARQL). I will introduce each of these technologies and discuss the ways these technologies are integrated into WMF projects.

The Resource Description Framework (RDF) is a standard that articulates knowledge representation techniques to describe resources on the web [Lehmann et al., 2012, 159]. RDF is a standard describing a data model [W3C, 2015]. A key component of RDF is the uniform resource identifier (URI). URIs are a specific type of hyperlink that points to web resources [W3C, 2006]. RDF leverages URIs to define relationships between entities, a basic unit of which is the *triple*, which is composed of two URIs as well as a named edge connecting the two URIs.

> "RDF introduces a simple language based on triples with subject/ predicate/ object entities, each of which is given meaning by the use of a unique identifier. Entities can be grouped together in ontologies, which define the relation between different concepts. These ontologies themselves are also expressed in RDF, so they are fully machine-interpretable as well" [Van Hooland and Verborgh, 2014, 126].

RDF triples are often published on the web. RDF triple stores are databases that collect facts marked up using RDF. The Web Ontology Language (OWL) is used to mark up ontologies, and this markup then becomes part of the RDF graph for the ontology (see Figure 2.3).

> "The Web Ontology Language (OWL) extends RDF with more fine-grained concepts, allowing to precisely define ontological concepts" [Van Hooland and Verborgh, 2014, 126].

While RDF is the data model, OWL is a language to describe additional semantic relationships between entities in the ontology. OWL is expressed in RDF triples and thus, information expressed in OWL can be leveraged in certain types of SPARQL queries.

The ontological concepts that OWL can express are relations between classes, cardinality, equality, property typing, describing characteristics of properties and enumerated classes [Groth et al., 2012].

An example of some of the advantages of using OWL are:

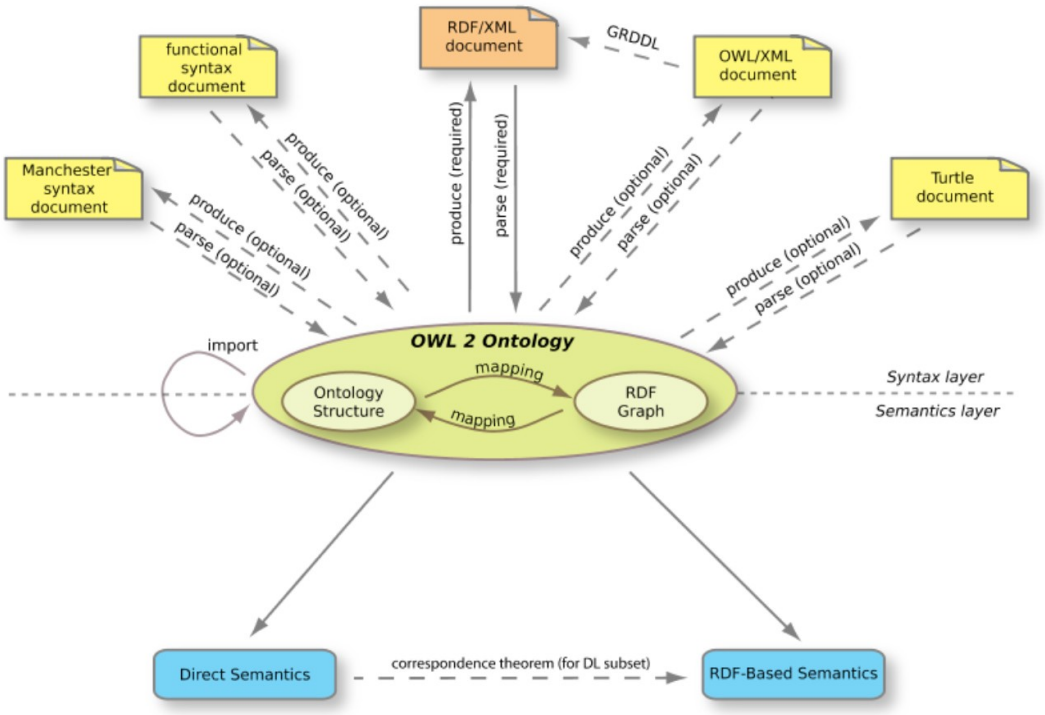

Figure 2.3: Using OWL2 to express an ontology

> "For example, when we know that the term "spouse" is symmetric (that is, that if A is the spouse of B, then B is the spouse of A), or that zip codes are a subset of postal codes, or that "sell" is the opposite of "buy," we know more about the resources that have these properties and the relationships between these resources" [DuCharme, 2013, 20].

As we see in this passage, using OWL allows us to create annotations for information that we want to share via the web. By using OWL effectively (such that it can be validated as adhering to the conventions of the language) we are formalizing the knowledge expressed in the information we sharing. This formalism allows for semantic interoperability which is a requirement for inferencing and machine computable logic as well as for federation [Gödert et al., 2014, 105]. This semantic enrichment is one of the ways that economic value can be created from structured data.

Many resources are described using RDF and OWL. To find specific information in a set of RDF triples people have created query languages. One query language that is frequently used to query RDF is SPARQL. The name SPARQL is a recursive acronym for SPARQL Protocol and RDF Query Language [DuCharme, 2013]. SPARQL queries RDF triple stores in order to identify and return sets of triples that meet the criteria specified in the query.

Mapping Wikidata properties to those described in RDF from the semantic web allows for the use of standard query languages to ask questions of the knowledge base [Hernández et al., 2015]. However, the mapping of Wikidata properties to RDF properties is not yet complete. This mapping requires human judgement, thus it is not something that can be entirely automated [Spitz et al., 2016].

2.3.2 Wikidata as Infrastructure of the Semantic Web

The Wikidata project was created with these design principles in mind. Wikidata makes structured data available in three formats: JSON, XML and RDF [Wikidata, 2015b]. The individual data elements and their relations are available via a Linked

Data interface, the MediaWikiAPI, the Wikidata Query service, and via several SPARQL end points [Wikidata, 2015a]. The semantics of the data are described in a formalism, a reification of the Wikidata data model in RDF [Erxleben et al., 2014].

Another term that is sometimes used to describe many of the same technologies is "web of data" [Hitzler et al., 2010, 15]. This term is used to foreground the importance of interactions of data exchange. This conceptualization of a web that is enriched with semantic information is the location of the community-led project of Wikidata. Wikidata is expected to become a very valuable component of infrastructure of the semantic web, there is already evidence of this, see [Mitraka et al., 2015].

Wikidata is a knowledge base [Müller-Birn et al., 2015]. One effort to estimate the importance of the knowledge base as a specific genre of infrastructural component argues:

> "The increased practical use of semantic technologies is witness to the fact that important base technologies are well-developed—their strengths and weaknesses understood much better than in the early years of the Semantic Web activity—and that they are useful for solving the problems encountered in practice. And indeed, recently we have seen major IT and venture capital companies investing in the segment, while the trend in research projects and funding drifts rather heavily from foundations to applications. New technologies continue to be developed, and it can be expected that they will lead to the solutions that will enable innovative applications with high impact in the next few years" [Hitzler et al., 2010, 15].

If this estimation proves to be correct, Wikidata will be one of the largest and most frequently queried sources of structured data on the web in the future [Mitraka et al., 2015].

Conclusion

I report on research that takes place at the nexus of social computing, knowledge organization and semantic web technologies. Across the projects of the Wikimedia Foundation (WMF) editors are working collaboratively to organize information and mark it up so that it can become structured data, an important component of the semantic web. The values of the free software movement emphasize the importance of sharing source code for software and making this code open to inspection. The structures of information organization of the WMF projects I describe in this research are created collaboratively by social groups, this collaboration is open to inspection and the work products (such as structured data) that are created by these communities play an important role in the semantic web.

Chapter 3

RESEARCH ACTIVITIES

This project consists of an investigation of structures of information organization in Wikimedia Foundation (WMF) projects. I report on three projects; (1) a qualitative content analysis of the discussions of categorization in the category system, (2) a quantitative analysis of patterns of editing among infobox template editors, and (3) a close reading of infrastructural components of three structures used for information organization in (WMF) projects. I introduce six research questions for the set of three projects. I then introduce each project and the methods I used for each project.

In order to understand some of the strategies used for information organization used by contributors to WMF projects, I posed the following motivating questions:

1. How does collaboration occur in relation to the creation of the category system?

2. What do editors see as the purpose of the category system?

3. How are infoboxes created?

4. How can we measure and distinguish different styles of infobox template editing?

5. In what ways does the knowledge base manifest as infrastructure?

6. How are KOS instantiated in infrastructure?

I describe each of the three projects separately. I first describe the qualitative content analysis of the discussions of categorization; then the quantitative analysis

of three original metrics to identify patterns in the editorship of infobox templates; and, finally, the close reading of infrastructure.

3.1 Qualitative Content Analysis

I report on a qualitative content analysis of the discussions related to categorization in English Wikipedia[1]. I describe the data I collected, and the qualitative content analysis I use to analyze the data.

I investigate the category system because it is one of the structures of information organization that is most visible as a knowledge organizing system (KOS). I investigate how the community worked together to create this novel system of labeling articles with tags within the wiki architecture. These category labels describe how the article is related to other content in Wikipedia by grouping sets of articles into categories.

The category system is a structure of information organization that allows editors to group content conceptually. In 2003, roughly two years after Wikipedia began, the community decided to create a category system to organize and tag the content of the site [Voss, 2006]. The category system has changed over time, as have conceptualizations of the role it should serve in Wikipedia. Rather than use an existing KOS, the community created a new system for organizing content. Any Wikipedia editor has the authority to apply category labels to pages, to remove category labels from pages, to create new categories, and to suggest categories to be considered for deletion. The incremental additions to, revisions of, and deletions of category labels to pages make up the category system. Discussions of this work take place on the talk pages of the wiki, such as the talk page for `Categorization` [Wikipedia, 2016b].

I investigate the category system due to the fact that it represents a controlled vocabulary of of concepts used to organize content, because it has been reused in numerous research applications [Bao et al., 2012], and it plays a important role

[1]This project concerns the discussions of categorization that took place in English Wikipedia. I will refer to English Wikipedia as 'Wikipedia' throughout my discussion of this project for brevity.

in the structure of DBpedia and Wikidata, knowledge bases that play important roles in the semantic web [Müller-Birn et al., 2015].

3.1.1 Data

I analyze the processes the Wikipedia community went through in the design, creation, and implementation of the category system of Wikipedia through their on-wiki discussions of categorization. Due to the fact that the category system of Wikipedia is one of the largest extant examples of a collaboratively-built system for the organization of digital content, I chose to document and learn from the experiences of the community who created it.

In order to understand this process, I conduct a qualitative content analysis of the talk page for categorization in Wikipedia [Wikipedia, 2016b]. The data consists of the text of the talk page for categorization from 2004-2007. Talk pages in Wikipedia are pages that are affiliated with article pages where members of the Wikipedia community can post comments in order to discuss related issues with one another.

The text on this particular talk page consists of discussions, proposals, disagreements, examples, hypothetical situations and opinions about the category system of Wikipedia. This page represents much of the deliberation that went on about the design and implementation of the category system and is evidence of the community's views of the system throughout its history. Viégas et al. demonstrated that talk pages are the part of Wikipedia where the majority of collaborative deliberation takes place [Viégas et al., 2007].

3.1.2 Content Analysis

I chose to perform a content analysis of the discussions on the talk page for `Categorization` because I am interested in the processes through which the category system is collaboratively created and the opinions and actions of the community in relation to the category system.

I follow the work of McMillan [McMillan, 2000] and Krippendorf [Krippendorff, 2012] to inform my approach to content analysis of this talk page in Wikipedia. McMillan expands upon Krippendorf's [Krippendorff, 2012] approach to content analysis to make it more suitable for web content analysis [McMillan, 2000]. McMillan outlines the ways that each of five steps of content analysis (formulation of question, sampling, determining unit of analysis, coding, analysis) differ when applied to web-based content.

In my project, I have addressed McMillan's five steps as follows: for the formulation of my research questions, I situated the problem in the context of the literature of the creation of systems of information organization. By analyzing a collaboratively-created system of information organization for digital content, I am looking at a novel system and comparing it to what we know about other systems of information of organization.

For the step of sampling, I chose to take a purposive sample in the form of the archive of a single talk page. I chose to scope my data collection to the text of the talk page for categorization because I decided that that page was the most centralized location for discussion on the topic. Although the topic was also discussed via internet-relay chat (IRC) and on listservs and in-person, talk page participants made efforts to summarize external conversations of relevance on the talk page. While the content of this talk page does not represent the full text of all discussion about this topic in the Wikipedia community, it is the most representative source to which we have open access.

For the step of coding, I engage in open coding as defined by the Grounded Theory approach of [Glaser and Strauss, 1967]. For the step of analysis, I conduct a thematic analysis of the codes which emerge through the open coding process. In her discussion of the advantages of open coding in computer-mediated discourse analysis, Herring states:

> "This approach is especially well suited to analyzing new and as yet relatively undescribed forms of CMC, in that it allows the researcher to

> remain open to the possibility of discovering novel phenomena, rather than making the assumption in advance that certain categories of phenomena will be found" [Herring et al., 2004, 354].

I collected the text of the Category talk page covering the six month period from June 2004 through January 2005, which I refer to as period 1. That marked the start of the community effort to implement a category system in Wikipedia. At this point the design of the category system was completely open; changes were still being discussed and codified. I coded the data thematically in order to begin a content analysis of each section of the archived talk page. Through open coding I developed a codebook of thirty one codes.

After coding the data and arranging codes into families, several themes emerged from the data. The editors who were participating in the discussions of how to design and implement the category system frequently returned to issues of hierarchy, scope, navigation and collaboration. By contrasting the editors' discussions of hierarchy with theory about constructing hierarchy in indexing languages, it is clear that many of the same conclusions reached through trial and error in Wikipedia have been described in the KO literature. The editors' discussions of how to scope the number of categories and the type of categories in the system are especially interesting when considered in the context of the recurring appeals to the community to consider extant schema for information organization as potential models. Discussions of navigation through Wikipedia via the category system reflect a very clear understanding of the category system as a navigational tool, something that no longer seems to be the case for many members of the Wikipedia community.

I then collected a second set of data, which I refer to as period 2, (spanning January 2005- January 2007) from the same talk page and coded that data using the codebook. The primary difference between this data collection and coding process and the first round is that the number of codes relating to the purpose of the category system increased. Enough new codes relating to purpose were needed that I decided to introduce a new theme, that of 'purpose'. The goal for collecting

data from the second period was to learn how the community understood these themes as the category system was expanded and changed. I was also investigating whether new themes would emerge and how the focus of the community would shift over time.

3.1.3 Connection to Research Questions

The analysis of this data will address my first and second research question. RQ 1: How does collaboration occur in relation to the creation of the category system? RQ2: What do editors see as the purpose of the category system? The category system is one of the structures of information organization in Wikipedia. This provides me the opportunity to see how the community discussed, designed, and implemented this structure of information organization. Through analysis of these discussions I learn about the category system as a structure. Editors discussed tools, ideas, previous organizational strategies, processes for category application, and other topics throughout this discussion.

3.2 Quantitative Metrics for Understanding Template Specialization

Working with colleagues [2], I collected a set of data related to the editing patterns of the set of all templates used to create infoboxes and then provided descriptive statistics related to this data set. I developed three original metrics to represent dimensions of editing activity: influence, longevity and diversity. I discuss these metrics in detail in Chapter 5. Infoboxes are structures of information organization that provide quick facts about a topic. Infoboxes are presented graphically on the right-hand side of articles (see Figure 3.1). This infobox has an image of a nautilus and information related to the classification of the organism. The Wikimedia community describes the purpose of information boxes to be tools to support readers [Wikipedia, 2015].

As the community of editors of Wikipedia embraced the infobox structure, they

[2]This project is the result of collaborative research with Martez Mott and David W. McDonald.

Figure 3.1: Nautilus Infobox

created numerous infoboxes on article pages. Infoboxes are created on pages by applying code that describes a template into the wikitext of an article. Due to the structured nature of the markup, each information box also represents a chunk of structured data presented in a machine readable format. As the number of information boxes grew, these chunks of machine readable structured data accumulated, and people recognized that alongside the wealth of unstructured information and data in Wikipedia, there exists a growing store of structured data.

3.2.1 Data

There are more than 11,000 infobox templates transcluded into more than 2 million pages in English Wikipedia. I collected the edit histories of each template for an infobox from the publicly-available data dumps provided by Wikipedia [Wikimedia, 2016]. I then counted the number of edits to these templates per editor. I used the Wikipedia webservice API to request all pages in the template namespace that begin with the string: "Template:Infobox". This request returns a broader set of pages associated with template including the template page itself, sandbox pages, and documentation pages. I cleaned the resulting list of pages to exclude (as best possible) any pages that were not explicitly the templates themselves. That is, I

removed documentation and sandbox pages from our list. This resulted in a list of 11,561 template pages, which includes some "duplicates" that result from name changes and template page moves. I then used the resulting list of infobox template pages to collect revision data, including the editors, edit timestamp, along with the text of each revision for the complete list of template pages. I then queried the API to find out how frequently each template is transcluded into pages in the article namespace. That is, I ignored transclusions of infobox templates into user sandboxes and other places where users may simply be experimenting with an infobox template. In sum, the dataset includes 234,989 unique revisions, 18,956 unique editors, and 11,508 unique infobox templates.

The characteristics of infobox template editors vary widely. I divided editors into 5 separate cohorts based on edit count to begin to unpack the different types of template editing activity. I decided not to include editors with less than ten revisions, as I was interested in editors who have demonstrated commitment to template editing work. The cohorts were constructed so that approximately 25 percent of the editors were placed in each cohort, with the exception of cohort 5, the outliers. Across my analysis of these cohorts I use three different metrics to help us understand template work: longevity, diversity, and influence. I explain the details of these metrics in Chapter 5.

3.2.2 Connection to Research Questions

This research project allows me to answer my third and fourth research questions. RQ 3. How are infoboxes created? RQ 4. How can we measure and distinguish different styles of infobox template editing?

This research also allows me to gain a technical understanding of how templating, transclusion, and wikitext are used with relation to how information boxes are rendered on article pages. The technical understanding of templating that I gained informed the design of the following study and also helped me learn about the infrastructure of WMF projects more generally.

3.3 Close Readings of Infrastructure

In order to understand how to read infrastructure, we must first operationalize the definition of infrastructure. I follow Bowker et al. who define infrastructure as "pervasive enabling resources in network form" [Bowker et al., 2010, 98]. The infrastructure I analyze for this third project consists of the technologies, information systems, standards, models, modules, scripts and programs that enable WMF projects.

After determining how I define infrastructure for this research, I selected a method that would allow me to analyze the infrastructure that is used to make three structures used to organize information in WMF projects. I call this method *close reading of infrastructure* and I borrowed elements of from ethnographic traditions [Star and Strauss, 1999, Star and Ruhleder, 1996, Geiger and Ribes, 2011, Tsing, 2015, Geertz, 1994].

The method of close reading of the infrastructure relies on the techniques presented in the ethnography of infrastructure [Star and Ruhleder, 1994, Star and Ruhleder, 1996, Star, 1999]. This method allows for analysis of these systems for the purpose of providing techniques to enable visibility of what is expressible in a given structure of information of organization and what remains inexpressible.

3.3.1 Connection to Research Questions

Close readings of infrastructure allow me to answer my fith and sixth research questions: RQ 5: In what ways does the knowledge base manifest as infrastructure? RQ6: How are KOS instantiated in infrastructure?

I addressed my research questions through identifying and analyzing relevant examples of the dimensions of infrastructure as observed in three settings. I chose these settings because of the important role that the category system and the set of all infoboxes played in the history of how the conditions in which an open, collaboratively-built commons-based peer production system centered around the creation and maintenance of a platform for semantic linked data came to exist.

3.3.2 Observing Infrastructure

Star outlines her recommendations for how to undertake an ethnography of infrastructure because of the need she saw for "...new methods to understand this imbrication of infrastructure and human organization" [Star, 1999, 379]. I based my analysis of the infrastructure of Wikimedia projects in Star's recommendations. I used these techniques to perform a close reading of the infrastructural components of these systems in order to discover how the infrastructure makes the expression of certain relationships and claims possible, while making the expression of other types of claims and relationships impossible.

Star argues that an ethnographic sensibility allows one to pay attention to: "an idea that people make meanings based on their circumstances, and that these meanings would be inscribed into their judgments about the built information environment" [Star, 1999, 383]. It is precisely a subset of these judgements, those that pertain to how information is organized, and the structures that are used to accomplish this work, and the ways in which these structures are instantiated into the infrastructure of WMF projects that drew my attention to the suitability of these peer production systems as the site of research.

Star and collaborators [Star and Ruhleder, 1996, Star, 1999, Star and Bowker, 2006] advanced and refined a theory of how to analyze infrastructure. They state that the following dimensions can be found infrastructure:

- Embeddedness- Infrastructure is "sunk" into other structures, social arrangements and technologies.

- Transparency- Infrastructure does not need to be reinvented each time or assembled for each task.

- Reach or scope- Infrastructure has reach beyond a single event or one-site practice.

- Learned as part of membership- strangers and outsiders encounter infras-

tructure as a target object to be learned about. New participants acquire a naturalized familiarity with its objects as they become members.

- Links with conventions of practice- Infrastructure both shapes and is shaped by the conventions of a community of practice.

- Embodiment of standards- Modified by scope and often by conflicting conventions, infrastructure takes on transparency by pugging into other infrastructures and tools in a standardized fashion.

- Built on an installed base- Infrastructure inherits the strengths and weaknesses, affordances and limitations of the technologies in which it is implemented.

- Becomes visible upon breakdown- The normally invisible quality of working infrastructure becomes visible when it breaks; the server is down, the bridge washes out, there is a power blackout. Even when there are back-up mechanisms or procedures, their existence further highlights the now-visible infrastructure.

- Is fixed in modular increments-Because infrastructure is big, layered, and complex, and because it means different things locally, it is never changed from above. [Star and Ruhleder, 1996, Star and Bowker, 2006, Star, 2010]

I will describe how I used each of these dimensions to investigate the infrastructure of WMF projects.

3.3.3 Reading Infrastructure

I used these dimensions of infrastructure as sensitizing concepts [Glaser, 1978]. They gave me a sense of what to look for when attempting to observe infrastructure. In order to gain familiarity with Wikimedia projects, I became a registered editor of English Wikipedia in 2011. Through the process of becoming and editor I was

able to enage in participant observation [Musante and DeWalt, 2010] of the infrastructure of these projects. In order to learn about the Wikidata system, I became a Wikidata editor in October, 2013, shortly after the Wikidata project launched. I read the documentation of the data model for Wikidata. I familiarized myself with the documentation for the technologies that constitute the installed base, as well as the documentation for the standards that inform the design of these technologies. I participated as an editor making edits to Wikidata several times a week for two years and on a daily basis for a third year, I read mailing list postings on the Wikidata mailing list, I monitored Wikidata: ProjectChat [Wikidata, 2016c], I attended Wikidata Office Hours [Wikidata, 2016b], and gained hands-on experience of using a range of automated and semi-automated tools for editing Wikidata. I took notes about interactions I observed that related to information organization. I took notes about where I found descriptions of structures that helped me clarify my understanding.

I wrote descriptive field notes about each example employing the approach of thick description [Geertz, 1994]. I used these notes to describe how these structures are implemented from a technical perspective. I explored documentation, or noticed the lack of documentation used to communicate these structures. I explored the text of talk pages, logs of revisions and edits to pages in the wiki using the approach of trace ethnography [Geiger and Ribes, 2011], a method of learning from computational logs, to help me understand the system through looking at records of interactions preserved in the software used to create the wikis.

My choice to become and editor of Wikipedia and then of Wikidata, and my use of participant observation [Musante and DeWalt, 2010] inform my understanding of infrastructure of WMF projects. As discussed in Chapter One, infrastructure is often hidden from view, making it difficult to observe. The infrastructure of WMF projects has been extended over a period of fifteen years and is the work of tens of thousands of contributors. An extended period of observation is necessary for anyone who seeks to understand where infrastructure is and how it plays a role in the WMF projects. Through my enagement as a participant observer I learned how

Table 3.1: Sites of Close Reading of Infrastructure

Structure	Perspective	Technical Infrastructure
category system	infrastructure used to create it	wikitext
		category template
		hotcat
	as infrastructure itself	dbpedia
		hidden categories
information boxes	infrastructure used to create them	wikitext
		template:infobox
	as infrastructure themselves	DBpedia
		Wikidata
Wikidata	infrastructure used to create it	MediaWiki
		WikiBase
	as infrastructure itself	RDF serialization
		SPARQL endpoint

to use tools to manipulate infrastructure, watched in awe as other editors wrote bots to perform instructions relating to manipulating infrastructure and observed moments of infrastructural breakdown. Just as Star and Ruhleder argue that infrastructure is "learned as part of membership" [Star and Ruhleder, 1996], it was only through becoming a member of these WMF project communities that I could learn enough about infrastructure to report my findings.

Based on my observations as a community member of English Wikipedia and Wikidata, I selected six sites (see Table 6.1 in which to perform close readings of infrastructure. I will address my research questions through the analysis of the dimensions of infrastructure.

I considered each of my three sites: (1) the category system, (2) the set of infobox

templates, and (3) the knowledge base of structured data, from two infrastructural perspectives: (1) the infrastructure used to create them, and (2) how they in turn serve as infrastructure for other applications, resulting in six sites for this close reading of infrastructure.

I identified examples of Star's dimensions of infrastructure from each of these sites by returning to the notebooks containing field notes I created while learning about WMF projects and conducting the research described in Chapter 4 (qualitative study of the category system) and Chapter 5 (quantitative study of the templates for infoboxes).

Using my field notes as the source for examples, I identified potential mechanisms that could be used to decide what types of knowledge can be represented and what types of knowledge will not be represented by looking at how the conceptual systems are expressed in the technical infrastructure. I wrote memos [Glaser and Strauss, 1967] about each of these dimensions of infrastructure, paying particular attention to how knowledge organization systems (KOS) are expressed in the technical infrastructure.

Conclusion

The research activities I conducted take place across three projects, a qualitative content analysis of the category system, a quantitative analysis of patterns of editing in the templates for infoboxes, and a close reading of infrastructure. The close reading of infrastructure takes places across three sites, including the sites of the first two projects. By revisiting the category system and the infobox templates from the perspective of infrastructure, I was able to deepen my understanding of how each of these projects informs the subsequent work. This enriched my overall understanding of how content and infrastructure are related. The knowledge base of structured data is more complicated than either the category system or the infobox templates, without the process of understanding those two structures in relation to infrastructure first, I may not have been able to undertake the close reading of the knowledge base.

These three projects represent three different ways to read infrastructure. In

the content analysis of the category system I examined discussions of infrastructure; in the quantitative analysis of patterns of editing I examined how editors take part in the creation of a type of infrastructure; and in the close reading of infrastructure I examined how KOS are encoded in infrastructure. The structure of the close reading project also allowed me to bring all three projects together by examining the first two projects again from an infrastructural perspective. Due to the fact that WMF projects are the products of vibrant communities continuing to thrive today, this also allows me to examine the ways infrastructures are expanded, reused, revised and extended. The knowledge base of structured data, Wikidata, could not have been built in 2004 when the category system was implemented. Nor could it have been built at the same time that infobox templates began to be transcluded across pages of Wikipedia. It was through the lessons learned from building the infrastructure for the category system and for the templating system that the community gained the ability to recognize the need for a knowledge base of structured data and began to envision its design.

Chapter 4

COLLABORATIVELY CREATING A CATEGORY SYSTEM

I selected this case in order to understand the structure of the category system in the context of English Wikipedia. The category system shares some characteristics of structures of information organization well known in classification theory, however there are also many areas of divergence. The category system is one of the most frequently re-used structures for information organization in Wikipedia. In this section I will describe the characteristics of the structure of the category system and discuss how this relates to other structures commonly used in information science. I will also discuss the power of the relationships created in the category system and how they are leveraged in other applications.

I pose two research questions for this project:

1. RQ 1: How does collaboration occur in relation to the creation of the category system?

2. RQ 2: What do editors see as the purpose of the category system?

I introduce the notion of tagging and how tagging is used to create this structure of information organization. I then introduce the setting for this research: English Wikipedia. As described in Chapter Three, I use the method of qualitative content analysis in this project. I qualitatively coded the text of the talk page for `Categorization` [Wikipedia, 2016b].

4.1 Browsing Wikipedia

Trent was recently browsing Wikipedia and stumbled on a new page for the indie folk band Aquabats. Trent decided he would help by categorizing this page. He had

never tagged in Wikipedia, but he had used del.icio.us before—how hard could this be? He looked for the category `Bands` *and found it was redirected to another category. He found* `Musical groups by genre` *but couldn't find* `Indie` *or* `Independent`*. He found* `Folk` *but couldn't categorize the Aquabats as* `Folk rock` *nor* `Folk punk`*. He considered trying to add a new category, but a category with only one item wouldn't be all that useful. Thinking to himself that this was going to be harder than he initially thought, he gave up on the idea of helping.*

Social tagging became wildly popular with the advent of del.icio.us [Wetzker et al., 2008]. Suddenly, the idea of visible and shared categories became popular with developers and researchers. Researchers could study how shared category metadata was applied, how it evolved, and how individual use differed from collective use. Yet, despite the large number of studies of tags and tagging behaviors, none have considered the collective rationale behind the tagging scheme because largely the rationale had to be divined from the tags themselves. This project specifically considers the collective effort and negotiation that generated one tagging scheme currently in use and analyzes those negotiations. Schemes for categorization or tagging have a direct impact on how people access information and how they can collaborate. These pieces of metadata can be used to help manage processes as tags are added and removed. Further, they form a type of communication system, a set of signals that collaborators can use. A better understanding of the rationale behind a tagging scheme will inform the future development of tools to improve the quality of the scheme, how the scheme is applied, and the way it provides access.

Wikipedia is an online encyclopedia created entirely of user-generated content. Roughly two years after Wikipedia began, the community decided to create a category system to organize and tag the content of the site. The category system has changed over time, as have conceptualizations of what role it should serve in Wikipedia. This study analyzes six months of discussion about the design and implementation of the category system in Wikipedia. The analysis reveals a set of important themes that related to category systems and tagging broadly.

In the following, I outline the prior work that considers tagging practices and relate that to some of the traditional work in information science. I describe my data collection, my approach to analysis and describe four key themes that emerge from the data. I then present a set of analytic frames for thinking about how category systems enable styles of collaboration and illustrate how individuals contributing to the discussion assumed these roles at different times.

4.2 Tagging and Wikipedia

In Wikipedia, categories are applied to pages as an internal hyperlink to a 'category' page. The WikiMedia system reads those special internal links and groups together all pages that include the link. Any number of category links can be applied to a page. The category system contains traditional content categories as well as management and historical categories that are often hidden when the page is requested. The simple application of a link to a page is similar to the metadata practice of adding a tag to a piece of content.

In an early analysis of tagging behavior, Golder and Huberman [Golder and Huberman, 2006] argue that sensemaking is one purpose of tagging and that individual conflicts at the level of tag specificity makes collaborative tagging systems fuzzy and less precise. Subsequent studies by [Sen et al., 2006], [Farooq et al., 2007], and [Millen et al., 2005] served to illustrate the tension between individual tag choice and group tag use, quantitative techniques for considering whether a tag had information value, and the different motivations for tagging within an organization as compared to public tagging. This project extends what we know about tagging systems by characterizing the decisions about how a tag-based category system should operate, and how best to address these known issues through the design of the system.

The category system in Wikipedia has been considered from a number of different perspectives. Some studies follow a quantitative approach. For example, Muchnik et al. tested five algorithms for automatic extraction of hierarchical re-

lationships among Wikipedia pages against the extant category system [Muchnik et al., 2007]. Kittur et al. quantitatively mapped the categories used in Wikipedia into seven topical areas [Kittur et al., 2009]. This work collapsed the category system using a shortest-path-to-root approach for determining where node content would be assigned. This obscures the significant work Wikipedians have undertaken to systematize category hierarchy and encourage editors to apply category labels according to community convention.

Holloway et al. conducted a quantitative study of the category system from 2004 through 2007 to produce a thorough description of the category system as it stood in 2007 [Holloway et al., 2007]. As these studies show, quantitative approaches can illustrate the depth and breadth of the category system and compare the relative prevalence of topical content. But the quantitative approach cannot illustrate the collaborative rationale behind the system, and some quantitative approaches may specifically obscure the community efforts to apply category tags in a systematic manner. Still, other studies have considered participation in Wikipedia from a qualitative perspective. Forte and Bruckman [Forte and Bruckman, 2008] outline social roles that exist in the Wikipedia community. They describe the negotiation that takes place in the wiki, and on related mailing lists between individuals who are trying to create community-wide policies. This highlights the negotiated aspects of much of the work in Wikipedia, including that around the category system.

In a study of the types of work valued in the Wikipedia community, Kriplean et al. point out that both category creation and category link application are types of work acknowledged and valued by the community [Kriplean et al., 2008]. This work supports their argument by providing rationales for elaborating content organization advanced by the community in the early stages of the development of the category system. Analyzing discussions related to the creation and implementation of the category system, we describe strategies that the Wikipedia community employed in the attempt to make the implementation of category designations consistent throughout the site.

In the field of information science, knowledge organization (KO) theorizes, analyzes and critiques systems designed to organize information. From a KO perspective, Voss described the Wikipedia category system as a thesaurus built through collaborative tagging [Voss, 2006]. While Voss notes that simply adding categories to other categories facilitates the creation of hierarchy, he did not outline how different implied relationships are created, and sometimes confused, between categories when they are nested. My analysis extends this finding by characterizing types of relationships present among super- and sub-categories in Wikipedia. There is surprisingly little evidence of the community prioritizing designs that would help users distinguish between different types of relationships in the early discussions of the category system in Wikipedia.

4.3 Category System Design

While working together to decide how a category system should function in Wikipedia, interested contributors shared their thoughts and debated the merit of many proposals. Several themes emerged from their discussion: hierarchy, scope, and navigation and collaboration. In terms of hierarchy, many editors felt that the best way to structure the category system would be through category tags structured into (what they called) hierarchical trees, and the Wikitext markup was designed to prompt users to always place a newly-created category into the supercategory of their choice. Scoping concerns are reflected in the way editors debated the appropriate number of category tags to apply to a given page in Wikipedia. Another theme addressed the question of whether navigation was considered as one of the purposes of the category system, and how best to address user needs related to navigation. Editors were also generally concerned with how the category system would facilitate forms of collaboration. I discuss each of these themes in detail.

4.3.1 Hierarchy as a Term of Art in Knowledge Organization

Hierarchy is a fundamental component of classification. Kwaśnik communicates the breadth of what classification can describe, saying:

> "Classification is a way of seeing. Phenomena of interest are represented in a context of relationships that, at their best, function as theories by providing description, explanation, prediction, heuristics, and the generation of new questions. Classifications can be complex or simple, loaded with information or rather stingy in what they can reveal. They can reflect knowledge elegantly and parsimoniously, or they can obfuscate and hinder understanding. Some classifications enable flexible manipulation of knowledge for the purposes of discovery; some are rigid and brittle, barely able to stand up under the weight of new knowledge" [Kwasnik, 1992, 46].

This passage provides a sense of the heterogeneity of types classifications. In order to provide background for the discussions of hierarchy that the Wikipedia editors who participated in the discussions on the talk page for Categorization had, I will introduce how hierarchy is theorized in the scholarly literature related to classification.

Classification theorist, Brian Vickery describes hierarchy saying: "Any given hierarchy consists of a series of terms of increasing intension, derived from the summum genus by the application of a succession of characteristics" [Vickery, 1958, 26]. This means that each sub-class is differentiated from a class through sets of qualifying criteria. Svenonius clearly articulates the several varieties of hierarchies saying:

> "Hierarchical relationships take a variety of forms. In the previous section the two most important were introduced: genus-species and perspective hierarchies. The genus-species relationship, also known as the inclusion relationship, is the classic hierarchical relationship with the

> properties of reflexivity, transitivity, and anti-symmetry. It has another property as well, which in the computer literature is called inheritance and in the classification literature hierarchical force, whereby what is true of a given class (Furniture) is true of all classes it subsumes (Chairs, Tables, and so on)" [Svenonius, 2000, 163].

Hope Olson summarizes one of the drawbacks of using hierarchy in the bibliographic tradition geared to producing linear organization.

> "Like the two-dimensional projections of the three-dimensional earth classifications, these must distort all knowledge in its infinite multidimensionality into a linear arrangement suitable for creating a browsable list or locations on shelves" [Olson, 1998, 240].

Olson reminds us that if any of the interpretive decisions that take place when items are classified had been made differently it is possible that alternate orderings of the items could have been produced. The complexity of this issue is magnified when we consider hierarchical structures designed to create a single linear order as a model for the organization of digital content. One linear order is not only not necessary, it is not desirable. Rather than having single, fixed path through the organizational structure to the location of the content, it becomes advantageous to create and articulate multiple paths so that content is discoverable from an increased number of original queries. Kwaśnik lists a several drawbacks of tree structures built on the principle of inclusion. She summarizes her critique by pointing out the trade-offs of establishing the single order these structures enable us to produce saying: "As with hierarchies, by emphasizing a certain relationship, a tree can mask, or fail to reveal, other equally interesting relationships" [Kwasnik, 2000, 34].

Following Svenonius, the genus-species relationship is one of the two most important types of hierarchies [Svenonius, 2000, 163]. In the genus-species hierarchical relationship there is the assumption of inheritance. The biological origin

of the concept of inheritance might be one reason that genetic metaphors are so commonly used in classification theory literature. Fischer notes that, standards as well as the academic literature:

> "more or less implicitly allow that these different types of hierarchy relations may be conflated into one hierarchical relationship in an actual thesaurus" [Fischer, 1998, 20].

While this is less and less true of the academic literature today, it is still the case in many systems.

The definitions of hierarchy are terms of art in KO. The editors who were contributing to discussions of categorization in English Wikipedia held many competing understandings of hierarchy, many of which would not qualify using these scholarly definitions. As Fischer noted, even scholars and the authors of standards are not always consistent in their enforcement of how heirarchy is operationalized.

4.3.2 Understandings of Hierarchy in the Wikipedia Category System

The community of editors who participated in the discussions was very concerned with ensuring that hierarchy would be a feature of the category system. The fact that the community felt hierarchy to be such an important element for organization is consistent with how the KO literature describes the advantages of hierarchical structures. Svenonius noted that hierarchical relationships provide excellent advantages for supporting user navigation and the ability to instrument the relevance and size of a set of results for a user query [Svenonius, 2000]. The fact that these advantages were apparent to the editors who designed the category system in Wikipedia is evident, even if diverse understandings of what hierarchy means proliferated among editors. When looking at the categories displayed at the bottom of any page it is possible to click on any of them (they are all hyperlinks) and see related categories and all subcategories. The editors who contributed to the design of the category system shared a vision of this feature that it would allow

users to more quickly understand the context of any individual article by making relationships between articles visible. Hierarchical structures allow users to see how different concepts relate to one another in a given system. Aitchison et al. [Aitchison et al., 2000] defines four types of hierarchical relationships: generic, whole-part, instance, and polyhierarchical [Aitchison et al., 2000].

Generic - A generic hierarchical relationship is defined as a conceptual transitive closure. There are very few examples of this in the category system. But there are other indexing tools in Wikipedia that are organized in this way, for example, the Wikipedia page for 'List of birds'. That is, if we take the class to be 'birds', all of the pages for which links are supplied in the list are pages for birds.

Whole-part - This relationship consists of a single concept or entity as the class with parts of that concept or entity as the subclass. An example of this type of hierarchical relationship in Wikipedia would be the pages listed under the category 'States of the United States'. Other than the page 'U.S. state', the other pages under this category are all parts of the category itself.

Instance - These are general concepts or classes which have specific instantiations as a subclass. It is difficult to find examples of categories for which the subcategories are all instances of the category. This is more often achieved through lists in Wikipedia. An example of this type of hierarchical relationship in Wikipedia would be the 'List of cathedrals' page. If we take 'cathedrals' to be the class, all of the cathedrals listed on that page are instances of the class.

Polyhierarchial - This type of relationship describes cases when one term is located underneath more than one category. Many categories in Wikipedia are located in more than one parent category. Polyhierarchy is very common in the category system of Wikipedia.

Much of the discussion in the data set illustrated a failure to discriminate between these different types of hierarchy, and multiple, sometimes contrasting, assumptions of what type of hierarchy was being proposed. All four types of hierarchy are in use in the category system of Wikipedia. In an ideal hierarchical structure a consistent relation would link terms. This recommendation is not followed in

the category system of Wikipedia, and is likely a source of unaddressed, perhaps unrecognized, conflict in discussions of how the categories should be managed.

Browsing through the category system in order to examine the types of relationships that exist between superclasses and subclasses, the predominant relationship is one of association. Strictly speaking, an associative relationship is not a type of hierarchical relationship. The category system of Wikipedia, although perhaps envisioned to be a hierarchical system by some in 2004, is now full of categories related to other categories by an associative relationship. This is significant because although the designers felt that they were creating a hierarchical category system, many of the relationships in the category system are not hierarchical.

The fact that the relationships between supercategories and subcategories in Wikipedia include both hierarchical relationships as well as associative relationships is due to the fact that the category system emerged from a community in which there were divergent views of what the system should look like. One of the editors who contributed to the discussions in the dataset stated:

> So I think we need a way of distinguishing between a category where (a) you are asserting that everything in the category is an example of the thing it is in (ie list categories), and (b) categories where you are just providing hierarchical links for convenience. (editor 1)[1]

The first type of category they describe encompasses the first three types of hierarchical relationships described by Aitchinson et al., generic, whole-part and instance [Aitchison et al., 2000]. The second type of category this editor describes would make use of the nonhierarchical, associative type of relationship. This editor is highlighting the need to be clear about the different types of relationships in the category system as it is being designed and created, and the differences this editor points to are elaborated in the KO literature.

---

[1] Quotations are from period 1 of data collection unless labeled 'period 2.'

Another editor provided the following example of why one page might need multiple category designations:

> I'm thinking about some of the dog topics. For example, dog is a member of pets; dog is also a member of mammals; both mammals and pets are members of animals but neither is a subcategory of the other. Now, how about dog agility? It needs to go under the dog sports category, which needs to be under the dog category, because it's related to dogs. It also needs to go under the sports category, because it's a sport. It probably also needs to go under the hobby category. But dog and sports do not at any higher point in the hierarchy have a common parent. (editor 2)

This statement is a clear articulation of the need for polyhierarchy. The editor would like there to be hierarchy, but would also like a single category to be able to belong to more than one superclass. This is an excellent example of the community working through the issues until they come to a point of recognizing what they need. The advantages and limitations of this type of hierarchical relationship are well-documented in KO, and the Wikipedia community recognized the same issues when planning out the category system. By 2006 the many editors were aware of the disjointed state of hierarchy in the category system:

> Welcome to the chaotic state of categorization on Wikipedia. As has been discussed extensively here and elsewhere, much of the chaos arises from two aspects: 1) there is no clear distinction between strictly hierarchical categories (article X is a type of/member of category Y) and associative categories (article X is somehow (often tangentially) related to category Y). The related to categories can turn into trails of free association. 2) Some editors will inappropriately

> remove an article from a parent category and instead place the article's eponymous subcategory within the parent category. This can lead to very strange hierarchies. Well, and aside from those two fundamental problems, there is always the problem of inexperienced users naively (mis-)applying categories. (editor 3)(period 2)

4.3.3 Scope of the Wikipedia Category System

Another theme that emerged from the discussions was the scope of the category system. The editors were very concerned about the number of category labels that would be applied to each page. One editor commented:

> Even if each of these categories is relevant (which can be doubted) the original page starts to clutter up rapidly. Logically, there is no almost limit to the extent to which categories can be applied to any page for the imaginative editor. (editor 4)

This editor was worried that categories would be assigned unevenly. Some individuals would choose to apply a large number of category labels, while others would apply few. Some shared the concern that the number of categories applied to different articles would be widely divergent. One argument put forth was:

> I suggest that there should be a Guideline for categorisation by which editors (1) exercise caution and err on the side of not ascribing a category unless the text of the page justifies it (2)limit the size of the categorisation link text so that it remains small in relation to the size of the page. (editor 4)

While this editor clearly wanted to create hierarchical relationships between categories, they had observed how inconsistently these hierarchies were constructed.

Other editors felt that the work of scoping the category system of Wikipedia was such a large task that it should be modeled on existing structures for information organization. One editor brought up the challenge of making relationship types explicit and suggested modeling the category system on the Resource Description Framework (RDF).

> The fix is to label the arrows: describe the relations. This is, in my limited understanding, what RDF does. That uses the terms subject, predicate, and object. The subject is the thing you're categorizing. The object is the category you're adding it to. And the predicate describes the relation. Predicates allow you to make semantic inferences programmatically. (editor 5)

Other editors suggested modeling the category system on extant directories created to index the World Wide Web.

> To minimise reinvention of wheels, consider the category structures of Web directories such as www.zeal.com, which have been painstakingly thought out over long periods. (editor 6)

This editor is pointing out how much effort could be saved if the category system were modeled on an extant structure. The design, creation, monitoring and evaluation of the category system required significant community effort. In order to create the most effective system, many discussions were based around how best to scope the category system. Members of the community expressed concern over inconsistency in the average number of categories that might get applied to a given page, made appeals to modeling the syntax of the system on RDF and suggested web directories as other potential models for the category system.

4.3.4 Navigation via the Category System

In the first period of data collection (June, 2004 through January, 2005) the hyperlinked category labels for each Wikipedia page were displayed at the top of each page. However, it is now the case that the category information is displayed in a box at the bottom of each page. There are many pathways to any individual page in Wikipedia. Users arrive to specific pages via a link from a results page from a search engine, from a link in the text of another page, from a link in an infobox located in the upper-left or upper-right- hand corners of many pages which typically contain pointers to a large amount of related content, from a list of links, or from the list of pages provided on the page for any category. At this time it is unclear how often the category system is used for navigation in Wikipedia. Regardless of the current reality, many of the editors who contributed to the design of the system in 2004 felt that navigation was a primary way the category system would be used. One contributing editor articulated the following vision of navigating through the category system:

> We have to think from the encyclopedia user's point of view. He/she is starting at the top level of the hierarchy with a subject in mind, and they need to know which blind path to go down to find an article on that subject. It might help to think of the problem as a game of twenty questions. The first question we may ask is, ''If they wanted to know about Stephen King's books, they might choose Category: Things, and have a choice of Category:Animals, Category:Vegetables, Category:Minerals, Category:Ideas, etc., and go down one of those paths. My point is, Categories link only as a hierarchy; Wikipedia articles link as a network to every related article. So as long as the user reaches the article on Steven King (the person), or the articles on Steven King's books using the categories, the articles themselves link to each other.

> (editor 7)

This statement clearly indicates that the editor felt people would begin their search by looking at the category system from top to bottom for desired content. Lee and Olson compared the hierarchical navigation structure of Yahoo! directories with information retrieval via keyword searching in a search engine. One of the factors that they consider in their study is the location of the hierarchical browsing tool on the Yahoo! main webpage [Lee and Olson, 2005]. They noted that the harder it was to find the directory/ category information the less it was used. From looking at the discussions in this dataset, it is clear that the decision to move the category box to the very bottom of each page predated community understanding of how the category system would be utilized. Another editor presented a contrasting vision for how the category system would be used.

> On the other hand, I think there's a good case to be made for a more bottom-up approach; let's take a look at how things are being categorized, and try to find the patterns in that. It's more the Wikipedia way, too. For example, I've noticed that there are a lot of categories that are nonplural, such as Category:Medicine, Category:Biology and Category:Law. In those cases, rather than being categories containing only one article (Medicine, Biology and Law, respectively) they are instead full of articles and subcategories that are about the indicated topic. (editor 8)

This editor is articulating a need for specific guidelines for term construction to facilitate vocabulary control, another example of the Wikipedia community echoing principles frequently discussed in the KO literature. This editor's comments suggest that some members of the community felt that the amount of effort that was being expended in the design of the category system could be reduced if the purpose of the category system were explicitly articulated.

4.3.5 Purpose of the Category System

In the initial period of data collection and coding I created several codes relating to the purpose of the category system. These codes were few in number and, in general, were used to label sections of text on the discussion page in which editors expressed desire to come to agreement on the purpose of the category system. In the second round of data collection, many editors expressed opinions about what they felt the purpose of the category system to be, thus I added more codes relating to purpose. This was such a frequently-discussed topic that I chose to add a fourth theme, that of purpose. Some editors expressed that categories were tools for browsing:

> Categories are intended to be an aid to browsing, rather than an general taxonomy (which would be POV and destined to fail). (editor 9)(period 2)

One editor stressed that their best use would be to support browsing, in particular, as opposed to search:

> You don't need categories to find a particular article. Categories are not a search tool, they're a search-for-related things tool. If you want to find a particular article, like Portland, Oregon, just type Portland, Oregon in the search box. (editor 10) (period 2)

Other editors raised the question of whether browsing was the primary purpose:

> A question concerning the purpose of categories: Is the primary purpose of categories to: Aid the reader in finding material that may be of interest, or relevant to a particular topic? Producing a taxonomy; wherein being included in one or more categories is an indication|nay, a declaration by the

> Wikipedia community|that the subject of an article is an instance of the category it is included in. I seem to suspect the latter... (editor 11)(period 2)

Contrasting purposes were raised:

> There seems to be a dichotomy between those who are looking to hone categories into encyclopedic taxonomies and those who are looking for a tagging system in which they can do keyword searches. The more we push at removing overcategorization, the more there is a need for a simpler tagging system. If we can answer that need, it might make everyone happier. (editor 12) (period 2)

The fact that multiple, sometimes conflicting, conceptualizations of the purpose of the category system were evident two years after the category system had been introduced is a challenge that the community frequently discussed.

4.4 Modes of Collaboration around Categories

Large-scale collaboration among editors of Wikipedia to create the category system is one of the primary differences between this system and development in the KO literature. Negotiations between editors take place around each decision that is made about the design and implementation of the system. While there is a far greater number of people contributing to the category system in Wikipedia than historically have labored over the conceptualization of a classification system such as the Dewey Decimal Classification (DDC), progress can sometimes be impeded by disagreements. In the discussions I observed several prominent themes related to collaboration. I identified four modes or styles of collaboration around the category system that were assumed by the participants:

Collaboration with the category system - This theme describes discussion in which editors were conceptualizing the category system as an entity that facilitates navigation and or retrieval, clustering or conceptual visualization.

Collaboration over the category system - This theme describes discussions in which editors were conceptualizing the category system as an object of work that individuals must use and manipulate. These include be debates about how categories should be applied, what types of relationships should exist and be made explicit between categories.

Collaboration through the category system - This theme describes discussion in which the editors were conceptualizing the category system as a mechanism for communicating and interacting with others.

Tools for category collaborations - This theme describes discussions in which editors discuss tools they are using to facilitate collaboration with, over and through the category system.

In the following section I illustrate how participants in the discussions appealed to each of these styles or modes when making a case for features in the emergent category system.

4.4.1 Collaboration with the Category System

The theme of collaboration with the category system is applicable to discussions in which editors discussed how the category system will be engaged with by users. One contributing editor articulated the need to balance the workflow of editors who are applying categories with the needs of Wikipedia users.

> People are creating categories from the bottom up because that's the easiest way for editors to work -- they put the four Beatles together in a category then lump them together into larger categories, because few people want to attempt to create a list of hundreds, or thousands, of articles. But however it's done, we have to make it easy for encyclopedia users to navigate from the top downward. Vegetarianism is fine within the discipline of Food and drink, as long as it isn't within a subcategory that's a list of foods or a list of drinks. (ed-

itor 7)

Another editor, replying directly to this comment, stated the importance of reducing the number of clicks users would be required to make in order to get to related content of interest.

> I think we have a philosophical difference here. You rightly talk of the importance of top-down, and of a properly understood and maintained hierarchy. I agree completely. Where we disagree is that I think the wiki can encode much more than that (without breaking the behaviour you would like). It's clear that this is what people are trying to do, but only by breaking the hierarchy in the process. I also take the view that people are more likely to start in the middle of the hierarchy than at the top: they'll google their way to a Wikipedia article, spot the categories, and jump into the tree. Where do they go from there? From a user's perspective, categories are primarily a navigational tool. Frankly it would annoy me intensely if I surfed from a footballer to Category:Football (soccer) players and then had to start from the top of the hierarchy to reach Category:Football (soccer) rather than follow a single link. That link could go in a ''Related links" section of the Football players category page, sure. But that's a kludge, I'm afraid, and throws away what could be meaningful data. I think the approach to take here is to add the ability to have relations in the category. It doesn't remove anything from the system that we have now and may add something. In the first implementation off this, all that should change is that the category page would have multiple lists, one for each relationship, rather than the single list it has now. This shouldn;t be too hard to code up and add

> extra possibilities without any downside (apart from the implementation time). (editor 13)

Another contributing editor highlighted the importance of categories as a support for browsing.

> My boosterism of functional categories has been in support of that skimming, browsing user. Reading the above description, although intriguing, I must confess I have never considered the category tree as supporting that sort of precision datamining search. Wikipedia strikes me as more of an imprecise, people-to-people exercise in information transfer, like any traditional encyclopedia in that sense. (editor 14)

Editors were very concerned with making decisions that would ensure the design of a category system to support navigation and retrieval, clustering of related content and conceptual visualization. In this way editors were designing a category system that would itself be a partner in the collaborative process of expanding Wikipedia.

4.4.2 Collaboration Over the Category System

The theme of collaborating over the category system is evident when editors discussed the category system as an object of work that individuals must use and manipulate. One editor expressed a need to find a solution to sorting issues within categories:

> Consider the situation of a Category containing 30 articles with Sort Keys Book 01, Book 02, ..., Book 30. These would all appear under B using the current system; this would still hold if the threshold was measured against the number of articles as opposed to the number of "sort buckets". The lat-

> ter is what I want to measure. The system which has been unilaterally adopted in Category:Harry Potter movies will only work for a series up to 9 items, since an article with Sort Key 10 will appear under 1 and screw up the sorting arrangement. I would prefer to use the system I originally installed, being more scalable, but am unwilling to impose it without some discussion. (editor 15)

This comment demonstrates concern for how design decisions affect user experience and also serves as an appeal to other interested parties to help reach a majority opinion on how to mediate this issue.

Editors also discussed how the category system might help users conceptualize topics in relation to one another.

> I'm not sure why ''ease of maintenance" is an issue on this, or why that overcomes the great navigation and classification benefits that have previously been mentioned. Articles that define categories are not only the parents of those categories but will also logically be a member of whatever parent categories their own categories belong to. The articles should reflect this, for navigation purposes, as well as to properly classify the article. These are the two functions of categories. A reader of an article may not want to read just more topics on that article, but to see others of the same kind, and he may not even know that such a parent category exists without the article being tagged with it. Categorizing Ohio only under Category:Ohio just tells you that there are more articles about Ohio. A non-U.S. reader in particular may not assume that clicking through that may take him to other categories on other states, plus he may wonder why Ohio isn't classified as a U.S. state, if that's what the

> article tells him it is. Why unnecessarily increase the steps required to find what should logically be right there? Why omit classifications on the articles that are obviously the most (or all equally) important instances of that classification by virtue of their having a subcategory? (editor 16)

Concerns over how the category scheme as an object of work that individuals must use and manipulate required the editors who participated in the design and implementation of the category system to collaborate over the system design. This collaboration entailed providing scenarios with different outcomes in relation to issues of concern. It also involved explicitly soliciting the input of others before a decision would be implemented. I discuss tools that were used to facilitate such calls for collaboration below.

4.4.3 Collaboration through the Category System

Collaboration through the category system is a theme that applies to discussions of how the categories themselves might serve as a mechanism for communicating and interacting with others. Feinberg argued that knowledge organization tasks are vehicles for the expression of the creators' beliefs [Feinberg, 2011]. The discussions about the Wikipedia category system expand her argument in the realm of systems in which the design has been massively collaborative. The participants must reconcile which points of view will be expressed through the organization and assignment of categories and how the resulting system affects adherence to the neutral-point-of-view policy.

Editors were concerned with how the assignation of categories might violate the neutral-point-of-view policy.

> We as a group have begun in a few different places around WP to identify a potential problem with POV in categorization. It seems this is happening when a category is created that

> has a negative connotation and no self-evident criteria for inclusion/exclusion. (It probably could also happen with a category having an extreme positive connotation, but I haven't seen that come up yet.) I think it may be useful to understand more fully just which actual, current categories are subject to this phenomenon, so that we may draw better-grounded conclusions after inspecting a more full set of actual examples. To that end, I'm starting an alphabetical list here (feel free to chip in) of categories I think are likely to cause POV controversy. (editor 14)

This comment provides evidence that the community was aware of the potential for expression in the act of naming and applying a category. This discussion also contains a direct appeal for collaborative effort toward the end of identifying examples of categories that might be contentious. Another example of how categories were seen to have the potential to be expressive of bias is evident in the following discussion:

> I really don't like the idea of categorizing people by race, religion, or sexual orientation, so Category:Gay people should go, and Category:Jews and Judaism should be just Category:Judaism. Gay rights activists would be a proper category, however, as would Jewish religious leaders, as long as it is categorized by something someone does rather than what they supposedly are. I think categorizing people under Category:African Americans or Category:Asian Americans is highly offensive and POV. Whether someone is one or not is largely a matter of self-identification (how do you label yourself if you are multiracial?), and it is inherently POV to think that people are appropriately classified based on what race they are, as if that is a defining trait. It is much less offensive but still prob-

> lematic to merely include this...information in list articles, because at least that way you're not slapping a classification on the subject of an article, saying ''this is what he is". (editor 17)

4.4.4 Tools for Category Collaborations

As mentioned above, there were multiple points in these discussions where appeals for collaboration were explicitly made. Several tools were named that editors were using to facilitate this collaboration. On the topic of how category links are highlighted if they are created but left without being assigned to a parent category, one editor invited others to help address the problem.

> If people created categories responsibly, there would never be a redlink category. A red linked category does not mean it doesn't have a parent. It means it has no description. (And if it has no description, it can't have a parent, but that's not the point.) There are many criteria that determine if a category ''exists"...does it have articles? Does it have a parent? Does it have an article (description)? Only the last of these means anything to the link color. If you hate red categories so much, maybe you'd like to join me in fixing them on Category:Orphaned categories where I have lots of them listed. (editor 18)

The function that such categories are highlighted with the color red could also be interpreted as a tool to support collaboration itself as it indicates to editors that the category lacks a parent, a vital part of the category assignation process. But as well, in the quote we see a range of ways to determine "existence" of a category that point toward a range of technical assistance for category creation and application.

Another way that editors collaborated was by sharing recommendations for how to accomplish certain types of tasks.

> In the meantime, one useful method I've found is to go ahead and edit the article or subcategory you'd like to classify. Type in your best guess at the name of the proper category/ies, AND the name of a larger category that you're sure exists, which could be a (grand)parent (i.e. Category:Medicine or Category:Music, or Category:Musical groups by genre, as specific as you can get). Then use the SHOW PREVIEW, not the Save page button. Look for the previewed categories at the very bottom of the page (it may be below the Preview edit box, depending on your Preferences settings). If your best guess is blue, you've hit upon an existing category tree; if red, it doesn't exist or is spelled or worded differently. (editor 19)

This advice was a helpful work-around when the software did not support any type of browsing of the category system other than via an alphabetical list of all categories. Editors also made direct reference to tools external to Wikipedia.

> In Bugzilla, there is 'Bug 450: Categories need to be structured by namespace'. To some extent the lists are already structured by namespace (their name sorts them together). Contrary to the deletion log sometimes included in categories or user pages, images aren't just noise in the category, but informative. As it's easy for readers and other users to distinguish them from articles, I'd include them. As another example, one could quote Category:Saint Helena. (editor 20)

This comment indicates that some editors made use of the bugtracking software, Bugzilla, to keep track of work that needed to be tackled within the Wikipedia project. Another editor mentioned Sourceforge:

> RfE 964667 is probably the closest task in sourceforge and is currently unassigned. (editor 21)

This response was made to address a question of whether anyone had proposed creating a visualization tool that would display all dependent subcategories for a given category to facilitate the accurate assignation of categories to pages.

4.5 Discussion and Implications

Tagging is a valuable and popular collaborative technology. Many prior studies have focused on the application of tags without a clear tie to the underlying assumptions of the users who are applying those tags. The act of tagging, or labeling, an item as a member of some category of things or concepts has profound implications for the way we see or understand that item. In a collaborative context, the ability of the members to successfully negotiate disagreements over instances of labeling directly influences the progress and success of the collaboration.

The content of Wikipedia is connected as a graph structure with many pages explicitly linked to other pages using largely associative relations. Early on participants argued for the creation of a category system with clear hierarchical relationships. Many discussions of hierarchy did not distinguish between different types of hierarchical relationships, nor did they cover the challenges users would face when trying to interact with a system in which many different types of relationships would exist between categories without being explicitly described. This points to specific opportunities for category tools. While Wikipedia has some extensions that allow for the exploration of the category system, there is nothing that specifically visualizes the assumed or real relationships among category nodes. A tool that in some way displayed the relation among categories would help regular users navigate with the category system and help individuals who wanted to tag pages.

Members of the community were also conscious of the issue of appropriately scoping the category system. They worried about consistency in terms of the num-

ber of categories and supercategories that might be applied to a page. They also suggested external models for the structure and syntax of the category system. They recognized that models and syntax structure would require work to create and maintain. Further, they also recognized that they might become a barrier to entry for newcomers who want to participate by categorizing uncategorized pages. Scoping tools represent yet another possible technical enhancement. A scoping tool could help users understand whether a page might be over or under categorized. Similarity measures between sets of category tags and the text of pages could suggest new categories that might be applied or categories that may be unnecessary.

The category system was initially conceived as a navigational tool to complement traditional search. Many recommendations were made as to how to facilitate navigation through Wikipedia using the category system. It is unclear what role the placement of the category tags at the bottom of articles plays in the utilization of the system for navigation, but studies of other systems suggests that it may have a negative impact.

Wikipedia currently relies on search as the primary navigation system. But with all of the labor that has gone into the category system, an obvious enhancement would be to leverage the category system to provide users more contextual information about the content that they are viewing. Visual snippets of the categories relating to the page alongside their super- and subcategories could help users understand a specific article or could be attached to internal page links to present more context about the possible target page.

There is much future work to be done in this area. In particular, while there has been work to look at the relative distribution of tags [Kittur et al., 2009] the growth of the category system itself has not been considered. Further how the growth of the category system mirrored or has not mirrored the application of category tags would also be important to know relative to existing studies of other collaborative tagging and folksonomic systems. The approach that Spinellis and Louridas employed to examine the growth of Wikipedia via links to pages that do

not yet exist and are subsequently created would be a useful model to explore for a similar study of the expansion of the category system [Spinellis and Louridas, 2008].

Conclusion

This study analyzes the early efforts to collaboratively create a category system for Wikipedia. Through analysis of group discussions, I saw the collective concerns for how the category system would be structured, how it would be applied, and how it could be useful for future users of Wikipedia. This analysis unpacks some of the collective and social concerns that individuals had about the creation and use of a category system. It also analysis extends what we know about social tagging systems and the collaborative creation of category systems. The analysis of how the category system functions to enable different styles of collaboration is important in relation to the other collaborative tagging studies. This rubric of collaboration styles should be tested in other large-scale collaborative projects where category schemes are fabricated and used.

While we know from the work of Kriplean, et al. that the labor invested in maintaining and improving the category system is valued work in the Wikipedia community, the system itself seems to be underutilized [Kriplean et al., 2008]. If we harness the power of the information structures built by thousands of editors to provide context for each article among related content in order to present such relationships visually to users, this would provide an alternate navigational option with the potential to support sensemaking. If we are able to display category information that would support tagging decision making, we could encourage increased participation in the expansion and refinement of the category system, especially among novice editors.

Chapter 5

THE SPECIALIZED WORK OF TEMPLATE CREATION AND MAINTENANCE

In the following sections I begin by briefly introducing Wikipedia templates and a few details about infobox template work. I cover some related work on studies of specialized work in Wikipedia. I then dig into an analysis of infobox editing activity. This analysis of editing activity was motivated by two questions:

1. RQ 3: How are infoboxes created?

2. RQ 4: How can we measure and distinguish different styles of infobox template editing?

I describe the data collection and the resulting dataset for the analysis. I describe three different metrics to unpack infobox editing activity and the different types of activity I was able to identify. I conclude by considering the role of rarified types of work and how they support large scale collaborations and structured data interchange across the web.

5.1 Boxes of Information

Elaine is eating her morning breakfast cereal and sees the gold medalist for women's ski jumping pictured on the box. Elaine has never heard of this athlete before, and she decides to enter the gold medalist's name into the search bar of her web browser to see what information she can find. Elaine's search generates a list of results on the left-hand side of the page and a rectangular box of information on the right-hand side of the page. This rectangular box has a picture of the athlete in mid-jump at the 2014 Sochi Winter Olympics. The box also has the athlete's name and some biographical information. How convenient! Elaine didn't even have to leave the search engine's

page to find her answers! Elaine wonders, where did the search engine get this information? Did her favorite search engine compose the information into this stylish presentation? Likely, this information is coming from somewhere else; but where? Further, this information was probably composed and curated by someone; but who?

Across the web there are thousands of places that collect and curate User Generated Content (UGC). These sites structure user content through internal links, references, and sundry markup languages that make bits and pieces of the information accessible. That these pieces of information live on the web makes them accessible to people and machines alike. The structure, in whatever form it might be present, facilitates the representation and re-representation to people across different systems, in different web presentations.

One of the largest sources for structured data presented by search engines is Wikipedia. Commonly the structured data that is re-represented by search engines comes from infoboxes. Infoboxes are a type of Wikipedia template that has become popular. Infoboxes are representative of one class of templates designed to support information presentation to readers of Wikipedia on article pages. Infoboxes are tailored to the specific content of the page, and provide a quick visual summary of many facts that are present in the article.

This project considers the individuals who create and maintain infobox templates as a specific example of a specialized type of hidden work in a large scale collaborative environment. Studying highly specialized work can be difficult because of how rare it may be in an environment. For example, access to participants for interviews could be difficult because of how few people may do the work while quantitative studies may not be able to enumerate enough observations to reach significance. The scale of participation and lifespan of Wikipedia is showing value for allowing the research community to begin understanding complex specialized work. Compared to article editing work, the work on templates is a tiny fraction of the total edits, making template work hidden; completely obscured by the sheer volume of other activity. Yet templates get reused and repurposed within Wikipedia and across the Internet.

5.2 Templates

Templates are used in Wikipedia to reduce the work it takes to maintain and elaborate structured content by reusing helpful pieces of structure. This structure is distinct from the content that is placed within it. Just like the creation of a new encyclopedia article, any editor of Wikipedia can create a new (or derivative) template. There are several different types of templates. Some templates post commonly used messages, such as welcoming a new user, or warning a vandal. Some templates, like userboxes, allow users to express individual attributes like their hometown or their current total edit count.

Templates are Wikipedia pages themselves; they are edited and discussed just like any other page in Wikipedia. The templating system works based on the principle of transclusion. Transclusion allows reuse of text, images, or structure by marking up the content to facilitate extraction and display on other pages. Figure 5.1 shows sample infobox template code (a), that same template on an appropriate page with specific parameters (b), and what it looks like when displayed on the article page (c). As of June 2013, there are more than 11,000 different templates for information boxes (infoboxes) in Wikipedia. For infoboxes, templates govern the structure of the information; the data parameters of the box, and how it is laid out when presented on a page. An editor using an infobox contributes the content that is relevant to the specific Wikipedia article by filling in the appropriate parameters. When the article page is requested the parameters are merged with the structure of the template and the result is transcluded into the article page before it is sent to the user.

The work of templates generally, and infoboxes specifically, is of two kinds. First, there is the specific work of editing the code or script of the template (left side of Figure 5.1). The templating language for the templates that we study in this paper was homegrown and has some quirks that make it difficult to learn and master. It is this first type of work that this paper unpacks, the scripting. I call the Wikipedians who do this template coding 'Templatiers'. The second form of tem-

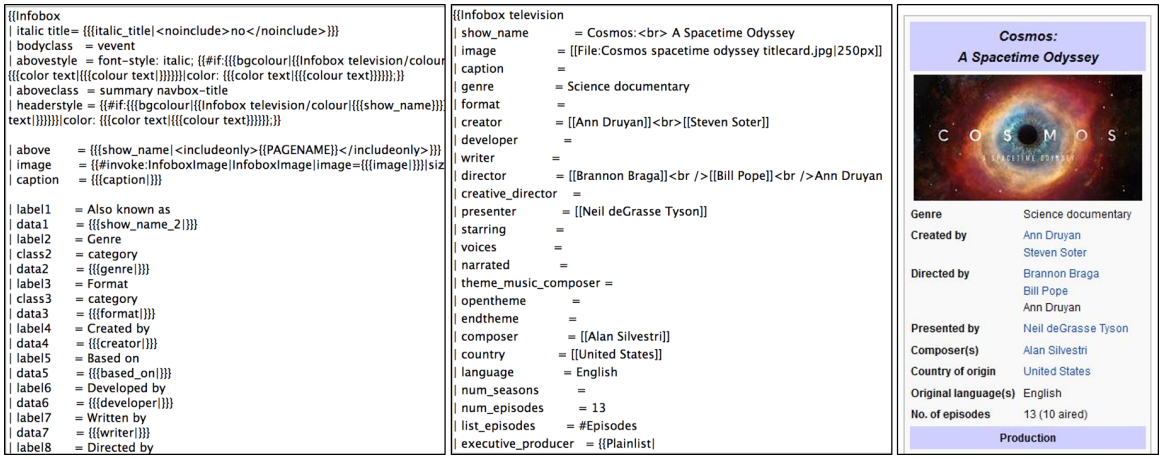

Figure 5.1: Shows sample infobox template code (a), that same template on an appropriate page with specific parameters (b), and what it looks like when displayed on the article page (c).

plate work is the effort necessary to place a template upon an article page. While this paper does not focus on that work, it is essential as the work to parameterize the template on a specific article page can also be quite difficult and is the crucial link between information content and information structure that makes infobox content reusable across the web.

5.3 Studies of Focused Work in Wikipedia

The editors who do the work related to templating in Wikipedia are participating in a type of work that is not as visible as article creation, contributing images, or fighting vandalism in the system. While the work is not highly visible, these templates are used on many pages in Wikipedia and are seen by users of the system every day. As a group, the infoboxes of Wikipedia comprise one of the largest publicly-available, creative-commons-licensed sets of structured data on the web. Considered from the perspective that editors who contribute to templates are creators of this set of structured data, they participate in work that is seen and or used by millions of people each day. However, the details of this work have not been studied.

At a general level, the work of these template editors is a form of scripting or coding. An early study of scripting in the context of spreadsheets illustrated a number of collaborative aspects of scripting work [Nardi and Miller, 1991]. In an organizational setting, individuals who had specialized skills often generated scripts or complicated formulas to solve a particular problem. Others, often with less-advanced scripting skills, would reuse scripts to tweak and modify them to fit a different, but similar problem. The patterns of reuse and sharing often follow relationship patterns of who knows whom. Scripting and coding exemplified in many of the studies of open source coding activities share similar insights about the specialized nature of this type of work. This insight lead me to frame my question about whether Wikipedia editors specialize in templates, or whether they are more likely to specialize in topical content and edit the related templates as needed.

There have been studies of specialized work within Wikipedia. The far majority of studies have focused on editing work. That is, the work of creating and refining articles in Wikipedia. With a few exceptions most of those studies have focused on quantitative understanding of editing. There has been some focused study of other specific forms of work, like vandal fighting [Geiger and Ribes, 2010]. Keegan et al. found that breaking news is a unique type of specialized work in Wikipedia, and that collaboration around these types of articles is different than collaboration around other types of articles [Keegan et al., 2012]. Wattenberg et al. found that there are patterns to how administrators in Wikipedia allocate their time and effort [Wattenberg et al., 2007]. The work of Kriplean et al. proposes a typology of work in Wikipedia [Kriplean et al., 2007]. Their typology identifies editing work, social and community support actions, border patrol, administrative, collaborative actions and disposition, and meta-content work. Under this typology, creating and editing templates is meta-content work. While the work of templating fits well into this category, there has not yet been a study of the types of work that goes into the design of these templates. This study focuses on the specialized work of template editing to elaborate an understanding of the type and diversity of work

which sustains large-scale collaborations like Wikipedia.

Star and Strauss describe a continuum of work from visible to invisible [Star and Strauss, 1999, 15]. They theorize three areas of the continuum: creating a non-person, disembedding background work, and abstracting and manipulating of indicators. Star and Strauss describe 'creating a non-person' as work contexts in which there is a power differential around who gets to define legitimate work. They describe 'disembedding background work' as the transformation of work that is expected but not formalized into legitimate, recognized work. Star and Strauss argue that the 'abstracting and manipulation of indicators' can influence how work is measured and quantified in two ways. First, indicators are created, and those who create them have the ability to make work more or less visible. Second, the work that goes into a product is rendered invisible to the consumer through removal from the setting in which the work takes place [Star and Strauss, 1999, 15]. Another research thread that I build upon is related to how to understand rarefied types of work. I consider the concept of hidden work as the intellectual transpose of visible work, a concept described by Suchman in her discussion of "unstructured" or "informal" work that is not represented in procedural models of work [Suchman, 1983]. In some related work Suchman notes that "What we acknowledge less frequently is that bringing such [hidden] work forward and rendering it visible may call into question the grounds on which different forms of work are differentially rewarded both symbolically and materially" [Suchman, 1995]. In another study, Nardi and Engeström propose four different categories of hidden or invisible work, paraphrased as:

- Work in locations or places that are less visible to outsiders.

- Work that is perceived of as routine or manual, but which requires special knowledge, skills or problem solving.

- Work done by individuals who are socially less visible.

- Informal work activities that are rarely included in formal job descriptions [Nardi and Engeström, 1999].

The work of people who edit templates in Wikipedia matches most closely the second category, specialized knowledge of how to work within the templating system is required. Additionally, the work of people who edit templates gets transcluded into other pages, where markup allows the information box to be rendered alongside article text. While many experienced members of the Wikipedia community would be able to trace the creation of the infobox to the history of the template for that particular infobox and see this hidden work, to many casual users, the box would just appear on the article page and they might not ever be aware of all of the work that went into the creation of the information box through the templating system of Wikipedia.

5.4 Understanding The Work of Templatiers

There are several types of information structures in Wikipedia that require work from Wikipedia editors to build and maintain. Intra- and inter-wiki links, navigation boxes (navboxes), information boxes (infoboxes), portal pages and the category system all facilitate information seeking and access by Wikipedia users and all require active creation and maintenance. Of these structures, infoboxes are one key system supporting structured data in Wikipedia. Structured data is marked up in such a way that becomes machine- readable, making it broadly reusable across the web. People who edit templates, Templatiers, edit the code of the template to change presentation of the data (resizing the columns or rows, fonts, colors), to add, or rename parameters, and change how the templates interact with dynamic values, like edit counts or how other templates are included. Templatiers are leveraging special knowledge of the templating environment and community standards of information presentation.

5.4.1 Collecting Meta Work Activities in Wikipedia

I used the Wikipedia webservice API to request all pages in the template namespace that begin with the string "Template:Infobox". This request returns a broader set of pages associated with a template including the template page itself, sandbox pages, and documentation pages. I cleaned the resulting list of pages to exclude (as best possible) any pages that were not explicitly the templates themselves. That is, I removed documentation and sandbox pages from our list. This resulted in a list of 11,561 template pages, which includes some "duplicates" that result from name changes and template page moves. I then used the resulting list of infobox template pages to collect revision data, including the editors, edit timestamp, along with the text of each revision for our complete list of template pages. I then queried the API to find how frequently each template is transcluded into pages in the article namespace. That is, I ignored transclusions of infobox templates into user sandboxes and other places where users may simply be experimenting with an infobox template. In sum, the dataset includes 234,989 unique revisions, 18,956 unique editors, and 11,508 unique infobox templates.

5.4.2 Metrics for Understanding Template Specialization

There are more than 11,000 infobox templates transcluded into more than 2 million pages in English Wikipedia. Much like article pages, a few Templatiers account for a disproportionate amount of the total edits to templates. Less than 1 percent of template editors (22 individuals) account for more than 23 percent of all template edits. Figure 5.3 shows the log transformed number of revisions made by editors. Both figures display a pattern common in other editing behaviors found in Wikipedia, which is, a small number of infobox templates that are edited frequently, and a large number that are edited infrequently. Similarly, a small number of editors, like the 22 individuals mentioned above, have completed many infobox template revisions and a large number of editors have relatively few.

The characteristics of infobox template editors vary widely. I divided editors

Cohort	N	Edit Range	Median Longevity	Median Diversity	Median Influence
1	768	10-14	284.51	.0003	38.36
2	689	15-23	469.07	.0004	61.88
3	705	24-50	636.30	.0006	124.00
4	687	51-1000	1098.05	.0016	533.16
5	22	>1000	2333.01	.0419	36781.70

Figure 5.2: A summary table of the different metrics for each cohort. Due to extreme outliers in each cohort, the median of longevity, diversity, and influence is shown.

into 5 separate cohorts based on edit count (see Figure 5.2) to begin to unpack the different types of template editing activity. I decided not to include editors with less than ten revisions, as I was interested in editors who have demonstrated commitment to template editing work. The cohorts were constructed so that approximately 25 percent of the editors were placed in each cohort, with the exception of cohort 5, which will be discussed in more detail later in this section. In my analysis, I will primarily discuss cohorts 3, 4, and 5, as cohorts 1 and 2 exhibit similar properties to cohort 3. Across my analysis of these cohorts I use three original metrics to understand template work: longevity, diversity, and influence.

5.4.3 Longevity

It is crucial to the process of understanding hidden work behind infobox templates to know how long editors have been engaging with the editing of templates. To evaluate how long editors spend in the infobox template space, I introduce a measure of longevity. Longevity describes how long an editor has been active in editing infobox templates. Longevity is determined by calculating the difference between the timestamp of an editor's first and last revision to any of the pages for any infobox template. As a measure of time, longevity is a simple, yet effective,

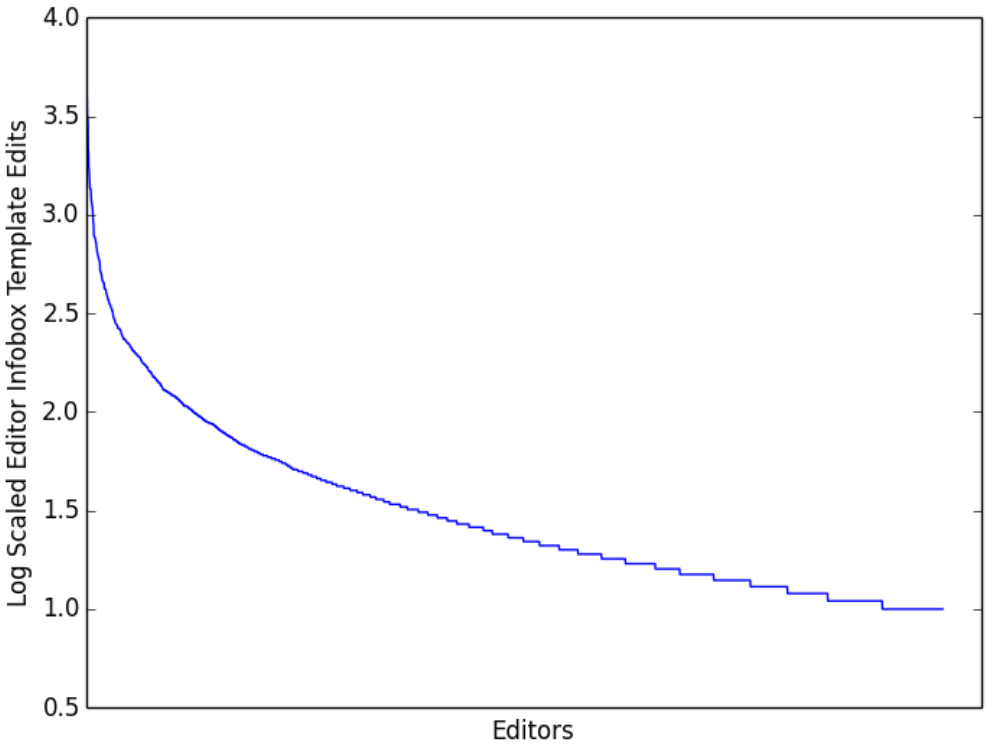

Figure 5.3: Log Scaled Editor Template Edits

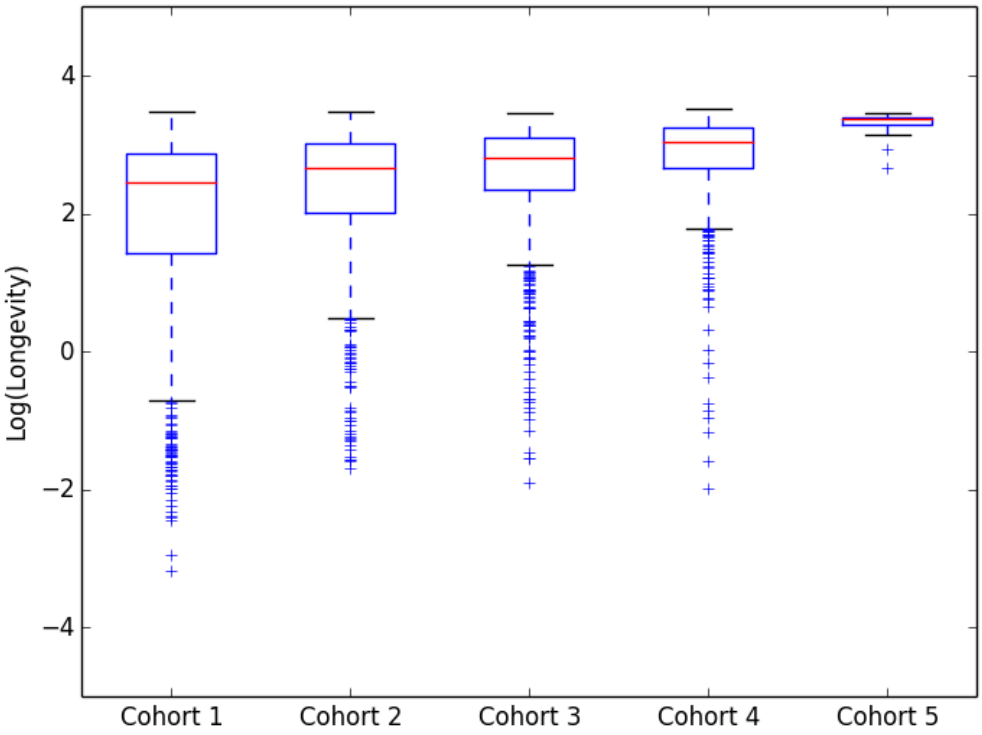

Figure 5.4: Longevity

means of understanding the lifespan of an infobox template editor. Longevity is also helpful when attempting to understand the retention rate of infobox template editors during a specific time frame (e.g. determining the retention rate between January 2010 and December 2010). It is important to note that longevity is not a measure of user activity, but rather a measure that describes the lifespan of an editor's time in the infobox template space.

For my analysis, I measured longevity in days. Thus, if an editor has a longevity of 1, that means one day passed between the editor's first revision and their last revision to any page in the infobox template space. To determine how long editors engaged in editing infobox templates, I calculated the longevity for each editor in

my dataset. As shown in Figure 5.2, and seen in Figure 5.4, the median longevity for each cohort increases as the cohorts progress from 1 to 5. Because the cohorts are delimited by edit count, this shows, essentially, that editors who have been in the template space longer usually have also made more edits.

Although the median increases with each cohort, it is interesting to see that many editors in the lower cohorts have high longevity scores. This set of editors can be viewed as having high longevity but low participation (as measured by template edit count), compared to the other members of their cohort who have low longevity, and low participation. Members in the higher cohorts tend to exhibit the behavior of high longevity, high participation. However, some outliers in cohort 4 (members of this cohort have contributed at least 51 edits) exhibit a rapid editing behavior, where editors completed many revisions in a short period of time, but never returned to edit in the infobox template space.

5.4.4 Diversity

An interesting facet of infobox editing behavior is the range of different infoboxes that a user edits. To further understand the range of infoboxes revised by editors we created a measure of diversity. Diversity represents the extent to which editors branched out and revised a range of different infobox templates. Diversity is determined by calculating the ratio between the number of unique infobox templates revised by a given editor and the total number of infobox templates. This diversity measure is useful when trying to understand if editors specialize in a specific set of infobox templates (perhaps clustered topically), or if they edit more broadly.

Diversity is capped for each user, and, consequently, for each cohort, as no editor can receive a diversity score higher than u / N, where u is the unique number of infobox templates an editor has edited, and N is the total number of infobox templates. As a result, I would expect to see diversity increase with the cohorts, as members in the higher cohorts have the potential for a higher u. The statistics shown in Figure 5.2, and the box plots in Figure 5.5 show this to be the case.

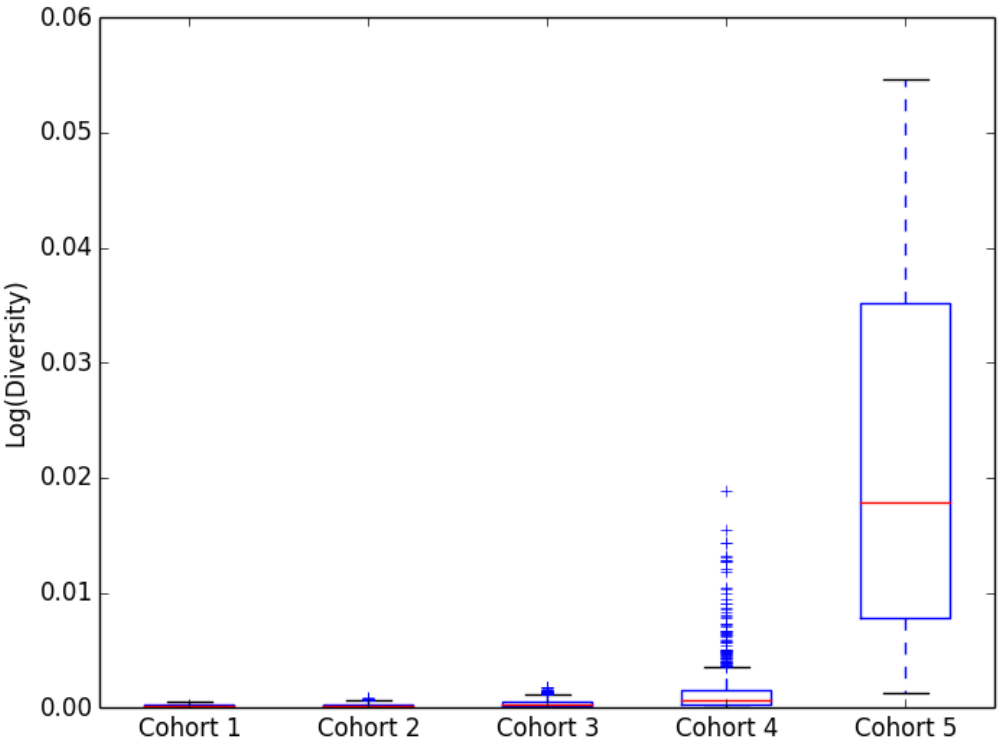

Figure 5.5: Log Transformed Diversity Measure

5.4.5 Influence

An edit made to an infobox template that is transcluded on thousands of article pages could have a more profound impact than an edit made to an infobox template that is only transcluded a few times. To understand the impact of an editor's contribution to the infobox template space, I introduce a measure of influence. Influence is calculated using the following equation 5.6:

$$Influence = \sum_{n=1}^{k} \frac{E_T}{E_E} * TR_T$$

Figure 5.6: In this equation, EE represents the number of revisions made by a given editor for the current template, ET is the total number of revisions made to the current template, and TRT is the number of times the current template has been transcluded. Influence is useful for understanding the contribution each editor has made to the infobox template space.

From Figure 5.7 it is clear that, like the other metrics, by cohort, individuals who have increasing numbers of edits, the median influence score increases. While that result may not be surprising, as it could be assumed that more editing results in higher influence, the plots in Figure 5.7 shed light on some interesting editing behaviors. The plots demonstrate that editors in lower cohorts can have extremely large influence scores. The outliers in cohorts 1 and 2 (maximum 14 and 23 edits, respectively) have higher influence scores than some editors in cohort 5. This is an astonishing finding, as this means editors with fewer than 23 total edits have higher influence scores than some editors with more than 1000 edits to the infobox template space. This finding suggests two things. The first is that some editors with relatively low participation in the infobox template space edit infobox templates with high visibility. The second is that highly active editors in the infobox template space do not necessarily edit templates with the highest visibility.

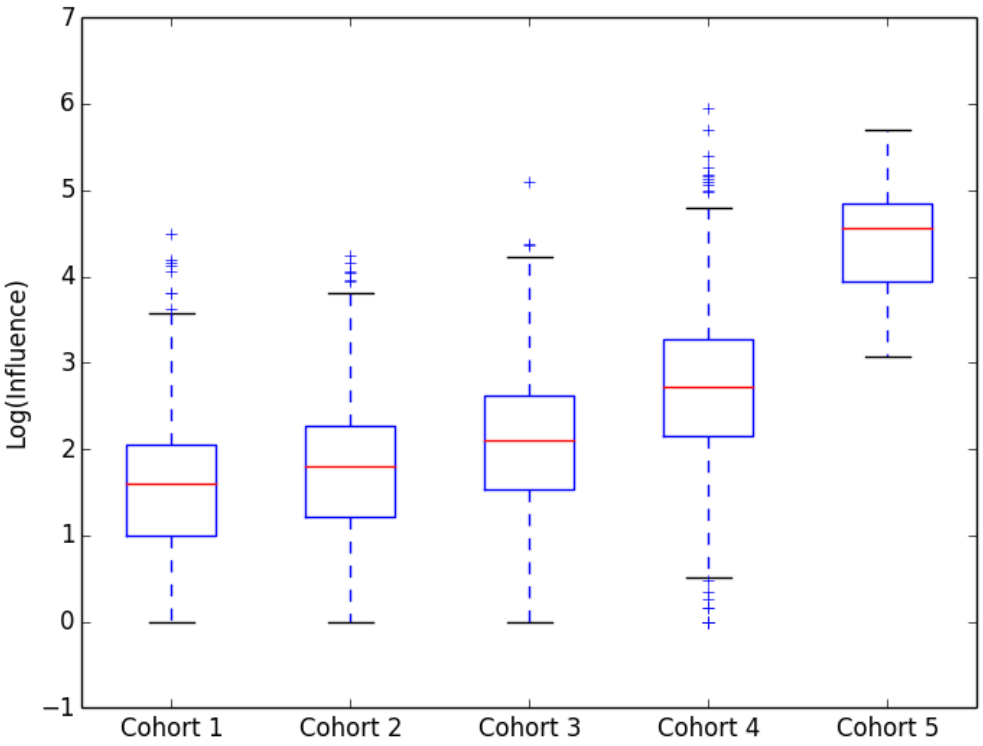

Figure 5.7: Log Transformed Influence Measure

5.5 Cohort Five

Cohort five is the smallest group, and the most unique in that many of the editors have made many contributions to the structure of information boxes through their consistent editing of many templates, or, as in some cases, the consistent editing of a small number of templates.

These twenty-two editors made between 107,225 edits and 69,969 edits to the template namespace by the time of data collection. This accounts for roughly twenty-three percent of the total number of edits to the template namespace at the time of data collection. I am especially interested in this cohort of editors because of the large volume of their edits. I found a statistically significant difference

	Longevity	Diversity	Influence
Cohort 1	$p < .0001$	$p < .0001$	$p < .0001$
Cohort 2	$p < .0001$	$p < .0001$	$p < .0001$
Cohort 3	$p < .0006$	$p < .0001$	$p < .0001$
Cohort 4	$p < .05$	$p < .0001$	$p < .0001$

Figure 5.8: The results of pairwise comparisons between Cohort 5 and the other cohorts for our three measures.

between their measures of longevity, diversity and influence when compared to the scores of the other four cohorts.

While these twenty-two editors have each contributed numerous edits to infobox templates, there are distinct patterns of work visible within this group. Some of the members of cohort five are template generalists, they have edited hundreds of templates over the span of their participation in Wikipedia.

5.6 Discussion

Templates in Wikipedia are an interesting type of information structure. They allow work to be reused, thus saving editors time and energy. They are also powerful tools for organizing and structuring information. The act of contributing to the creation of an infobox template is an act of information organization because it creates a structure to house information in context. Once the set of properties of the template for Infobox:Television has been established, we know quite a bit about how the different values contained in the infobox are related to the entity in question and to one another. Through looking at the activity of the editors who have contributed to infobox templates in the history of English Wikipedia, I uncover this human labor aimed at organizing and structuring the information in the system. I have demonstrated how to document this hidden work related to the

User Name	Infobox Template Edits	Total Edits	Infobox Edits / Total Edits
E1	5396	107225	5.03%
E2	4527	147097	3.08%
E3	4016	112461	3.57%
E4	3137	312113	1.01%
E5	3125	156759	1.99%
E6	2930	95575	3.07%
E7	2084	992473	0.21%
E8	2082	88966	2.34%
E9	1723	73168	2.35%
E10	1696	21930	7.73%
E11	1587	119975	1.32%
E12	1442	368177	0.39%
E13	1365	68726	1.99%
E14	1348	46069	2.93%
E15	1347	53157	2.53%
E16	1343	176016	0.76%
E17	1218	106732	1.14%
E18	1173	91860	1.28%
E19	1153	61976	1.86%
E20	1114	600484	0.19%
E21	1111	100754	1.10%
E22	1059	69969	1.51%

Figure 5.9: Edits to infobox templates as a percentage of total Wikipedia edits for editors in cohort 5.

creation and elaboration of templates for information boxes.

The purpose of creating metrics to help understand the work of creating templates for infoboxes in Wikipedia is that this work is currently hidden. Of the four types of hidden work outlined [Nardi and Engeström, 1999], the hidden work that goes into the creation of templates in Wikipedia best matches the definition:

> "work defined as routine or manual that actually requires considerable problem solving and knowledge" [Nardi and Engeström, 1999].

They go on to describe this work, providing the examples of telephone operators and bank tellers, pointing out that hidden work is often overlooked when new processes are implemented for increased-efficiency or cost-savings measures. The results I present here suggest that it is also important to consider hidden work that is ongoing, and that looking for such work purposively could allow it to be accounted for, rewarded, leveraged, enhanced or supported before the point at which it is replaced by new work processes. Do we now need an additional category of hidden work—one that describes work that is repurposed in contexts that entirely obscure the creators once it is discovered to be extraordinarily valuable, but is under-recognized when the work is being done due to the circumstances in which is it performed?

One strategy I considered for the analysis of the work of template editors was to create a metric that would allow me to calculate how impactful any individual editor's work is to the set of all infoboxes. I developed several metrics over the course of this research that I hoped would capture this impact. I have not yet identified such a metric. A primary reason that such a holistic metric is challenging to identify is that I did not find evidence of template specialists in the sense that individual editors were focusing on template editing as their main focus or specialization in Wikipedia. There have been studies of other types of specialized work in Wikipedia. Welsler et al. introduced four categories of social roles in the Wikipedia community: substantive experts, technical editors, counter vandalism and social networkers [Welser et al., 2011]. Of the twenty-two editors who make

up cohort five, templates make up less than one percent of their total editing activities in Wikipedia (see 5.9). This suggests that the individual editors who have contributed the largest number of edits to the set of all infobox templates are also extremely active in many other parts of the Wikipedia ecosystem.

Among editors with the highest influence scores within my data set (editors in cohort 5 and a small number of the editors from cohort 4) I did see evidence of template specialization. Template specialization is something like identifying an attribute that is working well in one infobox, and propagating it (or a similar attribute) across all similar infoboxes. As discussed above, influence scores need to be carefully interpreted along with several other factors to be understood since the number of times a template has been transcluded is an element of the score. Editors who make even a single edit to a frequently-transcluded template have very high influence scores.

Conclusion

Infoboxes were created to provide a summary of key characteristics of individual articles. Once the number of infoboxes increased, the community reached a tipping point where the sum of all structured data contained in Wikipedia was suddenly extremely valuable. This value was no longer tied to the individual articles, but rather to the full set of structured data that could be extracted from the system freely by anyone interested in its reuse.

Although it is hidden, the value of the structured information contained in the full set of infoboxes is profound. Structured information is frequently mined from Wikipedia to populate artificial intelligence applications or to populate repositories of structured data that other programs utilize. This information is then propagated to many other applications and programs. The importance of researching the structured information in Wikipedia lies in the frequency with which it is reused. As Bao et al. highlight, many information retrieval (IR), natural language processing (NLP), and artificial intelligence-based systems use data mined from Wikipedia [Bao et al., 2012]. While research into the optimization of these systems and their underlying algorithms is being conducted, [Wu and Weld, 2008, Lange

et al., 2010, Hecht, 2007, Syed and Finin, 2010, Tolksdorf and Simperl, 2006, Yu et al., 2007, Köhncke and Balke, 2010], little research has been done to investigate the work that goes into creating this vast set of structured data. This research discusses the hidden work that goes into the creation of these information boxes, which are then mined by many parties outside of Wikipedia to power systems such as Google's knowledge graph. Varandečić and Krotzsc discuss this particular instance of the reuse of Wikipedia structure data [Vrandečić and Krötzsch, 2014].

Out of the total number of people who have ever edited Wikipedia, the members of cohort five number only twenty-two individuals, yet their template edits account for 23 per cent of all template edits up to May, 2013, when I collected this data. These twenty-two individuals have had a dramatic impact on the creation of the set of structured data that has been produced by the Wikipedia community and reused in applications and projects across the web. Within the Wikipedia system, it requires effort (digging through the history of the template used to create the infobox and the history of page to which it has been transcluded) to determine who did the work of creating and populating the data represented within the infobox. Once the structured data has been mined out of the system, it has been entirely separated from the work that went into creating it. Perhaps this goes beyond hidden work in the context of reuse. Perhaps, once the structured data has been removed from the Wikipedia system it is no longer merely hidden work, but rather the product of hidden work, the documentation of which has been obscured. The documentation has been obscured in that it is not packaged alongside the data set, and remains hidden in the history pages of Wikipedia.

Chapter 6

CLOSE READINGS OF THE INFRASTRUCTURE OF WIKIMEDIA FOUNDATION PROJECTS

Contributors to Wikimedia Foundation (WMF) projects organize information as they carry out their editing of project wikis. Consider two hypothetical editors: Sasha and Sven. *Sasha is reading the Wikipedia article for 'Acetic acid' and notices that the* `Production` *section of the article is confusing. Sasha decides to log into her account and edit the article. She decides that an additional heading of* `Production through Fermentation` *would clarify for users how synthetically-produced acetic acid is distinct from biologically-produced acetic acid. Sasha writes several sentences describing the difference between these types of production processes and uses the citation tool to format a reference to a book called The Art of Fermentation. Sasha adds an internal link to the article for 'Vinegar' as well as an external link to an article on the web describing multiple fermentation techniques that produce vinegar. Meanwhile in Brisbane, Sven is looking up the population of Montréal. Sven notices that there are several category labels that could be added to the article for Montréal. Sven clicks edit and adds the labels* `Port settlements in Québec` *and* `Hudson's Bay Trading Company` *to the article so that the article about Montréal will now be connected to those categories. Additional editors are adding to, transforming, or removing content and structure from pages in Wikipedias across 290 languages. Editors are not only editing content, but they may also be editing the infrastructure of Wikipedia.*

How might we approach the challenge of getting a sense of where we might be able to perform close readings of infrastructure in a commons-based peer production system? Infrastructure is defined as the "pervasive enabling resources in network form" that make systems possible [Bowker et al., 2010, 98]. For Wikime-

dia Foundation projects infrastructure is open for inspection, but it is not always clear where to draw boundaries between content and infrastructure. In order to investigate infrastructure I follow the work of Star and collaborators who theorized an approach to analyze infrastructure [Star, 1999, Star and Ruhleder, 1996]. As discussed and defined in Chapter Three, Star and Ruhleder state that the following dimensions can be found in infrastructure: can be found infrastructure: embeddedness, transparency, reach or scope, learned as part of membership, links with conventions of practice, embodiment of standards, built on an installed base, becomes visible upon breakdown, is fixed in modular increments [Star and Ruhleder, 1996, Star and Bowker, 2006, Star, 2010].

I investigate these dimensions of infrastructure across three research sites: the category system, the infoboxes, and the knowledge base for structured data, Wikidata. For each site I will consider two perspectives: infrastructure used to create the structure, and these structures themselves as infrastructure as illustrated in Table 6.1. The research questions I pose for this investigation are: RQ5: In what ways does the knowledge base manifest as infrastructure? and RQ6: How are KOS instantiated in infrastructure?

6.1 Category System

The category system of English Wikipedia refers to the set of all category labels that are used in the wiki as well as the relationships between them (their placement on specific pages). This structure has been extracted from Wikipedia and reused as a convenient, machine-readable set of relationships between topics. One example of how the structure of the category system has been reused is the Omnipedia project. In this project the category system was used to align concepts across multilingual Wikipedias. For example, Omnipedia produces alignments across languages resulting in machine-readable correspondences such as: “World War II” (English), “Zweiter Weltkrieg” (German), and “Andre verdenskrig” (Norwegian) describing the same concept [Hecht and Gergle, 2010, 292]. These alignments were derived by re-

Table 6.1: Sites of Close Reading of Infrastructure

Structure	Perspective	Technical Infrastructure
category system	infrastructure used to create it	wikitext
		category template
		hotcat
	as infrastructure itself	dbpedia
		hidden categories
information boxes	infrastructure used to create them	wikitext
		template:infobox
	as infrastructure themselves	DBpedia
		Wikidata
Wikidata	infrastructure used to create it	MediaWiki
		WikiBase
	as infrastructure itself	RDF serialization
		SPARQL endpoint

lationships in the category system, and thus could be leveraged as infrastructure to then provide concept alignment in a new context, the Omnipedia application. In the following section I will discuss infrastructure in the context of how the category system was created and then discuss the category system itself as infrastructure.

6.1.1 Infrastructure Used to Create Category System

There are multiple ways to add a category label to a page. One way to add a category is to click on the `Edit` tab on the wiki interface and edit the article by typing: `[[Category:Dog breeds]]` at the bottom of the article.

Transparency

An alternative way to add a category is to use a tool such as the HotCat tool created by Magnus Manske [Wikipedia, 2016d]. The purpose of the HotCat tool is to provide relevant information about categories that are already in use. This tool reduces time that editors who are not yet familiar with the category system would need to spend in order to chose relevant categories. This tool also reduces errors in the form of misspellings. As of March, 2016 more than 450 editors self-identify as users of HotCat [Wikipedia, 2016c]. This number of editors who choose to share the information about their use of the tool is evidence of the support that the tool has in the community. HotCat is and example of the dimension of transparency because an editor does not need to know how the tool works at a technical level in order to use it to add category labels. In this way, the editor is using the HotCat tool in order to interact with infrastructure beyond the options presented in the default editing interface. The HotCat tool supports the task of adding relevant category labels to pages, allowing users to bypass the step of editing the wikitext themselves to add the category labels.

Both the manual method and the semi-automated method are made possible through the affordances of wikitext. Wikitext is text written in a wiki markup language that browsers then render [MediaWiki, 2016]. The infrastructure of Wikipedia, specifically wikitext and tools such as HotCat, see Figure 6.1, shapes

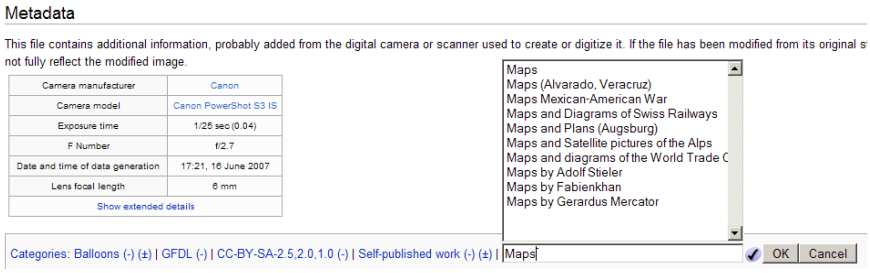

Figure 6.1: An example of HotCat in use, in this case offering completion options based on the next word in the category label [Siebrand, 2007].

the structure of information organization called the category system.

Learned as part of membership

Editors learn how to use wikitext to encode category labels into the text of pages through their membership in the community. Editors may read about how to apply categories on a help page, they may copy wikitext from an example, they may ask other editors how to work with categories, they may watch a video about working with categories, or they may learn how to work with categories at an in-person Wikipedia-themed event. The knowledge of how to add a category label to a page is not a skill that editors have before joining the Wikipedia community, they learn it as part of membership in the community.

6.1.2 Category System as Infrastructure

The category system itself is used as infrastructure. For example, in the field of artificial intelligence, researchers derived a taxonomy of concepts from subsumption relationships (*is a* relationships) in the category system. They used this taxonomy to compute semantic relatedness scores for words [Ponzetto and Strube, 2007]. The researchers found that that their machine-derived taxonomy performed competitively with human-curated resources like WordNet. The reuse of the category system in this new context of semantic relatedness is an example of how this infrastructure is leveraged, and how its use can reduce the need for human labor to

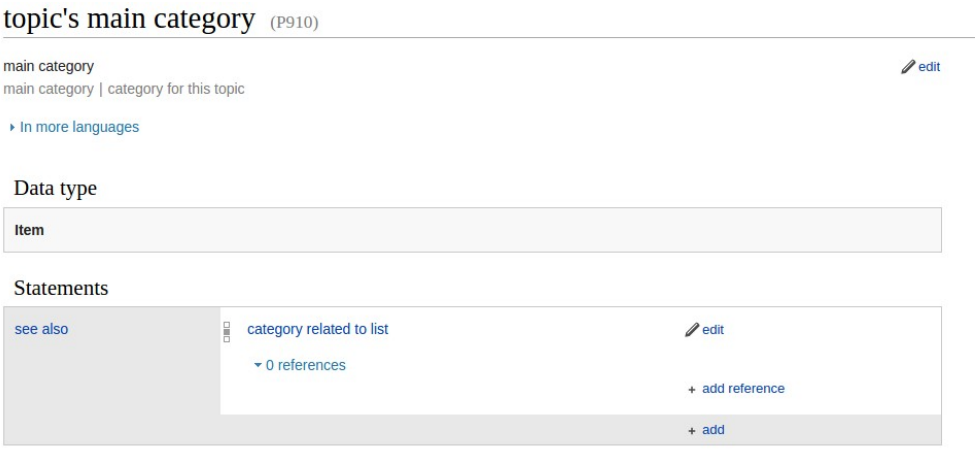

Figure 6.2: A screenshot of P910 in Wikidata [Wikipedia, 2016].

create such a taxonomy.

In addition to being applied in contexts beyond WMF projects, the category system is also reused within WMF projects. For example, the category system of English Wikipedia is referenced in the knowledge base of structured data, Wikidata, through use of Property P910, see Figure 6.2.

Embeddedness

Referencing the category system in Wikidata allows for yet another path for displaying or leveraging relatedness between concepts in the knowledge base [Spitz et al., 2016]. It is interesting to note that the category system of English Wikipedia and the category system of English Wikipedia as embedded in Wikidata are not identical. This is because both systems are constantly being revised and updated, and also because not all category information has been added to Wikidata at this point. The purpose of having a property to link up with the category system is to allow for Wikidata entities to be cross-linked to the category that describes pages about this entity in Wikipedia. For example, as seen in Figure 6.3, the Wikidata entity Q8 *happiness*, is cross-linked to the Wikipedia page *Category: Happiness*. A second example, as seen in Figure 6.4, the categories are also reused in the Wikidata Class Browser tool.

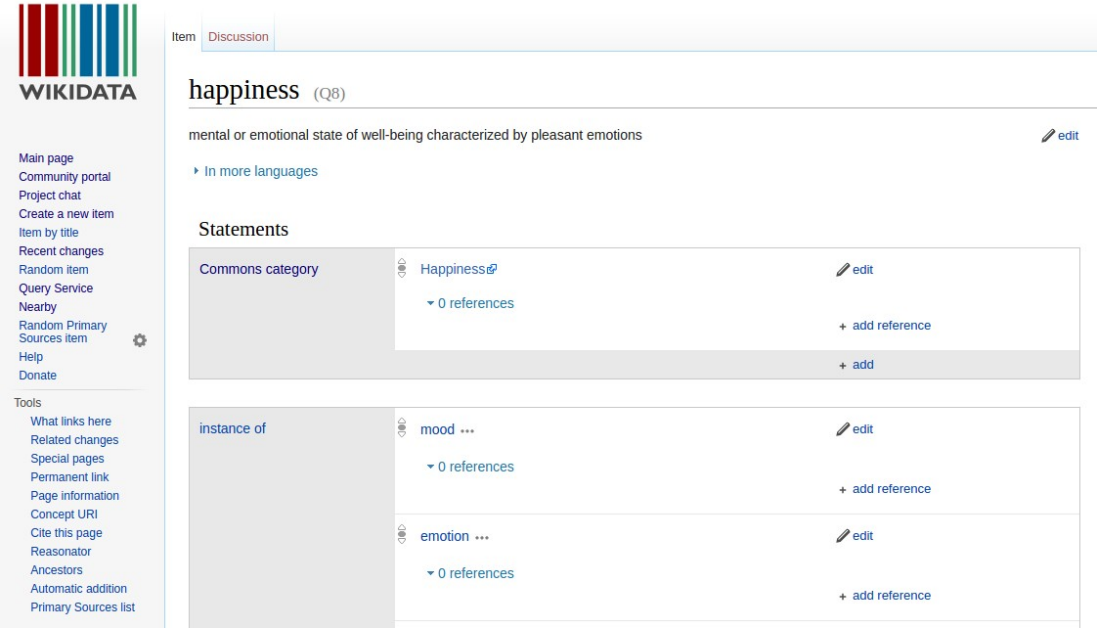

Figure 6.3: Screenshot of Wikidata item Q8 'happiness'.

In addition to the category labels that are displayed in the user interface for all users, there are also hidden categories in Wikipedia [Wikipedia, 2016a]. The purpose of hidden categories is to support administrative articulation work, for example these categories are used to tag pages in need of attention, administrators or members of wikiprojects who monitor processes, can then identify places in the wiki to focus their work [Burke and Kraut, 2008]. Hidden categories also help sort historical information such as "Accuracy disputes from November 2008", to see a full list, consult the Category:Hidden categories [Wikipedia, 2016b].

Embeddedness

Star and Ruhleder's dimension of embeddedness can be examined in the use of hidden categories by editors. In the wikis that enable WMF projects hidden categories are embedded into the wiki, and editors must have knowledge of the settings of their user accounts in order to display the hidden categories. This exemplifies the dimension of transparency in that administrative tasks are managed and tracked and assigned via the use of a structure of information organization

Figure 6.4: A screenshot of the Wikidata Class Browser reuse of Wikipedia categories.

(the category system) which also has its own social arrangements and technologies. Due to the fact that hidden categories are not visible by default, and that individual users must learn about the system in order to identify the process to display the hidden categories, we see yet another example of how infrastructure can be challenging to study because it is often pushed into the background.

Reach or scope

I observe this dimension of the category system in the fact that the category system is referenced in dbPedia [Lehmann et al., 2012, Auer et al., 2007, Bizer et al., 2009]. Through using markup languages to make explicit the relationships between categories and topics, this allows the taxonomic relationships to be reused as a feature of additional systems [Lehmann et al., 2012, 5]. The infrastructure of the category system thus has reach beyond WMF projects through reuse in additional applications.

6.2 Infoboxes

Infoboxes are laid out as a rectangular inset positioned on the upper right-hand side of a reader's screen. They display facts in the form of *Kingdom: Animalia, Phylum: Mollusca*, images, and graphics related to the content of the article, see Figure 6.5.

6.2.1 Infrastructure Used to Create Infoboxes

Information boxes are created using templates. Templates are Wikipedia pages themselves; they are edited and discussed just like any other page in Wikipedia. The templating system works based on the principle of transclusion. Transclusion allows reuse of text, images, or structure by marking up the content to facilitate extraction and display on other pages. As of June 2013, there are more than 11,000 different templates for infoboxes in Wikipedia. For infoboxes, templates govern the structure of the information; the data parameters of the box, and how it is laid out when presented on a page. An editor using an infobox contributes the

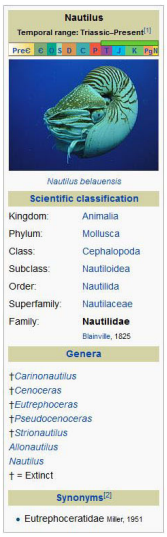

Figure 6.5: Nautilus Infobox

content that is relevant to the specific Wikipedia article by filling in the appropriate parameters, see Figure 6.6. When the article page is requested the parameters are merged with the structure of the template and the result is transcluded into the article page before it is sent to the user's browser.

Learned as part of membership

In the infobox template project described in Chapter 5, I report that 18,956 unique editors had made 10 or more edits to the templates for infoboxes by 2013. This is a small percentage of the total population of Wikipedia editors. Editing templates is specialized work that is learned as part of membership in the community. The process of transclusion is an affordance of wikitext. Wikitext is part of the infrastructure of Wikipedia that makes rendering infoboxes on articles possible. The technique of template transclusion is an example of how knowledge of how to manipulate the affordance of wikitext known as transclusion is something that people must learn.

Embodiment of standards

I observe this dimension of infrastructure in the infoboxes by the inclusion of information from standards in the infobox contents. For example, in the nau-

```
{{Taxobox
| name               =
| image              =
| image_alt          =
| image_caption      =
| regnum             = [[Animal]]ia
| phylum             =
| classis            =
| ordo               =
| familia            =
| genus              =
| species            =
| binomial           =
| binomial_authority =
}}
```

Figure 6.6: Template: Taxobox

tilus infobox information drawn from the scientific classification of the organism is included. This information is helpful to users because of the fact that it is a well-known standardized source of information in the scientific scholarly community.

6.2.2 Infoboxes Used as Infrastructure

The infobox structure has been directly borrowed by Google's Knowledge Graph [Juel Vang, 2013]. More than a decade after infoboxes had been created in Wikipedia, they began to show up on search results pages provided by Google. Google's reuse of Wikipedia content is possible due to the Creative Commons Attribution-ShareAlike 3.0 Unported License (CC BY-SA) [Wikipedia, 2016a]. This license allows for content reuse, and Google provides a citation of Wikipedia within the text of the infobox they display on search results pages. In this context the structure of information organization we call the "infobox" of Wikipedia serves as the infrastructure that allows for the rendering of infoboxes on Google search result pages.

Reach

Content from Wikipedia is also reused by Facebook [Wadhwa, Kul Takano,

2010, Park, Jane, 2010]. Facebook enriched their Social Graph with information about notable people who do not have Facebook accounts by creating "community pages", reusing information about these people from Wikipedia and displaying it as Facebook content. In this context the information boxes as well as unstructured text from Wikipedia articles are the infrastructure that allows Facebook to populate information to display on community pages. The reach of the infoboxes is beyond that of Wikipedia, the context in which they were created. The infoboxes now appear (with subtle differences) in systems that serve different purposes such as search, or social networking.

While the set of all infoboxes as rendered in Wikitext might be difficult to envision as infrastructure, the structured data that is extracted from each infobox, once aggregated as a set of all the infobox attribute value pairs from Wikipedia, itself can be considered infrastructure. Structured data is used to train machine learning algorithms related to natural language processing [Sen et al., 2014]. In this context the set of structured data that was extracted from the infoboxes of Wikipedia serves as infrastructure for third-party applications built around the specific machine learning algorithms that have been trained on the structured data. The "reach" dimension of the infrastructure related to templates can be observed through the migration of infoboxes and their content to systems beyond the WMF projects in which they were created.

Becomes visible upon breakdown

I observe this dimension of infrastructure in the templates used to generate information boxes. If a template is not working correctly, all infoboxes that are built using that template will manifest errors or will cease functioning. This is one reason that template pages are sometimes "locked", allowing only users with certain types of authority in the community to edit them, as a breakdown in a template can have widespread consequences for the pages to which the template has been transcluded.

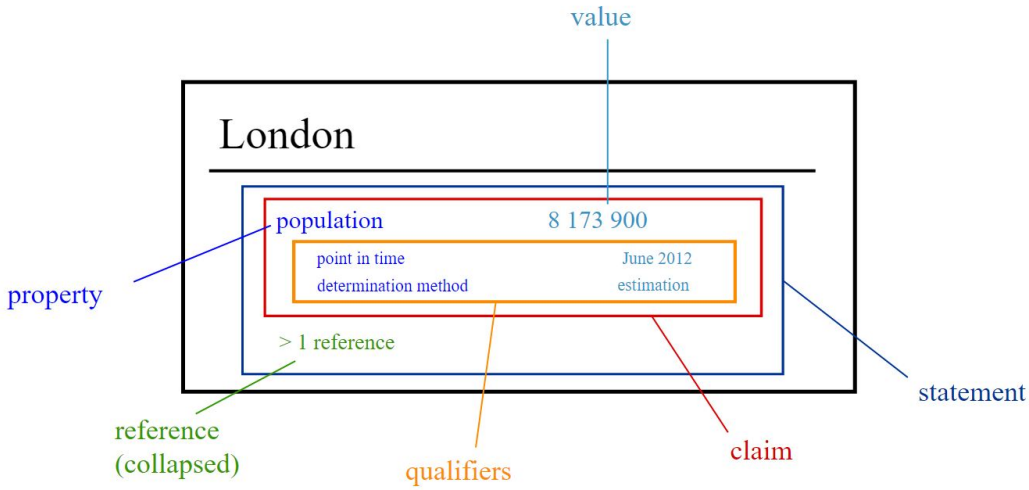

Figure 6.7: A Wikidata item labeled to indicate component parts from [Kaganer et al., 2013]

6.3 Wikidata

Wikidata is a information system that is optimized to interact with, and be part of the semantic web. In order to act as a hub of structured data, the design of the infrastructure of the information system must be negotiated in conversation with standards, accepted practices of other data hubs, and the technologies that enable the system. Thus the lines we draw around the infrastructure are somewhat arbitrary because the enabling technologies are in interaction. How do we study this infrastructure?

6.3.1 Overview of Wikidata

Wikidata is a database of linked items. In Figure 6.7 we see a labeled diagram identifying the components of a Wikidata item, Q84 for the city 'London'.

In Wikidata there are two primary components: *items* and *properties*. In Figure 6.7 the item is 'London'. The property is 'population'. The statement is an assertion relating the item to the property, in this case a claim about the value for the

property 'population' that is qualified to a particular point in time and supported by a reference to a source.

By entering the following URL for Q84[1] into a web browser, a user can see the full set of statements that pertain to Q84, 'London' stored in the Wikidata database. Each property is also a hyperlink, and via these links all items to which this property has been assigned are connected.

The Wikidata knowledge base is a database of linked structured data that contains facts about a broad range of topics. As of August, 2016, there are more than twenty million items in Wikidata [Wikidata, 2016].

6.3.2 Infrastructure Used to Create Wikidata

Built on an installed base

The designers of the Wikidata project wanted Wikidata to become integrated into other WMF projects, thus much of the infrastructure used to create Wikidata is the same as, or an extension of, the infrastructure used to create Wikipedia. The dimension of infrastructure named "built on an installed base" is evident in the context of the installed base of Wikidata. Star and Ruhleder extend the concept of "installed base" describing the pattern of how infrastructure inherits the strengths and weaknesses, affordances and limitations of the technologies in which it is implemented. In the case of Wikidata the technologies that comprise the installed base are, most notably, the software of Mediawiki, Semantic Mediawiki, and Wikibase. The choice to use and extend software that was already in use in other WMF projects is an important aspect of understanding the infrastructure of Wikidata. I introduce these three related pieces of software below.

MediaWiki is software that allows for the implementation of wikis, and is the software in which Wikipedia is created [Barrett, 2008]. The MediaWiki software has now been extended to allow for the implementation of semantic structures into the wiki engine of Mediawiki [Krötzsch et al., 2006]. This extension is known as

[1] https://www.wikidata.org/wiki/Q84

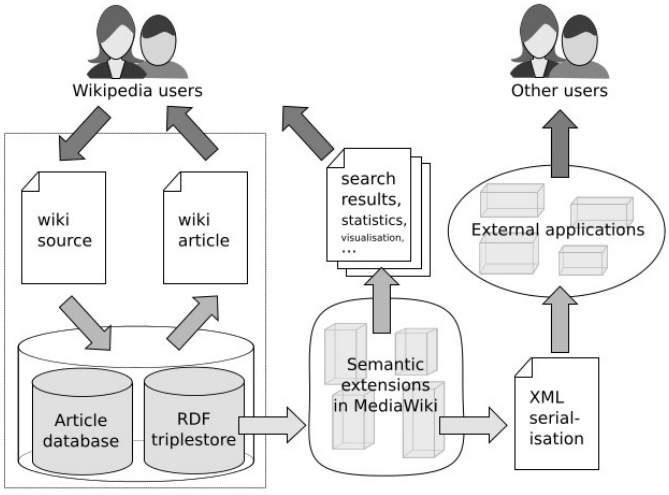

Figure 6.8: Basic Architecture of Semantic MediaWiki [Völkel et al., 2006]

Semantic MediaWiki and it allows for the enhancement of semantic information in a machine-readable format [Völkel et al., 2006]. The basic architecture of Semantic MediaWiki is represented Figure 6.8.

The relationship between Semantic MediaWiki and Wikibase is articulated here:

> "The software that powers Wikidata is a set of MediaWiki extensions collectively known as Wikibase, and though Wikibase has similarities to Semantic MediaWiki, it is a distinct set of software. However, some of SMW's backend code has been spun off into a separate library, called "DataValues", that is used by both SMW and Wikibase as a framework for storing data" [Semantic MediaWiki, 2015].

The shared history between Wikibase and Semantic MediaWiki is an important feature to understand. Wikidata is implemented using Wikibase: "Wikibase is a collection of applications and libraries for creating, managing and sharing structured data" [Wikibase, 2015]. It is software in which the knowledge base is instan-

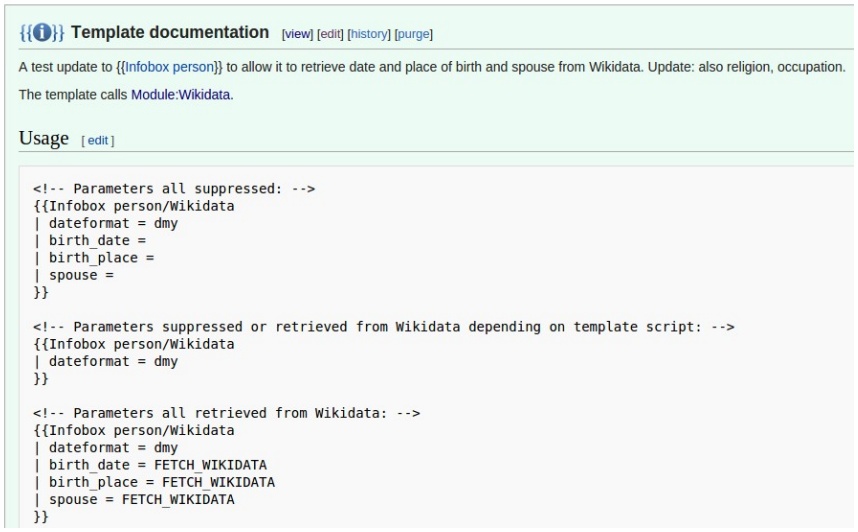

Figure 6.9: Template Infobox Person Wikidata used in Wikipedia

tiated. These technologies in interaction are the installed base on which Wikidata is built.

6.3.3 Wikidata as Infrastructure of the Semantic Web

The Wikipedia projects written in each of 294 human languages [Wikipedia, 2016] are the most visible reusers of data from Wikidata. Wikidata serves as a repository of structured data for the web. Returning to the example of infobox templates in Wikipedia, Wikidata is a central technology for the future of infoboxes. Many infobox templates are created using the workflow of editing each template individually and populating values by hand. Now that Wikidata exists, there is an additional workflow available that involves use of the programming language, Lua, to execute a script designed to find specific pieces of relevant structured data in Wikidata for display in an infobox (see Figure 6.1).

While reuse by other WMF projects is visible now, reuse by third parties requires additional research in order to determine how best to measure the phenomena. The data contained in Wikidata in addition to the structure of that data

in relation to other sets of structured data, is made available for reuse via several technologies. This allows developers of third-party applications to programmatically include structured data from Wikidata in their own systems.

Embeddedness

Embeddedness can be examined in Wikidata through the fact that the architecture of the web of data includes links involving seventeen properties in Wikidata to as direct links to other RDF data sets [Erxleben et al., 2014, 64]. The linkages between these seventeen Wikidata properties to other RDF data sets allow for conceptual interoperability between Wikidata and the other data sets. Statements about these properties that are contained in the Wikidata system are then embedded in the cloud of RDF triples that systems interacting with those other RDF data sets can access and reuse. For example, the property P686 is the ID for the GeneOntology, allowing for alignment of concepts between Wikidata and the GeneOntology. This allows Wikidata to perform some of the functions of a crosswalk that allows for interoperability across information systems [Gödert et al., 2014].

Transparency

Transparency emerges in Wikidata in greatest relief when considered by third parties that reuse structured data from Wikidata. For example, in 2012 the Google search engine began displaying infobox-like tiles of structured information related to searchers' queries [Google, 2012]. By reusing data from Wikidata and Wikipedia for display on their own web pages, Google makes the infrastructure used in the creation of this structured data less visible.

Reach or Scope

The scope of Wikidata is beyond a single site or event. Not only is data reused by WMF sister projects, but it is reused by proprietary graphs (Google's Knowledge Graph and Facebook's Social Graph among others), as well as by many other projects and systems, such as the application Histropedia [Histropedia, 2016]. In terms of domain, Wikidata is relevant across many application areas [Erxleben et al., 2014, 64].

Learned as part of membership

For example new editors of Wikidata may only be able to edit a single statement at a time in Wikidata, as they get acquainted with the community, they may then begin to use tools to support their edits and allow them to complete more edits more quickly [2] What Ribes and Bowker claim of ontologies is also the case in the Wikidata community:

> "The software tools of ontology are not encountered in everyday work with computers (i.e., word processors, browsers or internet searches) nor have they usually been part of the core training of scientists. Introducing ontologies, then, required a fairly high order conceptual discussion about the heterogeneity of languages, the categories of scientific data collection, and the structure of databases" [Ribes and Bowker, 2009, 204].

Few people beyond the Wikidata community would find themselves needing to use the tools that have been developed by the community. This is an example of how knowledge of the infrastructure of Wikidata (which requires the use of computational tools to explore) is learned as people join the community.

Links with conventions of practice

This dimension of infrastructure is at play in Wikidata in terms of how the conventions of practice about formal logic affect the strategies of knowledge representation of community members who rely on reasoning in their work processes (which could be external to Wikidata). This is due to the fact that many applications demand formalized knowledge in order to take advantage of logical inferencing and federated querying in their code. An example of this is how the formalization of OWL is related to particular strands of logic.

> "In turn, because knowledge representation has a language of its own (often called predicate logic, although in fact, it is a specific breed of

[2] For an example of how tools can be used to automate work see [Good, 2015] for the tool-building blog of Wikidata's most prolific tool author see [Manske, 2016].

> code, such as the Ontology Web Language (OWL)) the domain must learn to speak their knowledge in a language accessible to machine encoding. While the actual work of encoding is that of the ontologists, the knowledge must first be articulated in ways that can be parsed in the language of logic. Learning the practice of ontology is thus a fitting process between the epistemic conventions of the domain and the demands of formal modeling" [Ribes and Bowker, 2009].

This passage alludes to the conventions of practice about which ontologists must become knowledgeable in order to participate in the negotiations of the data model. We also see evidence for how the limitations of what formal logical modeling can express shape the potential expressivity of the knowledge base, in that only the knowledge which is structured in such a way that it conforms to the constraints of formal logic can be systematically leveraged for reuse by automated tools. Tools only have the ability to operate on that which can be represented in OWL. For perspectives discussing the limitations of OWL in terms of knowledge representation see [Krisnadhi et al., 2015, Meditskos et al., 2013b, Meditskos et al., 2013a]. The limitations of OWL do not preclude information that is not formalized according to the OWL specification from being created in Wikidata, but it is likely to shape the priorities of what work gets accomplished in what order. Due to the economic value of structured data, members of the community may choose to prioritize the needs of the biomedical domain, for example [Leonelli, 2010].

Embodiment of standards

Standards are the documentation representing the current working agreements within, between, and among communities of practice. I found evidence of many standards embodied in Wikidata: HTTP, HTML, OWL, RDF, SPARQL. Standards are technical documents that describe how technologies, in this case related to the web, work and interact with other technologies. Standards play a role in the structuring of information systems because they play the role of boundary objects for coordination of technologies.

Star and Lampland [Lampland, 2009] build on Thévenot [Thévenot, 1984] to elaborate the concept of nested standards. The notion of nested standards is characterized by recursiveness in terms of how standards interact with one another. Star and Lampland explain that standards:

- are nested inside one another
- are distributed unevenly across the sociocultural landscape
- are relative to communities of practice; that is, one person's well-fitting standard may be another's impossible nightmare
- are increasingly linked to and integrated with one another across many organizations, nations, and technical systems.
- codify, embody, or prescribe ethics and values, often with great consequences for individuals (consider standardized testing in schools, for example) [Lampland, 2009, 4].

Standards have influenced how Wikidata is designed and implemented. For example, the SPARQL endpoints that have been built for Wikidata adhere to the SPARQL standard [Wikidata, 2016a], and the RDF standard is reflected in the availability of database dumps of Wikidata data in the RDF format [Wikimedia, 2016]. This is the process by which interoperability is achieved. The importance of the ability to exchange data across the web is described:

> "Whereas Web data is indeed independently published and maintained in many sources, it is still universally accessible based on global addressing schemes and standardized protocols. More specifically, the Web emphasizes the importance of clearly specified, standardized languages that can be used to exchange data across software boundaries. Although there are some examples of earlier standardization activities

> around knowledge representation formalisms, the Semantic Web clearly has increased the practical importance of standardization in this area" [Van Hooland and Verborgh, 2014].

The role that standards play also has ethical implications. Star and Bowker explain that:

> "However, as we have already seen, the development and maintenance of standards are complex ethical and philosophical problems: ethical since decisions taken at the stage of development standards can have enormous implications for user communities; and philosophical since standards frequently deal with very basic ways of splitting up the world" [Star and Bowker, 2006, 238].

Thus, the way infrastructure encodes standards is a useful seam along which to perform close readings of infrastructure in order to discover evidence of implications of these design decisions for users.

Becomes visible upon breakdown

Wikidata is visible upon breakdown as all applications that link to Wikidata would be unable to pull data from the site were the Wikidata servers to breakdown[3]. Many systems that reuse data from Wikidata could be built to avoid this breakdown through the use of cached data. Such breakdowns would also be visible if all Wikidata SPARQL endpoints [DuCharme, 2013, 14] were to go down, as anyone using those endpoints would be unable to access data from Wikidata. Wikidata, when considered as an information system, is itself a infrastructural component of the semantic web. Star and Ruhleder note that infrastructure is a relational concept:

> "It becomes infrastructure in relation to organized practices. Within a given cultural context, the cook considers the water system a piece of

[3]To see a list of applications that are built with a workflow of reusing data from Wikidata see [Wikidata, 2016].

> working infrastructure integral to making dinner; for the city planner, it becomes a variable in a complex equation. Thus we ask, when—not what— is an infrastructure" [Star and Ruhleder, 1996].

The relationality of infrastructure is also recursive [Friedman et al., 1987, Hofstadter, 1980]. Simply put, infrastructure can be used to contain, manipulate, describe, and build other pieces of infrastructure in a system. Any analysis of infrastructure must operationalize the context of infrastructural analysis, because different vantage point will involve different views of the infrastructure. As we saw in the dimensions of infrastructure section, the dimensions of embeddedness, built on an installed base, links with conventions of practice, and embodiment of standards all point to this recursive nature of how infrastructure is built on and inside of other types of infrastructure.

Fixed in modular increments

The dimension of modularity is observed in Wikidata in the tools used to track manage and monitor changes to the technical infrastructure. The primary tool used to manage development of Wikidata infrastructure is Phabricator [phabricator, 2016]. Development tasks are coordinated via a Phabricator Workboard visible to the public [wdw, 2016]. Parts of Wikidata are repaired or extended collaboratively. Changes take time (as observed in the timestamps on the workboard tasks), and are negotiated (as observed in the discussion threads that are attached to each workboard task).

Conclusion

One challenge of performing a close reading of infrastructure is that it can never be complete. The examples I provide are not the only examples of these dimensions, there are many other examples that could be discussed. Due to the size and complexity of the infrastructure of WMF projects, the work of many close readers is needed in order to further elucidate the conceptual infrastructure and how that conceptual infrastructure is encoded in the technical infrastructure of WMF projects. I analyzed three structures of information organization in WMF

projects and discussed dimensions of infrastructure I observed.

Chapter 7

THE IMPORTANCE OF INSPECTABLE INFRASTRUCTURE

Performing infrastructural analysis of Wikimedia Foundation (WMF) projects allows us to identify where and how conceptual structures are encoded in technical infrastructure. When we read the code that makes up the technical infrastructure, we can see evidence of the choices that have been made about how to formalize the conceptual systems used to organize information. We can then consider implications for what types of knowledge can be expressed and what types cannot yet be expressed through these infrastructures.

I investigate the dimensions of infrastructure outlined by Star et al. [Star and Ruhleder, 1996] as they are observed in WMF projects. The dimensions of infrastructure are: embeddedness, transparency, reach or scope, learned as part of membership, links with conventions of practice, embodiment of standards, built on an installed base, becomes visible upon breakdown, and fixed in modular increments [Star, 2010, Star and Ruhleder, 1996, Star, 1999]. I identify three structures of information organization across WMF projects: the category system; the information boxes; and the knowledge base for structured data. Through my analysis of sites where these dimensions of infrastructure can be observed in WMF projects, I demonstrate an analytical approach that makes infrastructure more legible.

While it is certainly challenging to inspect the infrastructure of WMF projects, it is more possible in WMF projects than at many other sites of knowledge structuring. There are many knowledge graphs that are not openly available: Google's Knowledge Graph, and Facebook's Social Graph are among the corporately-owned closed systems [Färber et al., 2015]. Knowledge graphs that are designated proprietary corporate infrastructure can only be inspected at the discretion of the corporation. This means that the choices about how information is represented

and how to data is structured are not inspectable by the public. The fact that these choices are not inspectable means that users of these systems do not have the information necessary to determine where systemic biases may be encoded into these systems. For example, how will we pose decolonizing critiques of the information organization of these systems such as those advocated for by [Duarte and Belarde-Lewis, 2015, 681]? Duarte and Belarde-Lewis illustrate examples of how systems of classification and systems of library cataloging have represented indigenous knowledge in ways that that causes harm to indigenous people. They critique how this neo-colonial oppression is enacted on indigenous communities because of the fact the creators of these classification systems fail to hold themselves responsible for the implications of their decisions about how information is organized.

Both the infrastructure of and the structured data contained within the commons-based peer production system, Wikidata, are free and open and, thus, inspectable. Not only can the public read the content, the public has access to the code that governs how that content is organized and presented. All of the code that makes up the software of MediaWiki, Semantic MediaWiki, and Wikibase are also published under open licenses ensuring that the code is open for inspection. The fact that the content stored within WMF projects, as well as the software that is used to enable them, are governed by open licenses distinguishes Wikidata as a very unusual knowledge base. Wikidata is unusual in that it is open for inspection, and both the content and the software are freely available for reuse. As Lessig explains, the architecture of a computational system is also the politics of the computational system [Lessig, 2006]. While many for-profit companies seek to make both the content and the enabling software of a knowledge base proprietary and private, the WMF knowledge base is created by a commons-based peer production community, and the content created within WMF projects as well as the technical infrastructure which enables these projects are designated as intellectual commons of humanity. The fact that Wikidata is inspectable, is in harmony with calls for support of user-owned and commons-based infrastructure as coun-

terbalances to increasing centralization of technical control by corporations and state governments [Benkler, 2016]. Simply put, the way that information is organized and data is structured in WMF projects is publicly inspectable. The ability to inspect the infrastructure of WMF projects is a powerful location where we can observe knowledge production and knowledge structuring. It is powerful because it is the collaborative work of contributors to a peer production system. Thus this structuring happens outside of the control of any single corporate interest or any state government. While the public does not have the opportunity to inspect information organization or data structuring within private knowledge graphs, the public does have the opportunity to inspect information organization and data structuring within WMF projects.

Wikidata is a knowledge base of structured data governed by an open license. Many application developers are choosing to reuse structured data from the Wikidata knowledge base rather than create their own knowledge base. The ability of the public to inspect the infrastructure of Wikidata is a unique opportunity that allows us to create spaces where we can make ourselves aware of how information is being organized and data is being structured that is beyond the control of corporations and state governments.

In this sense, the Wikidata project is an important event in the history of information structuring in WMF projects. The process of information structuring is the route through which knowledge organization systems (KOS), which I refer to as conceptual infrastructure, are encoded into technical infrastructure. I analyze the category system and the infobox templates as structures used for the organization of information. While the category system and the templates for information boxes can encode some features of KOS in them (associative relationships, class-subclass relationships, etc.), in order to analyze Wikidata we must consider additional layers of complexity and representation. In Wikidata ontologies can be created to describe other ontologies, thus entire KOS can be represented in Wikidata, and can exist alongside other KOS also represented in Wikidata. Not only is Wikidata an information system comprised of multiple technologies used in interaction, it

also has its own infrastructure and serves as infrastructure. Wikidata is a knowledge base that serves an infrastructural role in the semantic web. The purpose of the semantic web is to reduce barriers to information sharing among information systems. The semantic web is a conceptualization of a possible future of the world wide web. This vision, only parts of which have been realized, are comprised of design principles, as well as a large number of technologies that have been engineered to create, manipulate, store, and search the structured data that is then shared via the semantic web. The creation and curation of structured data is the purpose around which the Wikidata community grew.

An example of of how Wikidata plays a role in information structuring that is relevant at the scale of the semantic web is the story of how it became possible to query Wikidata for information related to pharmaceutical drug interactions. The structured data about drug interactions stored in Wikidata is result of the work of at least two independent teams, one of which created a bot to add identifiers for biomedical entities to Wikidata, and another group that entered data about drug interactions to Wikidata. The data from both groups is leveraged to answer certain questions relevant to the intersection of these two graphs [Mitraka et al., 2015]. Simply put, the infrastructure creates synergistic effects of the distributed work that the independent teams performed by allowing more complex queries to be answered by the intersections of the two graphs than by either graph on its own.

The fact that the Wikidata system uses software that is created within the Free-Libre Open Source Software (FLOSS) community and is governed by an open license, means that not only the content of Wikidata is free for reuse and inspection, the software that enables it is freely available as well [Vrandecic, 2013]. The inspectability of the infrastructure of the systems that enable collaboration was called for in the early 1990's by some members of the social computing community [Schmidt and Bannon, 1992, 26]. The commons-based peer production systems of the WMF projects are an example of what Schmidt and Bannon describe as an effective computational infrastructure for collaboration, in that both the in-

frastructure of the projects as well as the work products created therein are open to editing by the community, and the actions the community takes are fully inspectable for the population of interested people who have the technical literacy to read the encodings of the software used to create the infrastructure, and technical documents describing these technologies.

One additional level of complexity of this challenge is that each of the pieces of software in this software stack (MediaWiki, Semantic MediaWiki, Wikibase, etc.) are themselves being created in the context of their own FLOSS project communities. The members of these FLOSS communities are actively creating this unique type of computing infrastructure through their collaborative development efforts. Thus, to a certain extent, the only people who can describe these technologies are their creators. These situations occur when contributing developers write code to extend functionality of some aspect of the system. Until someone takes time to review, comprehend, and document the code, the developer may be the only person who understands it. With many contributing developers working in a distributed context, there are many parts of the system about which only a subset of people have a full technical understanding. Alongside the development of the technical systems, we have the expansion and elaboration of techniques for communicating about these systems through documentation. Schmidt and Bannon prescribed that:

> "The system should support the documentation and communication of decisions to adapt, circumvent, execute, modify etc. the underlying model" [Schmidt and Bannon, 1992, 26].

Thus the infrastructure is described and documented by the community itself. Approaches to documentation in open source projects are in flux and are changing rapidly as commons-based peer production systems change and grow [Gentle, 2012, McCanse, 2012, Foundation, 2016].

The fact that documentation is being created by this commons-based peer production community means that members of the community self-select the task of

creating documentation. In enterprise software engineering development contexts employees are assigned the task of creating documentation and may be held accountable for the timeliness of its creation. In the Wikidata community members hold themselves accountable for creating documentation, and it sometimes does not get created in a timely fashion.

> "If the structure of the data is not well documented and intuitive enough to allow researchers to use it without first researching the structure itself, then the initial energy that is required to use Wikidata is likely too high. This constitutes the major problem that we currently see with Wikidata, as the hierarchical relations between entities in the knowledge base are still evolving" [Spitz et al., 2016, 2].

This assessment of the importance of clear, accessible documentation of the conceptual infrastructure of the knowledge base of structured data is easier to recognize than to solve.

Organizing systems involve choices and decisions about how relationships are described between entities. With Wikidata poised to be a major infrastructural component of structured data for the semantic web, the structures used for organizing information in Wikidata are the primary site where audits of the data structuring can be performed and evaluated.

7.1 Who Can Read Infrastructure Now

In the process of recording field notes as I researched the category system, the infobox templates, and the knowledge base of structured data, I used a range of approaches to attempt to understand how these structures were created. In order to perform a close reading of infrastructure in WMF projects I had to gain literacy in understanding computational systems and technical documents. I mention this to provide information about how I came to understand the technical skills that are required as background knowledge in order to unpack the question of which elements of KOS or entire KOS themselves may be encoded in infrastructure.

In short, before we can decode infrastructure, we need to know the elements, structures, and syntaxes of the code itself.

> "In information infrastructure, every conceivable form of variation in practice, culture, and norm is inscribed at the deepest levels of design. Some are malleable, changeable, and programmable-if you have the knowledge, time, and other resources to do so. Others–such as a fixed-choice category set–present barriers to users that may only be changed by a full-scale social movement" [Star, 1999, 389].

In the context of the spectrum of designs that are malleable to those that are fixed, Wikidata is a malleable design. It is entirely editable, and the full stack of software that Wikidata is built on is also openly editable. If we turn to see what percentage of the population has the knowledge, time and other resources to edit Wikidata, we discover that the percentage is quite small [1]. This means that anyone who is concerned about how information is represented and data is structured must have the technical skills to inspect the infrastructure or access to open dialogue with mediators who are willing and available to translate.

The number of people who have this type of technical literacy is a very small subset of the population. Bowker et al. call for the creation of a set of infrastructural professionals who would have a mandate to hold themselves accountable to a commitment to thinking about the question of who is marginalized by standards [Bowker et al., 2010, 107].

What this means in concrete terms is that people must have XML literacy [Boiko, 2005] in order to read RDF and to comprehend from these encodings what conceptual structures are being used to organize information and structure data. The call for infrastructural professionals made by Bowker et al. can also be understood as an identification of an opportunity for the communities of library science and information science. Scholars in the fields of library and information science

[1] For a discussion of editorship of a another knowledge base cf. [Kochhar et al., 2010]

have produced bodies of research documenting how conceptual structures can marginalize people, can introduce bias, or exclude entire systems of knowledge [Olson, 2013, Duarte and Belarde-Lewis, 2015, Education, , Feinberg, 2007, Bowker and Star, 1999, Bowker, 2005, Zachry, 2008]. Scholars from library and information science who also have XML literacy are well-positioned to answer Bowker's call. If educational programs in library and information science continue to expand opportunities for students to become technically literate in XML, RDF, and other technologies of the semantic web, information professionals of the future will be well-prepared to fulfill Bowker's call for infrastructural professionals.

A knowledge base, such as Wikidata, organizes so much information that it is unlikely that individual humans may ever be able to understand all of the ways structures are in interaction with one another. Most often we interact with subsets of information from a knowledge base, or by querying it in some fashion. Querying a knowledge base is asking a question of the knowledge base. Both the structure of the knowledge base and the structure of the query language influence which queries will be possible and which are not possible, and what kind of answers are returned to users.

The organizational structure of the knowledge base of structured data, Wikidata, impacts all of the information systems and applications that interact with it. Due to the potential scale of reuse, the expressivity of these information structures require focused critical attention from people who can understand and evaluate the impacts. If we follow the conceptualizations of the implications of artificial intelligence advanced by [Lanier, 2014], [Shroff, 2013], and [Bostrom, 2014], then this knowledge base is a likely to play a influential part of machine-based artificial intelligences because of the structure between, around, and among entities in these complex assemblages of ontologies, databases, and algorithms. The inferencing capabilities that power machine-based artificial intelligence are direct products of these structures, since algorithms must be written to leverage extant data structures. Thus systems of artificial intelligence that use structured data from a knowledge base and the the structure of that knowledge base are intimately

related. The research I describe in previous chapters elucidates locations where structures of information organization are observed in the commons-based peer production communities of the WMF projects. I argue that it is vital that we constantly interrogate the design decisions that go into these systems and engage diverse standpoints in the evaluation of these systems and their roles as co-creators of the future of collective intelligence and knowledge representation.

Bowker concludes his book *Memory Practices in the Sciences* saying:

> "I have argued for a deeper consideration of the role of our memory practices as the site where ideology and knowledge fuse.... We need to hold the past open so that we do not hypostasize and freeze the present, and by extension limit our own future" [Bowker, 2005, 228].

In this passage Bowker draws and explicit connection between our memory systems and our ability to represent knowledge. Bowker would like to see humans emphasize multiple forms of knowledge representation. Inspired by Bowker's work, I see the knowledge base of structured data as a particularly important genre of memory system for us to audit. Auditing the knowledge base for what types of knowledge it can enable the expression of, and reading the silences of the knowledge it cannot yet express is crucial for evaluating the structured data contained within. The knowledge base of structured data maintained by the Wikidata community is unusual in that it is possible to inspect its infrastructure. The structured data that is reused from this knowledge base will be consumed by users in many contexts beyond Wikidata.

7.2 Illuminating Queries

In addition to the fact that there are a limited pool of individuals who have the technical literacy to inspect infrastructure, I identify another factor that contributes to our current situation of inattention to how conceptual systems are encoded in infrastructure. The lack of tools to support the work of inspecting infrastructure is a barrier to our ability to hold ourselves accountable for the implications of our

decisions in this area. Bowker and collaborators remind us that "The design of infrastructure itself can make its effects more or less visible" [Bowker et al., 2010]. If we want to have tools to inspect infrastructure of Wikidata, members of the Wikidata community will need to build these tools.

Well-designed tools for inspection can reduce the amount of work that it takes to monitor infrastructure. The close reading approach that I describe in Chapter 6 is not practical as an approach for the day-to-day monitoring of Wikidata infrastructure. My argument that additional tools are needed to help make the conceptual infrastructure of the knowledge base of structured data echos that of Spitz et al. who note that these tools are necessary to the growth, acceptance and reuse of data from Wikidata [Spitz et al., 2016].

An example of the type of tools that I envision for the work of monitoring infrastructure in order to examine how conceptual systems of information organization are encoded in infrastructure is a battery of SPARQL queries that could be sent to the Wikidata SPARQL endpoint. In the field of software engineering there is a process for determining the integrity of a program called regression testing [Leung and White, 1989]. Once a bug has been discovered in a piece of software, software engineers use the information they gained through understanding that bug (and the solution used to eliminate it) to create regression tests that can be run against the code for the software to ensure that the bug does not get recreated or reintroduced as the code base for the software is extended. The battery of SPARQL queries I am describing as a type of tool that could be created to support the work of infrastructural inspection is analogous to the concept of regression testing in software engineering.

The battery of SPARQL queries I am imagining can be thought of as "Illuminating Queries" in that they will make infrastructure more visible, especially those areas of infrastructure that allow us to inspect how conceptual systems are encoded into technical infrastructure. These queries could be run against the certain subsets of the data in Wikidata, against schema stored in Wikidata, or against the data model of Wikidata as a whole to identify conflicts, data gaps, or structural

limitations within the database. This is one example of how the set of infrastructural professionals [Bowker et al., 2010] describe might use tools to inspect infrastructure.

Conclusion

Star recounts how she became attuned to the opportunities of paying attention to what others take for granted by following Anslem Strauss' urging to "study the unstudied" [Star, 1999]. While many semantic web researchers are interested in knowledge bases, very few are researching from perspectives that center ethical considerations. The field of ontology evaluation is primarily concerned with understanding the structures that have been instantiated.

I emphasize the importance of holding the Wikidata community accountable for the implications of expressivity of knowledge representation in the knowledge base of structured data. I chose this approach because of the overwhelming economic pressure of capitalist gain via salvage accumulation that critiques of how this knowledge base is being created are unlikely to emerge from the neoliberal academy [Cotera, 2010].

The technical literacy required to read infrastructure in order to understand how a commons-based peer production system functions has been mastered by a subset of the current human population. It is imperative that we require additional strategies to communicate how knowledge is represented in technical systems to audiences beyond the subset of the population that has already achieved the necessary level of technical literacy to inspect the infrastructure personally.

Chapter 8

CONCLUSION

When we search for information on the web, we often evaluate this information for reliability, accuracy, and timeliness. When we ask questions on the web, we are sometimes answered by people, and sometimes the response to our query is generated by machines that have been programmed by people to execute commands according to some set of instructions. As we have shifted from web 1.0, to web 2.0, and beyond, the trend toward automated question answering on the web has grown.

Structured data is one of the raw materials necessary for systems of automated question-answering, collective intelligence, and artificial intelligence. Many of the largest technology corporations have begun to collect structured data in knowledge bases. Intelligent agents such as Apple's Siri, Amazon's Echo, or Microsoft's Cortana, use knowledge bases to provide answers to their users' queries [Bissig, 2015]. These knowledge bases are used to supply facts and information to users. As each of these companies compete to provide their unique products and information services to larger and larger segments of consumers, more people will be exposed to information from these proprietary sources.

Many corporations are building their own infrastructures for these knowledge bases. They are populating their knowledge bases through the reuse of existing corpora of structured data, like those drawn from WMF projects. Across WMF projects, structures of information organization are encoded in infrastructure. Structures of information organization encode things such as how concepts are related to one another through the use of class-subclass relationships.

It has been argued that the creations of the WMF community, such as English Wikipedia, influence "what people all over the world know, believe, and think"

[Jemielniak, 2014, 4]. If the conceptual schemes for information organization are where decisions about how knowledge can and cannot be represented are established, and if the structures of information organization that encode these decisions are only inspectable to a small group of people, there is risk. The risk I describe in this research is the potentially hazardous situation of a small group of people having the ability to determine, revise, and extend strategies for knowledge representation that will influence what many other people know, believe, and think. The reuse of structured data from Wikidata means that the decisions of the community not only have implications for users of Wikidata, but also have implications for all systems that reuse structured data from Wikidata.

8.1 Salvage Accumulation

Anna Tsing introduces the concept of *salvage accumulation*, the process by which corporations incorporate goods produced outside of capitalism into capitalist supply chains.

> "In capitalist farms, living things made within ecological processes are coopted for the concentration of wealth. This is what I call 'salvage', that is, taking advantage of value produced without capitalist control. Many capitalist raw materials (consider coal and oil) came into existence long before capitalism. Capitalists also cannot produce human life, the prerequisite of labor. 'Salvage accumulation' is the process through which firms amass capital without controlling the conditions under which commodities are produced. Salvage is not an ornament on ordinary capitalist processes; it is a feature of how capitalism works" [Tsing, 2015, 62].

We see evidence of salvage accumulation in the story of structured data available on the web. Through the creation of infoboxes, as the Wikimedia community structured an increasingly large set of structured facts in Wikipedias, capitalist

firms such as Google, Facebook, Yahoo and many others identified an opportunity for salvage accumulation of structured data. Due to the terms of the licence governing the content of Wikipedia, content is free to be reused in external applications. This allowed capitalist firms to turn freely-available structured data, computational systems to manage structured data, and applications making use of this data into commodities with economic value [Dong et al., 2014b, Steiner et al., 2012]. Google's Knowledge Graph is legally allowed to display a infobox of Google's design populated with structured data from Wikimedia projects. If users find answers in Google's display of structured data, this encourages users to remain on the Google search results page rather than click through to any of the other websites described in the results. By encouraging users to remain on Google pages, Google is able to display more advertisements for longer periods of time to those users, and thus able to charge companies paying for those advertisements more money.

This is an example of salvage accumulation because the structured data used in Google's Knowlege Graph was curated in the context of WMF project communities, but is displayed in an entirely new context by Google. Google uses the raw material (machine-readable structured data) to create their own product (Google's Knowledge Graph) which they can monetize (through selling ads to be displayed to users interacting with the Knowledge Graph).

In Figure 8.1, we see a screenshot of data from Google's Knowledge Graph as displayed on a search results page. A user has searched for the word 'nautilus,' and has been presented with a set of web links and a rectangular tile of quick facts (reminiscent of Wikipedia's information box for nautilus as seen in Chapter 1, Figure 1.3). If a user can find the information they are looking for on Google's search results page without having to consult any of the web results then Google has additional opportunities to display advertisements to the user, resulting in increased revenue for Alphabet, the holding company that owns Google [Metz, 2015]. This is an example of salvage accumulation in the context of structured data and the web.

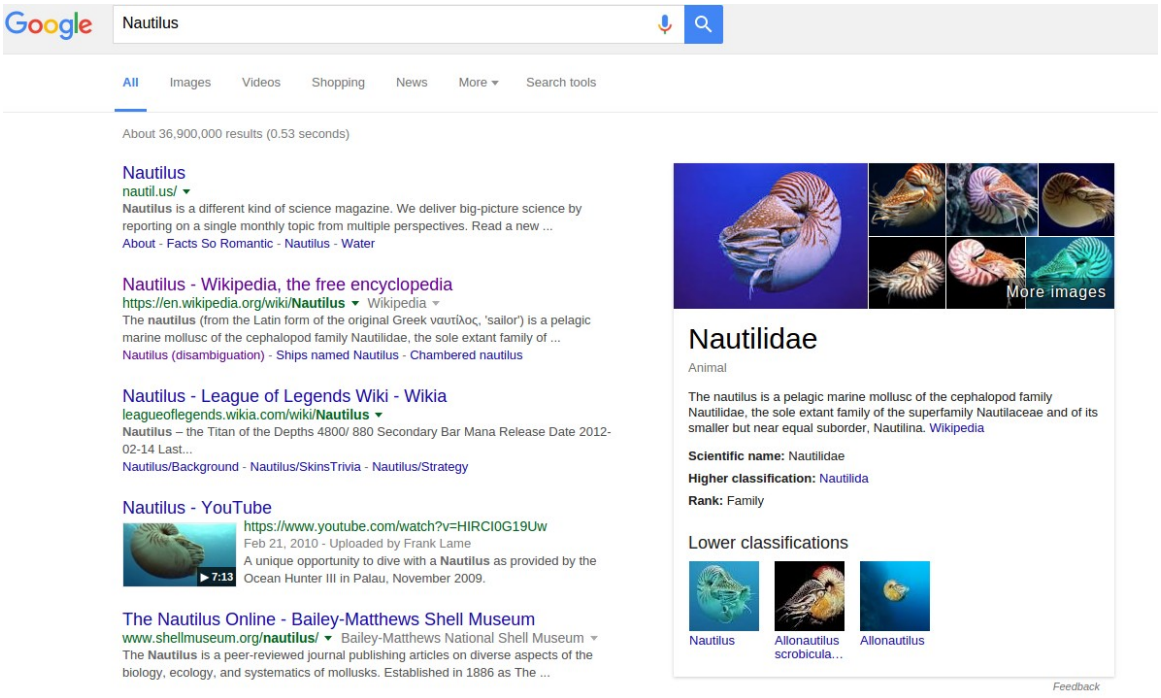

Figure 8.1: Google search results for 'Nautilus'

The full story of how salvage accumulation opperates in the context of structured data and the web is yet to fully unfold. In 2013 McKinsey and Company estimated that the value of structured data was between three to five trillion dollars of potential revenue [McKinsey and Company, 2013]. The infrastructure that is necessary for this creation of value is the infrastructure of the semantic web, and the commons-based peer-production systems, such as Wikidata, which create that infrastructure. Case studies of the value of open structured data can be found throughout the business literature [Open Knowledge, 2016]. Within capitalist economies, the financial incentives to monetize some aspect of structured data creation, management, or reuse are impressive because of the fact that structured data can be reused at no cost, but the user's interaction with the product can be monetized.

Tsing describes how neoliberal capitalism identifies locations for salvage accumulation. What will the the movements of structured data through neoliberal capitalist economies look like? Cotera defines neoliberalism saying:

> "Often referred to as the "privatization of everything," neoliberalism promotes individualism as the primary locus of freedom, and deploys this ideological argument to transfer ownership of public goods from the community to the private sector with the ultimate aim, as David Harvey has argued, of shifting the balance of capital accumulation to economic elites and the ruling class [Cotera, 2010, 331].

The gap between the haves and have-nots also comes into play in terms of the knowledge base. Hansson describes the logic by which value is created among documents in the online context:

> "In the late capitalism we live in today, economic value of documents in a simple Google search govern the pattern of retrieved documents. We see a new Matthew effect—the documents most likely found in a specific search pattern are the ones that give economic revenue, which in turn make them the ones most likely to be retrieved" [Hansson, 2013, 390].

Hansson is referring to how the Matthew effect, the phenomenon of preferential attachment to extant nodes in a network [Barabási and Albert, 1999], is observed when using algorithms of information retrieval. The implications of the Matthew effect are that the economic forces of our globalized capital flows shape how information flows through computational networks on the web. As Barbara Smith reminds us, capitalism is designed to create economic inequality, and that social injustice is a precondition for the growth of economic inequality [Smith, 1998, 169]. If left unexamined, will the human decisions related to information organization and data structuring in knowledge bases such as Wikidata mimic current patterns of globalized capital flows?

Commons-based peer production communities, like the Wikidata community, provide certain alternatives to some capitalist modes of production. Benkler argued that there were both allocation gains (community members self-identifying work they wanted to do), as well as information gains (more information, higher-quality information) for peer production communities over traditional firms.

> "Peer production has an advantage over firms and markets because it allows larger groups of individuals to scour larger groups of resources in search of materials, projects, collaborations, and combinations than is possible for firms or individuals who function in markets. Transaction costs associated with property and contract limit the access of people to each other, to resources, and to projects when production is organized on a market or firm model, but not when it is organized on a peer production model" [Benkler, 2002, 376].

Applying this line of thinking to the context of the WMF projects I have described here, the characteristics of the commons-based peer production phenomena seem to be a crucial element in this story. This story is the history of the creation of the largest set of human-curated structured data, and how this structured data—alongside the infrastructure built by and used by the WMF community to create it—are now being incorporated into capitalist economies through the process of salvage accumulation.

8.2 Limitations

The research I describe in this dissertation has several limitations. I have divided the limitations into three types: complexity, mutability, and technical literacy. I will explain each of these limitations in order.

8.2.1 Complexity

Infrastructure is complex in any computational system. When multiple computational systems are in interaction (for example, the relationship between Wikipedia

and Wikidata) the interaction of these infrastructures increases their complexity. WMF projects are multiple and complex. In this research, I considered only three structures of information organization. There are many additional structures of information organization in WMF projects that I did not consider. Infrastructure is inherently difficult to study because it operates in the background of a system. Bowker and Star discuss the challenge of researching infrastructure saying:

> "...infrastructure is not absolute, but relative to working conditions. It never stands apart from the people who design, maintain and use it. Its designers try to make it as invisible as possible, while leaving pointers to make it visible when it needs to be repaired or remapped. It is tricky to study for this reason" [Star and Bowker, 2006, 230].

This quote captures the limitation of my research I name 'complexity.' In order to understand infrastructure I had to pay close attention to the social interactions of the communities creating infrastructure. This increased the complexity of data collection and data analysis.

8.2.2 Mutability

The Wikimedia Foundation (WMF) projects I describe in this research are active communities of commons-based peer production. Hundreds of thousands of edits are being made to these projects each month, and this research took place over the course of several years. Spme of these edits are to the infrastructure of these projects, the very object of my research. The mutability inherent in the object of research is a limitation of theses projects. Selecting WMF projects as the site of my research involved adjusting my research activities to the context of a rapidly changing commons-based peer production system. A limitation of this work is that, upon publication, the technical descriptions of infrastructure may fall behind development.

8.2.3 Technical Literacy

Technical literacy is a precondition for infrastructural inspection. One of the ways that I came to understand that technical literacy is necessary to explore these structures is through my lived experience of posing questions about the conceptual structures of information organization utilized in WMF projects, and discovering that I would need to expand my technical literacy in order to investigate the answers to my questions. A limitation of this work is that I have mastered a subset of technical competencies, and there remain additional technologies to learn. It is possible that a researcher with a different set of technical skills would observe and understand and analyze these examples differently.

8.3 Contributions

Each of the three pieces of research I describe in this dissertation generate findings. I introduce findings from each project in order: the category system; the templates for information boxes; and the close reading of infrastructure.

8.3.1 Category System

My analysis of the category system produced a description of how members of the Wikipedia community collaboratively created this novel structure of information organization. I describe how collaboration around the creation of the category system, a structure of information organization, took place in a commons-based peer production system. This is the first study of the design decisions and design goals for the category system of English Wikipedia.

I Identified four themes related to collaboration around the category system that were assumed by the participants:

- Collaboration with the category system -
  This theme describes discussion in which editors were conceptualizing the category system as an entity that facilitates navigation and or retrieval, clustering or conceptual visualization.

- Collaboration over the category system -
  This theme describes discussions in which editors were conceptualizing the category system as an object of work that individuals must use and manipulate. These include be debates about how categories should be applied, what types of relationships should exist and be made explicit between categories.

- Collaboration through the category system -
  This theme describes discussion in which the editors were conceptualizing the category system as a mechanism for communicating and interacting with others.

- Tools for category collaborations -
  This theme describes discussions in which editors discuss tools they are using to facilitate collaboration with, over and through the category system.

This study provides additional context for understanding of how infrastructure is collaboratively created. The details of the infrastructure of the category system also inform the close reading of infrastructure research.

8.3.2 Infobox Templates

I discovered patterns in editing of the templates for infoboxes through the creation of three original metrics: influence, diversity and longevity. Templates in Wikipedia are an interesting type of information structure. They allow work to be reused, thus saving editors time and energy. They are also powerful tools for organizing and structuring information. The act of contributing to the creation of an infobox template is an act of information organization because it creates a structure to relate pieces of information in context. I discuss the patterns of template editorship within the context of theorizations of hidden work, and I demonstrate how to document this hidden work.

Templates are part of the technical infrastructure of WMF projects. Analysing them in the context of theorizations of hidden work allows me to unpack the layers

of work, especially highly-influential work, that enable WMF projects.

The purpose of creating metrics to help us to understand the work of creating templates for infoboxes in Wikipedia is that this work is currently hidden. Of the four types of hidden work outlined [Nardi and Engeström, 1999]. the hidden work that goes into the creation of templates in Wikipedia best matches the definition: "work defined as routine or manual that actually requires considerable problem solving and knowledge" [Nardi and Engeström, 1999]. The results I present suggest that it is also important to consider hidden work that is ongoing, and that looking for such work purposively could allow it to be accounted for, rewarded, leveraged, enhanced or supported before the point at which it is replaced by new work processes. I extend the theories of hidden work to an additional category of hidden work—one that describes work that is repurposed in contexts that entirely obscure the creators once it is discovered to be extraordinarily valuable, but is under-recognized when the work is being done due to the circumstances in which is it performed.

8.3.3 Close Reading of Infrastructure

I introduce three structures of information organization from across Wikimedia projects: the category system; the templates for information boxes; and the knowledge base of structured data. This is a descriptive contribution that documents specific aspects in the information organization processes and behaviors taking place in the socio-technical system [Niederer and Van Dijck, 2010] of WMF projects. I demonstrate how to investigate the infrastructure of a commons-base peer production system in order to discover how conceptual structures are encoded into technical infrastructure.

I demonstrated the need for inspectability of these systems in order to audit expressive constraints. I identified the need for non-technical human-readable documentation of the infrastructure of any project that wishes to build trust with users through transparency of communication of design decisions related to the

representation of knowledge in a technical system.

I theorize that capitalist ventures like Google have identified Wikidata as a site of salvage accumulation [Tsing, 2015]. I argue that the technical details of the infrastructure of Wikidata are likely to have relevance to our understandings of the proprietary knowledge bases. Due to the fact that we will not be able to inspect proprietary knowledge bases, those that are open to inspection will serve to help us to form educated guesses about what is going on in the closed systems.

I demonstrate that infrastructure is inspectable, and that only a small number of people who have had the opportunity to develop technical literacy are in a position to undertake such inspection. We must require of the communities that create these resources to consider issues of intellectual and social responsibility. Issues of responsibility come along with providing an infrastructural resource that will work in the background, nearly invisibly. If we do not constantly work to read it, make it legible, find ways to illuminate it, find ways to communicate about it, and find ways to remind ourselves that infrastructure is a powerful force in shaping our realities, then we face the risk of uncritically accepting views of the world that system designers and software developers display to us as we use their computational systems.

I demonstrate that ontologies serve as boundary infrastructure [Star, 2010] in the knowledge base of structured data. I argue that our shared understanding of what an 'ontology' is can be clarified by disambiguating 'ontology' used to refer to conceptual infrastructure and 'ontology' used to refer to technical infrastructure. Although these can not be separated our in computational systems, when different communities of practice collaborate around this boundary infrastructure, this additional precision in the use of 'ontology' is necessary.

Considering all three projects in relation to one another provides context for another finding related to capability building in WMF communities. WMF communities learned many lessons about information organization through the creation of the category system. The templates for information boxes were designed and implemented by the community with the lessons of the category system in

mind. The knowledge base of structured data, Wikidata could not have been built at the time the category system was created. The community had to integrate the lessons of creating the category system and the templates for information boxes, as well as the lessons learned from observing how structured data was mined out from the information boxes in order to design and implement Wikidata. Thus, when considered together, these three projects contribute to our understanding of how WMF project communities apply lessons from their experience to extend and further develop infrastructure.

8.4 Future Work

I have formulated additional questions through the course of this research. These questions suggest potential directions of future work. I group the future work I envision for this topic into three sections: illuminating conceptual infrastructure; improving documentation; and responding to ethical accountability. I will discuss each of these sections in order.

8.4.1 Illuminating Conceptual Infrastructure

In Chapter 1 I introduced Figure 1.5 to demonstrate how knowledge organization systems (KOS) are the conceptual infrastructure of computational systems. The fact that human decisions about how to encode conceptual infrastructure into technical infrastructure are not emphasized in many discussions of infrastructure is one of the signs of the extent of our need for tools to make conceptual infrastructure more visible and more legible as we design tools to support the inspection of infrastructure. How will we design solutions for the task of increasing the inspectability of infrastructure? How will we highlight conceptual infrastructure as we create and use these tools to help us inspect infrastructure?In future work I will develop a set of specialized SPARQL queries designed to highlight limitations of our current strategies for knowledge representation in knowledge bases like Wikidata.

8.4.2 Improving Documentation

One barrier to inspection of the infrastructure of computational systems like Wikidata is the limited documentation currently available. How will we present documentation of these technologies in such a way that a broad range of people can participate in the creation and evaluation of such knowledge bases of structured data? The Wikidata community is aware of this challenge, but it is unclear how the community will respond to improve documentation. How will we create documentation that can be understood by people at various stages of technological literacy?

8.4.3 Responding to Ethical Accountability

We have not yet seen enough research into the ethical implications of a small community of people organizing information and structuring data that will later be reused by a broad range of applications to provide information to broad audiences on the web. How will we communicate these issues to users who may be interacting with structured data the context of third-party applications reusing structured data from Wikidata? How will we communicate infrastructure through documentation that is accessible people beyond those who can already read code? How will we provide mechanisms for users to take action to address biases in the information stored in knowledge bases of structured data? How can we communicate the systems of values claimed by organizations in positions of control of infrastructural components? What strategies can we use to raise awareness of where information organization and data structuration is happening and who is performing it in knowledge bases of structured data? How can we raise awareness of the potential risks I describe relating to the potential for private companies to seek to protect their ability to present slices of information that advance their corporate objectives?

The transparency that has been designed into the community culture and technical infrastructure of this commons-base peer production system is a powerful

venue for inspecting Wikidata, a hub of infrastructure for the web, the semantic web, and beyond. How can we draw attention to the potential for this venue to become a space where we interrogate the implications of how we choose to organize information and structure data? How can we do this in a way that involves stakeholders from communities that are not currently represented among the technical developers of these computational systems?

How will we address concerns of communities that are illegible to a system of categorization, communities that Star terms 'residuals' [Zachry, 2008].

The framework I am using for justice is based on that of [Reardon et al., 2015]. Inspired by conceptualizations of ethics from the field of feminist science studies (FSS), this way to engage ethics relies on the concept of *response-ability*.

> "More than a clever play on words, response-ability, Donna Haraway argues, is not about aligning one's actions with a set of universal ethical principles. Instead, it requires cultivating practices of response. These practices are developed and done with others, both human and non-human, in a process of ongoing exchange" [Reardon et al., 2015, 12].

This way of engaging ethics is appropriate for a study of differential reading of the infrastructure of commons based peer-production systems for the structures of information used to represent knowledge. As we design, implement, revise, and elaborate a knowledge base in an information system we are in relation with it, the knowledge we are attempting to represent and we are in relation with many other actors. There is mutual articulation in interactions throughout the community creating this knowledge base.

The importance I place on finding strategies that could be used to consider ethical implications of the design of this knowledge base is inspired by the recommendations of the Pluralizing the Archival Curriculum Group [Education and Institute, ired]. I chose this as a source of information about possible strategies because of the work that has already been done to articulate the implications of information organization and data structuring for indigenous communities today.

As people have encouraged the archival tradition to reexamine practices related to indigenous knowledge, they have shared their findings and recommendations.

> "Recent events, developments, and research findings highlight areas that need to be addressed and the often stark differences in worldviews and power structures relating to information and knowledge management and memory-keeping between Indigenous and non-Indigenous communities" [Education and Institute, ired, 76].

What this group is calling for is a form of engaged scholarship. Due to the economic reality described above as salvage accumulation, it is unlikely that this critique will be posed from within organizations that earn financial profit in providing innovative applications that reuse structured data from Wikidata. I find it more likely that the communities of people who will ask questions about the ethical implications of the design of this new type of memory practice will be motivated to hold the communities responsible for the creation this infrastructure for reasons other than financial gain.

This research documents a situation that has been identified as an opportunity for the salvage accumulation of capital, and thus will impact our economies, information systems, and human lives. As the details of how to model and implement knowledge representations are created by a technological elite, it is difficult to imagine how ethical responsibility can be enacted. An epistemological stance that involves mobilizing marginalized perspectives at the same time as, and alongside, dominant perspectives could allow for a self-reflexive criticality. This type of critical analysis could identify likely points of rupture and dissonance for those who are marginalized by the dominant paradigms. By getting of sense of where these ruptures are likely to occur, we can make more well-informed decisions about how to engage with these technologies, and how we hold them accountable for what their infrastructure makes possible and what their infrastructure makes impossible.

8.5 Hermeneutic of Hope

I began this text by reviewing the urgent claims that Bostrom [Bostrom, 2014] makes about our responsibility for the choices we make about superintelligences to convey the stakes involved in the decisions we make about how to represent knowledge. Artificial intelligence systems require structured data to function. Machine-readable versions of many of our conceptual systems of information organization are being created and extended within the boundary infrastructure of Wikidata. I am hopeful that the opportunity to inspect infrastructure will be safeguarded by free/libre open source software communities such as those that create the MediaWiki software. I see the Wikidata community and the larger WMF communities as groups working to provide inspectability of their system.

The infrastructure of the semantic web is a resource that broad sectors of human population who regularly access the web rely on, even if it remains invisible to many. Those of us who are familiar with the power of KOS to shape information systems must work to understand this infrastructure and hold ourselves accountable to constructively critiquing how KOS are encoded in this infrastructure.

In order to remain positive and cognizant of my agency in these conversations about information organization and data structuring in the context of communities of commons-based peer production, like Wikidata, I rely on the theory of enacting a hermenutic of hope as conceptualized by Sandoval [Sandoval, 2000]. The potential that I see for the Wikidata community is to embrace these challenges and to develop tools to support critical, reflexive, and ethical evaluation of the implications of our decisions about how information is organized and data is structured in Wikidata in order to monitor the risks I describe in Chapter 7. Through the process of foregrounding the needs of communities most marginalized by capitalism, we can detect situations where harm is being done through activities related to knowledge organization.

8.6 Conclusion

The act of designating the infrastructure of Wikidata as belonging to an intellectual commons of humanity holds great potential for scholars of knowledge organization. I appreciate the values of the communities of commons-based peer production that collaborate to create Wikidata for their commitment to maintain transparency at all levels of the infrastructure of the system. I appreciate the work of members of these communities who contribute to these systems and make the code for the software (MediaWiki, Semantic MediaWiki, Wikibase, etc.) free to reuse, to modify and to share. Being able to inspect technical infrastructure is a precondition for being able to inspect how conceptual systems like knowledge organization systems are encoded into computational systems. The infrastructure of Wikidata is a powerful structure in that it governs how we organize information on the web. Let's work together to hold ourselves accountable for our decisions about information organization and data structuring in the knowledge base of structured data that anyone can edit.

BIBLIOGRAPHY

[med, 2016] (2016). Mediawiki. https://www.mediawiki.org/wiki/MediaWiki. Visited July 1, 2016.

[wdw, 2016] (2016). Wikidata workboard. https://phabricator.wikimedia.org/project/board/71/. Visited June 24, 2016.

[Aitchison et al., 2000] Aitchison, J., Gilchrist, A., and Bawden, D. (2000). *Thesaurus construction and use: a practical manual*. Psychology Press.

[Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). The semantic web: 6th international semantic web conference, 2nd asian semantic web conference, iswc 2007 + aswc 2007, busan, korea, november 11-15, 2007. proceedings. chapter DBpedia: A Nucleus for a Web of Open Data, pages 722–735. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Bao et al., 2012] Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., and Gergle, D. (2012). Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1075–1084. ACM.

[Barabási and Albert, 1999] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

[Barrett, 2008] Barrett, D. J. (2008). *MediaWiki*. " O'Reilly Media, Inc.".

[Beek et al., 2014] Beek, W., Rietveld, L., Bazoobandi, H. R., Wielemaker, J., and Schlobach, S. (2014). Lod laundromat: a uniform way of publishing other people's dirty data. In *The Semantic Web–ISWC 2014*, pages 213–228. Springer.

[Benkler, 2002] Benkler, Y. (2002). Coase's penguin, or, linux and" the nature of the firm". *Yale Law Journal*, pages 369–446.

[Benkler, 2006] Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

[Benkler, 2016] Benkler, Y. (2016). Degrees of freedom, dimensions of power. *Daedalus*, 145(1):18–32.

[Benkler et al., 2013] Benkler, Y., Shaw, A., and Hill, B. M. (2013). Peer production: a modality of collective intelligence. *Collective Intelligence*.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5):28–37.

[Bissig, 2015] Bissig, F. (2015). Drawing questions from wikidata. Master's thesis, ETH Zurich.

[Bizer et al., 2009] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web*, 7(3):154–165.

[Boiko, 2005] Boiko, B. (2005). *Content management bible*. John Wiley & Sons.

[Bostrom, 2014] Bostrom, N. (2014). *Superintelligence: Paths, Dangers*. Oxford.

[Bowker and Star, 1999] Bowker, G. and Star, S. L. (1999). Sorting things out. *Classification and its*.

[Bowker, 2005] Bowker, G. C. (2005). *Memory practices in the sciences*. Mit Press Cambridge, MA.

[Bowker et al., 2010] Bowker, G. C., Baker, K., Millerand, F., and Ribes, D. (2010). Toward information infrastructure studies: Ways of knowing in a networked environment. In *International handbook of internet research*, pages 97–117. Springer.

[Bratt, 2007] Bratt, S. (2007). Semantic web, and other technologies to watch. `https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(24)`. Visited August 21, 2016.

[Burke and Kraut, 2008] Burke, M. and Kraut, R. (2008). Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 27–36. ACM.

[Christian, 2012] Christian, B. (2012). *The most human human: what artificial intelligence teaches us about being alive*. Anchor.

[Coleman, 2013] Coleman, E. G. (2013). *Coding freedom: The ethics and aesthetics of hacking*. Princeton University Press.

[Cotera, 2010] Cotera, M. E. (2010). Women of color, tenure, and the neoliberal university: Notes from the field. In Nocella, A. J., Best, S., and McLaren, P., editors, *Academic repression: Reflections from the academic industrial complex*, pages 328–336. AK Press.

[Dong et al., 2014a] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014a). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610. ACM.

[Dong et al., 2014b] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014b). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 601–610, New York, NY, USA. ACM.

[Duarte and Belarde-Lewis, 2015] Duarte, M. E. and Belarde-Lewis, M. (2015). Imagining: creating spaces for indigenous ontologies. *Cataloging & Classification Quarterly*, 53(5-6):677–702.

[DuCharme, 2013] DuCharme, B. (2013). *Learning Sparql*. " O'Reilly Media, Inc.".

[Education,] Education, A. Research institute (aeri), pluralizing the archival curriculum group (pacg)(2011) educating for the archival multiverse. *Am Arch*, 73:69–101.

[Education and Institute, ired] Education, T. A. and Institute, R. (required). Educating for the archival multiverse. *American Archivist*, 74(1):69–101.

[Erxleben et al., 2014] Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., and Vrandečić, D. (2014). Introducing wikidata to the linked data web. In *The Semantic Web–ISWC 2014*, pages 50–65. Springer.

[Färber et al., 2015] Färber, M., Ell, B., Menne, C., and Rettinger, A. (2015). A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web Journal, July*.

[Farooq et al., 2007] Farooq, U., Kannampallil, T. G., Song, Y., Ganoe, C. H., Carroll, J. M., and Giles, L. (2007). Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 351–360. ACM.

[Feinberg, 2007] Feinberg, M. (2007). Hidden bias to responsible bias: an approach to information systems based on haraway's situated knowledges. *Information Research*, 12.

[Feinberg, 2011] Feinberg, M. (2011). Personal expressive bibliography in the public space of cultural heritage institutions. *Library Trends*, 59(4):588–606.

[Fischer, 1998] Fischer, D. (1998). From thesauri towards ontologies? *Advances in Knowledge Organization*, 6:18–30.

[Forte and Bruckman, 2008] Forte, A. and Bruckman, A. (2008). Scaling consensus: Increasing decentralization in wikipedia governance. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pages 157–157. IEEE.

[Foundation, 1996] Foundation, F. S. (1996). What is free software?

[Foundation, 2016] Foundation, F. S. (2016). Texinfo - the gnu documentation system. `https://www.gnu.org/software/texinfo/`. Visited August 14, 2016.

[Friedman et al., 1987] Friedman, D. P., Felleisen, M., and Bibby, D. (1987). *The little LISPER*. Mit Press Cambridge, Massachusetts.

[Geertz, 1994] Geertz, C. (1994). Thick description: Toward an interpretive theory of culture. *Readings in the philosophy of social science*, pages 213–231.

[Geiger and Ribes, 2010] Geiger, R. S. and Ribes, D. (2010). The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 117–126. ACM.

[Geiger and Ribes, 2011] Geiger, R. S. and Ribes, D. (2011). Trace ethnography: Following coordination through documentary practices. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10. IEEE.

[Gentle, 2012] Gentle, A. (2012). Documentation and my former self. In Pintscher, L., editor, *Open Advice FOSS: What We Wish Had Known When We Started*, pages 223–225.

[Georges-Pompidou, 2016] Georges-Pompidou, C. (2016). The building. `https://www.centrepompidou.fr/en/The-Centre-Pompidou/The-Building`. Online; accessed 21 April 2016.

[Glaser and Strauss, 1967] Glaser, B. and Strauss, A. (1967). The discovery of grounded theory. *Strategies for qualitative research. London: Weidenfeld and Nicolson*.

[Glaser, 1978] Glaser, B. G. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Sociology Press.

[Gödert et al., 2014] Gödert, W., Hubrich, J., and Nagelschmidt, M. (2014). *Semantic knowledge representation for information retrieval*. Walter de Gruyter GmbH & Co KG.

[Golder and Huberman, 2006] Golder, S. A. and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208.

[Good, 2015] Good, B. (2015). Poof it works – using wikidata to build wikipedia articles about genes. http://sulab.org/2015/10/poof-it-works-using-wikidata-to-build-wikipedia-articles-about-genes/. Online; accessed 18 Feb 2016.

[Google, 2012] Google (2012). Introducing the knowledge graph: things, not strings. https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html. Online; accessed 23 Feb 2016.

[Greve, 2012] Greve, G. (2012). Froeward. In Pintscher, L., editor, *Open Advice FOSS: What We Wish Had Known When We Started*, pages vii–ix.

[Groth et al., 2012] Groth, P., van Harmelen, F., Hoekstra, R., and Antoniou, G. (2012). A semantic web primer.

[Hansson, 2013] Hansson, J. (2013). The materiality of knowledge organization: epistemology, metaphors and society. *Knowledge organization*, 40(6):384–391.

[Hecht and Gergle, 2010] Hecht, B. and Gergle, D. (2010). The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 291–300. ACM.

[Hecht, 2007] Hecht, B. J. (2007). *Utilizing Wikipedia as a Spatiotemporal Knowledge Repository*. PhD thesis, University of California, Santa Barbara.

[Hernández et al., 2015] Hernández, D., Hogan, A., and Krötzsch, M. (2015). Reifying rdf: What works well with wikidata. In *Proceedings of the 11th International Workshop on Scalable Semantic Web Knowledge Base Systems*, volume 1457, pages 32–47.

[Herring et al., 2004] Herring, S. C., Barab, S., Kling, R., and Gray, J. (2004). An approach to researching online behavior. *Designing for virtual communities in the service of learning*, 338.

[Histropedia, 2016] Histropedia (2016). Histropedia: The timeline of everythin. `http://www.histropedia.com/`. Online; accessed 22 July 2016.

[Hitzler et al., 2010] Hitzler, P., Krotzsch, M., and Rudolph, S. (2010). *Foundations of semantic web technologies*. Chapman and Hall CRC Press.

[Hodge et al., 2003] Hodge, G. M., Zeng, M. L., and Soergel, D. (2003). Building a meaningful web: from traditional knowledge organization systems to new semantic tools. In *International Conference on Digital Libraries: Proceedings of the 3 rd ACM/IEEE-CS joint conference on Digital libraries*, volume 27, pages 417–417.

[Hofstadter, 1980] Hofstadter, D. R. (1980). Godel, escher, bach.

[Holloway et al., 2007] Holloway, T., Bozicevic, M., and Börner, K. (2007). Analyzing and visualizing the semantic coverage of wikipedia and its authors. *Complexity*, 12(3):30–40.

[Ismayilov et al.,] Ismayilov, A., Kontokostas, D., Auer, S., Lehmann, J., and Hellmann, S. Wikidata through the eyes of dbpedia.

[Jemielniak, 2014] Jemielniak, D. (2014). *Common Knowledge?: An Ethnography of Wikipedia*. Stanford University Press.

[Juel Vang, 2013] Juel Vang, K. (2013). Ethics of google's knowledge graph: some considerations. *Journal of Information, Communication and Ethics in Society*, 11(4):245–260.

[Kaganer et al., 2013] Kaganer, Kolja21, Bjankuloski06en, Pintscher, L., and Addshore (2013). Wikidata statment. `https://commons.wikimedia.org/w/index.php?curid=25322043`. Visited August 28, 2016.

[Keegan et al., 2012] Keegan, B., Gergle, D., and Contractor, N. (2012). Staying in the loop: structure and dynamics of wikipedia's breaking news collaborations. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, page 1. ACM.

[Kittur et al., 2009] Kittur, A., Chi, E. H., and Suh, B. (2009). What's in wikipedia?: mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1509–1512. ACM.

[Kochhar et al., 2010] Kochhar, S., Mazzocchi, S., and Paritosh, P. (2010). The anatomy of a large-scale human computation engine. In *Proceedings of the acm sigkdd workshop on human computation*, pages 10–17. ACM.

[Köhncke and Balke, 2010] Köhncke, B. and Balke, W.-T. (2010). Using wikipedia categories for compact representations of chemical documents. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1809–1812. ACM.

[Koren, 2012] Koren, Y. (2012). *Working with MediaWiki*. WikiWorks Press.

[Kriplean et al., 2008] Kriplean, T., Beschastnikh, I., and McDonald, D. W. (2008). Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 47–56. ACM.

[Kriplean et al., 2007] Kriplean, T., Beschastnikh, I., McDonald, D. W., and Golder, S. A. (2007). Community, consensus, coercion, control: cs* w or how policy mediates mass participation. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 167–176. ACM.

[Krippendorff, 2012] Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

[Krisnadhi et al., 2015] Krisnadhi, A. A., Hitzler, P., and Janowicz, K. (2015). On the capabilities and limitations of owl regarding typecasting and ontology design pattern views. In *Proceedings of the 12th International Workshop on OWL: Experiences and Directions (OWLED 2015) co-located with 14th International Semantic Web Conference on (ISWC 2015), Bethlehem, PA, USA*.

[Krötzsch et al., 2006] Krötzsch, M., Vrandečić, D., and Völkel, M. (2006). Semantic mediawiki. In *The Semantic Web-ISWC 2006*, pages 935–942. Springer.

[Kwasnik, 1992] Kwasnik, B. (1992). The role of classification structures in reflecting and building theory. *Advances in Classification Research Online*, 3(1):63–82.

[Kwasnik, 2000] Kwasnik, B. H. (2000). The role of classification in knowledge represantation and discovery. *Library trends*, 48(1).

[Lampland, 2009] Lampland, M. (2009). *Standards and their stories : how quantifying, classifying, and formalizing practices shape everyday life*. Cornell University Press, Ithaca.

[Lange et al., 2010] Lange, D., Böhm, C., and Naumann, F. (2010). Extracting structured information from wikipedia articles to populate infoboxes. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1661–1664. ACM.

[Lanier, 2014] Lanier, J. (2014). *Who owns the future?* Simon and Schuster.

[Lee and Olson, 2005] Lee, H.-L. and Olson, H. A. (2005). Hierarchical navigation: An exploration of yahoo! directories. *Knowledge Organization*, 32(1):10–24.

[Lehmann et al., 2012] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al. (2012). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

[Leonelli, 2010] Leonelli, S. (2010). Documenting the emergence of bio-ontologies: Or, why researching bioinformatics requires hpssb. *History and philosophy of the life sciences*, pages 105–125.

[Lessig, 2006] Lessig, L. (2006). Code: And other laws of cyberspace, version 2.0.

[Leung and White, 1989] Leung, H. K. and White, L. (1989). Insights into regression testing [software testing]. In *Software Maintenance, 1989., Proceedings., Conference on*, pages 60–69. IEEE.

[Luczak-Rösch et al., 2014] Luczak-Rösch, M., Simperl, E., Stadtmüller, S., and Käfer, T. (2014). The role of ontology engineering in linked data publishing and management: An empirical study. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(3):74–91.

[Mai, 2004] Mai, J. (2004). Classification of the web: challenges and inquiries. *Knowledge organization*, 31(2):92.

[Manske, 2016] Manske, M. (2016). The whelming. `http://magnusmanske.de/wordpress/`. Online; accessed 18 Feb 2016.

[McCanse, 2012] McCanse, S. (2012). Stop worrying and love the crowd. In Pintscher, L., editor, *Open Advice FOSS: What We Wish Had Known When We Started*, pages 109–112.

[McKinsey and Company, 2013] McKinsey and Company (2013). Open data: Unlocking innovation and performance with liquid information. `http://www.mckinsey.com/business-functions/business-technology/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information`. Online; accessed 18 Feb 2016.

[McMillan, 2000] McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the world wide web. *Journalism & Mass Communication Quarterly*, 77(1):80–98.

[MediaWiki, 2016] MediaWiki (2016). Markup spec/dtd. `https://www.mediawiki.org/wiki/Markup_spec/DTD`. Online; accessed 31 March 2016.

[Meditskos et al., 2013a] Meditskos, G., Dasiopoulou, S., Efstathiou, V., and Kompatsiaris, I. (2013a). Ontology patterns for complex activity modelling. In *Theory, Practice, and Applications of Rules on the Web*, pages 144–157. Springer.

[Meditskos et al., 2013b] Meditskos, G., Dasiopoulou, S., Efstathiou, V., and Kompatsiaris, I. (2013b). Sp-act: A hybrid framework for complex activity recognition combining owl and sparql rules. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pages 25–30. IEEE.

[Metz, 2015] Metz, C. (2015). A new company called alphabet now owns google. `http://www.wired.com/2015/08/new-company-called-alphabet-owns-google/`. Visited August 14, 2016.

[Millen et al., 2005] Millen, D., Feinberg, J., and Kerr, B. (2005). Social bookmarking in the enterprise. *Queue*, 3(9):28–35.

[Mitraka et al., 2015] Mitraka, E., Waagmeester, A., Burgstaller-Muehlbacher, S., Schriml, L. M., Su, A. I., and Good, B. M. (2015). Wikidata: A platform for data integration and dissemination for the life sciences and beyond. *bioRxiv*, page 031971.

[Morell, 2011] Morell, M. F. (2011). The wikimedia foundation and the governance of wikipedia's infrastructure: Historical trajectories and its hybrid character. *Critical point of view: A Wikipedia reader*, pages 325–341.

[Muchnik et al., 2007] Muchnik, L., Itzhack, R., Solomon, S., and Louzoun, Y. (2007). Self-emergence of knowledge trees: Extraction of the wikipedia hierarchies. *Physical Review E*, 76(1):016106.

[Müller-Birn et al., 2015] Müller-Birn, C., Karran, B., Lehmann, J., and Luczak-Rösch, M. (2015). Peer-production system or collaborative ontology engineering effort: What is wikidata? In *Proceedings of the 11th International Symposium on Open Collaboration*, page 20. ACM.

[Musante and DeWalt, 2010] Musante, K. and DeWalt, B. R. (2010). *Participant observation: A guide for fieldworkers*. Rowman Altamira.

[Nardi and Engeström, 1999] Nardi, B. A. and Engeström, Y. (1999). A web on the wind: The structure of invisible work. *Computer Supported Cooperative Work (CSCW)*, 8(1):1–8.

[Nardi and Miller, 1991] Nardi, B. A. and Miller, J. R. (1991). Twinkling lights and nested loops: distributed problem solving and spreadsheet development. *International Journal of Man-Machine Studies*, 34(2):161–184.

[Niederer and Van Dijck, 2010] Niederer, S. and Van Dijck, J. (2010). Wisdom of the crowd or technicity of content? wikipedia as a sociotechnical system. *New Media & Society*, 12(8):1368–1387.

[Noy and McGuinness, 2001] Noy, N. and McGuinness, D. (2001). Ontology development 101: A guide to creating your first ontology tech. rep. no. Technical report, KSL-01-05, SMI-2001. Stanford, CA: Stanford Knowledge Systems Laboratory and Stanford Medical Informatics.

[OED, 2004] OED (2004). ontology, n. `https://www.oed.com/`. Visited June 30, 2016.

[Olson, 1998] Olson, H. A. (1998). Mapping beyond dewey's boundaries: Constructing classificatory space for marginalized knowledge domains. *Library trends*, 47(2):233–254.

[Olson, 2013] Olson, H. A. (2013). *The power to name: locating the limits of subject representation in libraries*. Springer Science & Business Media.

[Open Knowledge, 2016] Open Knowledge (2016). Open Data Handbook. `https://opendatahandbook.org/`. Online; accessed 18 February 2016.

[Parameswaran and Whinston, 2007] Parameswaran, M. and Whinston, A. B. (2007). Social computing: An overview. *Communications of the Association for Information Systems*, 19(1):37.

[Park, Jane, 2010] Park, Jane (2010). Wikipedia on new facebook community pages. `https://blog.creativecommons.org/2010/04/21/wikipedia-on-new-facebook-community-pages/`. Online; accessed 31 March 2016.

[phabricator, 2016] phabricator (2016). A complete software development platform. `https://www.phacility.com/`. Visited June 24, 2016.

[Ponzetto and Strube, 2007] Ponzetto, S. P. and Strube, M. (2007). Deriving a large scale taxonomy from wikipedia. In *AAAI*, volume 7, pages 1440–1445.

[Priedhorsky and Terveen, 2011] Priedhorsky, R. and Terveen, L. (2011). Wiki grows up: arbitrary data models, access control, and beyond. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 63–71. ACM.

[Proto, 2005] Proto, F. (2005). The pompidou centre: or the hidden kernel of dematerialisation. *The Journal of Architecture*, 10(5):573–589.

[Reardon et al., 2015] Reardon, J., Metcalf, J., Kenney, M., and Barad, K. (2015). Science and justice: The trouble and the promise. *Catalyst: Feminism, Theory, Technoscience*, 1(1).

[Ribes and Bowker, 2009] Ribes, D. and Bowker, G. (2009). Between meaning and machine: Learning to represent the knowledge of communities. *Information and Organization*, 19(4):199–217.

[Ribes et al., 2013] Ribes, D., Jackson, S., Geiger, S., Burton, M., and Finholt, T. (2013). Artifacts that organize: Delegation in the distributed organization. *Information and Organization*, 23(1):1–14.

[Roush, 2005] Roush, W. (2005). Social machines: Computing means connecting. *MIT Technology Review*, 108(8):44.

[Sandoval, 2000] Sandoval, C. (2000). *Methodology of the Oppressed*, volume 18. U of Minnesota Press.

[Schmidt and Bannon, 1992] Schmidt, K. and Bannon, L. (1992). Taking cscw seriously. *Computer Supported Cooperative Work (CSCW)*, 1(1-2):7–40.

[Schuler, 1994] Schuler, D. (1994). Social computing. *Commun. ACM*, 37(1):28–29.

[Semantic MediaWiki, 2015] Semantic MediaWiki (2015). Faq. `https://www.semantic-mediawiki.org/wiki/FAQ/`. Online; accessed 28 January 2016.

[Sen et al., 2006] Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., and Riedl, J. (2006). Tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 181–190. ACM.

[Sen et al., 2014] Sen, S., Li, T. J.-J., Team, W., and Hecht, B. (2014). Wikibrain: Democratizing computation on wikipedia. In *Proceedings of The International Symposium on Open Collaboration*, page 27. ACM.

[Shaw and Hill, 2014] Shaw, A. and Hill, B. M. (2014). Laboratories of oligarchy? how the iron law extends to peer production. *Journal of Communication*, 64(2):215–238.

[Shroff, 2013] Shroff, G. (2013). *The Intelligent Web: Search, smart algorithms, and big data*. Oxford University Press.

[Siebrand, 2007] Siebrand (2007). Hotcat.png. `https://commons.wikimedia.org/w/index.php?curid=2431080`. Online; accessed 31 March 2016.

[Smith, 1998] Smith, B. (1998). *The Truth That Never Hurts: Writings on Race, Gender, and Freedom*. Rutgers University Press.

[Sowa, 1983] Sowa, J. F. (1983). Conceptual structures: information processing in mind and machine.

[Spinellis and Louridas, 2008] Spinellis, D. and Louridas, P. (2008). The collaborative organization of knowledge. *Communications of the ACM*, 51(8):68–73.

[Spitz et al., 2016] Spitz, A., Dixit, V., Richter, L., Gertz, M., and Geiß, J. (2016). State of the union: A data consumer's perspective on wikidata and its properties for the classification and resolution of entities. In *Tenth International AAAI Conference on Web and Social Media*.

[Stallman, 2009] Stallman, R. (2009). Viewpoint why open source misses the point of free software. *Communications of the ACM*, 52(6):31–33.

[Stallman and Lessig, 2010] Stallman, R. M. and Lessig, L. (2010). Free software, free society: selected essays of richard m. stallman.

[Star, 1999] Star, S. L. (1999). The ethnography of infrastructure. *American behavioral scientist*, 43(3):377–391.

[Star, 2010] Star, S. L. (2010). This is not a boundary object: Reflections on the origin of a concept. *Science, Technology and Human Values*, 35(5):601–617.

[Star and Bowker, 2006] Star, S. L. and Bowker, G. C. (2006). How to infrastructure. In Lievrouw, L. and Livingstone, S. M., editors, *Handbook of new media : Social shaping and social consequences of ICTs*, pages 230–245. SAGE.

[Star and Ruhleder, 1994] Star, S. L. and Ruhleder, K. (1994). Steps towards an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 253–264. ACM.

[Star and Ruhleder, 1996] Star, S. L. and Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information systems research*, 7(1):111–134.

[Star and Strauss, 1999] Star, S. L. and Strauss, A. (1999). Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer supported cooperative work (CSCW)*, 8(1-2):9–30.

[Steiner, 2014] Steiner, T. (2014). Bots vs. wikipedians, anons vs. logged-ins (redux): A global study of edit activity on wikipedia and wikidata. In *Proceedings of The International Symposium on Open Collaboration*, page 25. ACM.

[Steiner et al., 2012] Steiner, T., Verborgh, R., Troncy, R., Gabarro, J., and Van de Walle, R. (2012). Adding realtime coverage to the google knowledge graph. In *11th International Semantic Web Conference (ISWC 2012)*. Citeseer.

[Suchman, 1995] Suchman, L. (1995). Making work visible. *Communications of the ACM*, 38(9):56–ff.

[Suchman, 1983] Suchman, L. A. (1983). Office procedure as practical action: models of work and system design. *ACM Transactions on Information Systems (TOIS)*, 1(4):320–328.

[Svenonius, 2000] Svenonius, E. (2000). *The intellectual foundation of information organization*. MIT press.

[Syed and Finin, 2010] Syed, Z. and Finin, T. (2010). Unsupervised techniques for discovering ontology elements from wikipedia article links. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 78–86. Association for Computational Linguistics.

[Thévenot, 1984] Thévenot, L. (1984). Rules and implements: investment in forms. *Social Science Information*, 23(1):1–45.

[Tolksdorf and Simperl, 2006] Tolksdorf, R. and Simperl, E. P. B. (2006). Towards wikis as semantic hypermedia. In *Proceedings of the 2006 international symposium on Wikis*, pages 79–88. ACM.

[Tsing, 2015] Tsing, A. L. (2015). *The Mushroom at the End of the World: On the Possibility of Life in Capitalist Ruins*. Princeton University Press.

[Van Hooland and Verborgh, 2014] Van Hooland, S. and Verborgh, R. (2014). *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata*. Facet.

[Vedral, 2010] Vedral, V. (2010). *Decoding reality: the universe as quantum information*. Oxford University Press.

[Vickery, 1958] Vickery, B. C. (1958). Classification and indexing in science. *Classification and indexing in science.*

[Viégas et al., 2007] Viégas, F. B., Wattenberg, M., Kriss, J., and Van Ham, F. (2007). Talk before you type: Coordination in wikipedia. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, pages 78–78. IEEE.

[Völkel et al., 2006] Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., and Studer, R. (2006). Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web*, pages 585–594. ACM.

[Voss, 2006] Voss, J. (2006). Collaborative thesaurus tagging the wikipedia way. *arXiv preprint cs/0604036*.

[Vrandecic, 2013] Vrandecic, D. (2013). The rise of wikidata. *IEEE Intelligent Systems*, (4):90–95.

[Vrandečić and Krötzsch, 2014] Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

[W3C, 2006] W3C (2006). Naming and addressing: Uris, urls, ... `https://www.w3.org/Addressing/`. Visited August 21, 2016.

[W3C, 2015] W3C (2015). RDF. `https://www.w3.org/RDF/`. Online; accessed 15 January 2016.

[Wadhwa, Kul Takano, 2010] Wadhwa, Kul Takano (2010). heads up:wikipedia on facebook. `https://lists.wikimedia.org/pipermail/foundation-l/2010-April/057598.html`. Online; accessed 31 March 2016.

[Wattenberg et al., 2007] Wattenberg, M., Viégas, F. B., and Hollenbach, K. (2007). Visualizing activity on wikipedia with chromograms. In *Human-Computer Interaction–INTERACT 2007*, pages 272–287. Springer.

[Welser et al., 2011] Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., and Smith, M. (2011). Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, pages 122–129. ACM.

[Wetzker et al., 2008] Wetzker, R., Zimmermann, C., and Bauckhage, C. (2008). Analyzing social bookmarking systems: A del. icio. us cookbook. In *Proceedings of the ECAI 2008 Mining Social Data Workshop*, pages 26–30.

[Wikibase, 2015] Wikibase (2015). Wikibase. http://wikiba.se/. Online; accessed 15 January 2016.

[Wikidata, 2015a] Wikidata (2015a). Wikidata:data access. https://www.wikidata.org/wiki/Wikidata:Data_access. Online; accessed 15 January 2016.

[Wikidata, 2015b] Wikidata (2015b). Wikidata:databse downloads. https://www.wikidata.org/wiki/Wikidata:Database_download. Online; accessed 15 January 2016.

[Wikidata, 2016] Wikidata (2016). Statistics.

[Wikidata, 2016a] Wikidata (2016a). Wikidata query service user manual. https://www.mediawiki.org/wiki/Wikidata_query_service/User_Manual. Online; accessed 1 Feb 2016.

[Wikidata, 2016b] Wikidata (2016b). Wikidata:Past IRC office hours. https://www.wikidata.org/wiki/Wikidata:Events#Past_IRC_office_hours. Online; accessed 17 March 2016.

[Wikidata, 2016c] Wikidata (2016c). Wikidata:Project Chat. https://www.wikidata.org/wiki/Wikidata:Project_chat. Online; accessed 17 March 2016.

[Wikidata, 2016] Wikidata (2016). Wikidata:tools/external tools. https://www.wikidata.org/wiki/Wikidata:Tools/External_tools. Online; accessed 23 Feb 2016.

[Wikimedia, 2016] Wikimedia (2016). Wikimedia downloads. https://dumps.wikimedia.org/. Online; accessed 1 Feb 2016.

[Wikipedia, 2015] Wikipedia (2015). Wikipedia:Manual of Style. https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Infoboxes. Online; accessed 2 Feb 2016.

[Wikipedia, 2016a] Wikipedia (2016a). Categorization: Hiding categories. https://en.wikipedia.org/wiki/Wikipedia:Categorization#Hiding_categories. Online; accessed 31 March 2016.

[Wikipedia, 2016b] Wikipedia (2016b). Category: Hidden categories. https://en.wikipedia.org/wiki/Category:Hidden_categories. Online; accessed 31 March 2016.

[Wikipedia, 2016c] Wikipedia (2016c). Category:wikipedians who use hotcat. https://en.wikipedia.org/wiki/Category:Wikipedians_who_use_HotCat. Online; accessed 31 March 2016.

[Wikipedia, 2016d] Wikipedia (2016d). Hot cat. https://en.wikipedia.org/wiki/Wikipedia:HotCat. Online; accessed 31 March 2016.

[Wikipedia, 2016] Wikipedia (2016). List of wikipedias. https://meta.wikimedia.org/wiki/List_of_Wikipedias. Visited August 21, 2016.

[Wikipedia, 2016] Wikipedia (2016). propoerty:p910. https://en.wikipedia.org/wiki/Wikipedia:HotCat. Online; accessed 31 March 2016.

[Wikipedia, 2016] Wikipedia (2016). Wikimedia foundation — wikipedia, the free encyclopedia. [Online; accessed 22-April-2016].

[Wikipedia, 2016a] Wikipedia (2016a). Wikipedia: Copyrights. https://en.wikipedia.org/wiki/Wikipedia:Copyrights. Online; accessed 18 April 2016.

[Wikipedia, 2016b] Wikipedia (2016b). Wikipedia talk:Categorization. https://en.wikipedia.org/wiki/Wikipedia_talk:Categorization. Online; accessed 17 March 2016.

[Wilson, 1968] Wilson, P. (1968). *Two kinds of power: An essay on bibliographical control*. Univ of California Press.

[Wright, 2014] Wright, A. (2014). *Cataloging the World: Paul Otlet and the Birth of the Information Age*. Oxford University Press.

[Wright, 2007] Wright, S. E. (2007). Coping with indeterminacy: Terminology and knowledge representation resources in digital environments. In Antia, B., editor, *Indeterminacy in terminology and LSP: studies in honour of Heribert Picht*. John Benjamins Publishing.

[Wu and Weld, 2008] Wu, F. and Weld, D. S. (2008). Automatically refining the wikipedia infobox ontology. In *Proceedings of the 17th international conference on World Wide Web*, pages 635–644. ACM.

[Yu et al., 2007] Yu, J., Thom, J. A., and Tam, A. (2007). Ontology evaluation using wikipedia categories for browsing. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 223–232. ACM.

[Zachry, 2008] Zachry, M. (2008). "an interview with susan leigh star". *Technical Communication Quarterly*, 17(4):435–454.