Blogs as Infrastructure for Scholarly Communication

by

Matt Burton

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2015

Doctoral Committee:

Associate Professor Paul L. Conway, Chair
Assistant Professor Megan S. Ankerson
Senior Lecturer Maria S. Bonn, University of Illinois
Associate Professor Carl Lagoze
Associate Professor Qiaozhu Mei

Dedication

To Shannon.

Acknowledgements

Thanks to Shannon Riffe for her everlasting patience and support. Many thanks to my parents, Lex and Corky Burton, who supported me through this long journey. To my committee, Megan Ankerson, Carl Lagoze, Qiaozhu Mei, and Maria Bonn, I give thanks for their feedback. Special thanks to my chair, Paul Conway, whose heroic effort helped me jump through the various hoops of fire. Many thanks to the UMSI doctoral students, especially Adam Kriesberg, Melissa Chalmers, and Rebecca Frank, whose feedback was invaluable. Also thanks to the crew in Digital Scholarship Services, Nora Mattern, Aaron Brenner, and Michael Bolam, who let me juggle two jobs. Thanks to Amelia Acker and Brian Beaton as SIS for bouncing ideas for me. Thanks to Sandford C Bledsoe III, Mark and Kate Morris, and everyone else who served me copious amounts of coffee. Thanks to John Davi and the guys at Diffbot for the use of their services and support.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

This project systematically analyzes digital humanities blogs as an infrastructure for scholarly communication. This exploratory research maps the discourses of a scholarly community to understand the infrastructural dynamics of blogs and the Open Web. The text contents of 106,804 individual blog posts from a corpus of 396 blogs were analyzed using a mix of computational and qualitative methods. Analysis uses an experimental methodology (trace ethnography) combined with unsupervised machine learning (topic modeling), to perform an interpretive analysis at scale. Methodological findings show topic modeling can be integrated with qualitative and interpretive analysis. Special attention must be paid to *data fitness*, or the shape and re-shaping practices involved with preparing data for machine learning algorithms. Quantitative analysis of computationally generated topics indicates that while the community writes about diverse subject matter, individual scholars focus their attention on only a couple of topics. Four categories of informal scholarly communication emerged from the qualitative analysis: *quasi-academic, para-academic, meta-academic,* and *extra-academic*. The quasi and para-academic categories represent discourse with scholarly value within the digital humanities community, but do not necessarily have an obvious path into formal publication and preservation. A conceptual model, the *(in)visible college,* is introduced for situating scholarly communication on blogs and the Open Web. An (in)visible college is a kind of scholarly communication that is *informal*, yet *visible at scale*. This combination of factors opens up a new space for the study of scholarly communities and communication. While (in)invisible colleges are programmatically observable, care must be taken with any effort to count and measure knowledge work in these spaces. This is the first systematic, data driven analysis of the digital humanities and lays the groundwork for subsequent social studies of digital humanities.

Chapter One

Introduction

Many, if not all, graduate students have existential doubt in the final stages of the dissertations. Intellectual and geographic isolation, conflictions about the subject matter, anxieties about the job market escalate the existing emotional burden of writing the damn thing. Amid all of this, there are some who think, “I know, I’ll start a blog!”

Matt Gold, now an associate professor of English and digital humanities at the City University of New York, decided starting a blog would be a good idea while finishing his dissertation and living in another city, isolated from his academic peers. The desire to write and engage a public audience overcame the stress and fear the discovery of his pseudo-anonymous political blog would torpedo his chances at a tenure track job. In 2005, scholarly bloggers were told they “need not apply” to tenure track positions because, “what is the purpose of broadcasting one’s unfiltered thoughts to the whole wired world?”[1]

Now, a decade later, many hearts and minds have softened on scholarly blogging in the humanities. Fear has been replaced with enthusiasm and anonymity with professional identity. The expectations of the scholarly blog as a format have expanded. While blogs are certainly home to “unfiltered thoughts” they are also the home of a scholarly community, the *digital humanities*.

Digital humanities (DH) is an emerging, heterogeneous collective of activities oriented around the use and study of digital technology within the humanities.

[1] Tribble, Ivan. “Bloggers Need Not Apply.” *The Chronicle of Higher Education*, July 8, 2005. https://chronicle.com/article/Bloggers-Need-Not-Apply/45022

For five decades under the moniker “humanities computing,” scholars leveraged the power of computational analysis on traditional humanities research objects, like historical texts.

While debates on disciplinary and interdisciplinary practice are always concomitant with the emergence of a new scholarly discipline (Gold 2012), the digital humanities is unique in the places where these conversations unfold. Digital humanists love social media, they blog and tweet with a vigor unrivaled elsewhere in academia; indeed they blog so much that media scholar Tara McPherson (2009) has characterized the community as the “blogging humanities.” The enthusiastic adoption of the medium raises important questions about the complex social and technical relationships between disciplines and the infrastructures within which they practice.

This project has two aims: first, to explore how the digital humanities community uses blogs as an infrastructure for scholarly communication; and second, to explore how new methods and techniques of textual data analysis, such as topic modeling, can be utilized to better understand traces of scholarly communication whose scale defies traditional interpretive analysis. This project resulted in the production of a dataset of scholarly blogs, cleaned and prepared for quantitative and qualitative analysis. The data preparation practices, the quantitative models, and content were analyzed and interpreted as a set of traces from a new methodological perspective, *trace ethnography* (Geiger and Ribes 2010; Geiger and Ribes 2011; Ribes 2014).

Studies of *scholarly* blogs are few and those that do exist reveal a widely varied and rich source of data that has yet to be investigated (Halavais 2006; Saper 2006; Luzón 2009; Kjellberg 2010; Hank 2011; Hank 2013). New modes of scholarly communication inspires fresh conceptual models of knowledge production and dissemination, and informs the design of research, teaching, and publishing infrastructure for the humanities and social sciences (Welshons 2009).

Examining and exploring these phenomena is important for several reasons. Digital humanities is only the beginning of a larger shift in scholarly communication approaches the *Open Web* as a discursive infrastructure. The adoption and use of blogs occurs within the context of a "crisis in scholarly publishing" (Waters 2004; Lynch 2010) and digital scholarship as a viable alternative to the academic monograph. There is much to be learned from the ways in which the community interacts and experiments with knowledge production afforded by digital technology.

Beyond a description of technology adoption in scholarly practice, there is also an under-scrutinized *infrastructural* story to be told about digital humanities blogs. By *infrastructure* I mean the material and sociotechnical assemblage of people, technology, and practices within, upon, and through which human interactions and technological processes are carried out. An infrastructural perspective looks "under the hood," seeking out the boring and banal, the taken-for-granted dynamics embedded within both social practice and technical structure. Blogs are form of communication that has emerged from the Open Web, an information infrastructure that affords *open access* and *open publishing* using *open standards*. The use of blogs in the digital humanities raises questions about the changing nature of scholarly communication and its relationship to the Open Web.

Why study blogs as scholarly communication?

Scholarly communication, as a subject of research, is mainly focused on bibliometrics or scientometrics, a field whose research orientation studies structures of published literature and citation networks (Garfield 1955; Price 1963; Borgman and Furner 2002). These fields focus their research on bibliographic databases (like Scopus or Web of Science) that contain abstracts and citations of *formally* published scholarly communication. Yet, there is another world of *informal* scholarly communication that has previously been programmatically unavailable to scholars (let alone packaged and sold as a database). *Informal* scholarly communication is less well articulated and understood. The dominant framings of informal scholarly communication as "invisible colleges" (Price 1963; Price and Beaver 1966) highlight the difficulty in gaining access to study (Lievrouw 1989). However, the new platforms and technologies render *visible* the previously *invisible* networks of communication.

The Internet and the social web have changed the ways and means by which scholars communicate (Cronin et al. 1998). Scholars in some communities and disciplines have adopted public modalities such as blogs, and social media platforms (i.e., Facebook, and Twitter). Scholars constitute themselves through material and discursive acts (Latour 2005). To study such work is to come to a deeper understanding of the nature of knowledge and scholarship, and the practices of their production. What does it mean that scholarly activity is public and visible rather than being hidden behind paywalls and cloistered in the ivory tower?

Why study digital humanities?

Digital humanities continues a centuries-old scholarly tradition of investigating human knowledge and the collective cultural record. The adoption of innovative new media and methods for research, publication, and communication has important implications for how humanity comes to understand the world. But *where are the social studies of digital humanities* (Borgman 2009)?

Collectively, the digital humanities represent an emerging sociotechnical configuration of people, technology, and practice. This project characterizes the digital humanities as a loosely defined community, but it could equally be considered a field, a discipline, a movement, or all of the above. Above all the wonky definitional exercises, digital humanities is pregnant with the hopes and dreams of humanities scholars seeking a way out of scholarly practices that have ossified along disciplinary and epistemological lines inhibiting innovation in the exploration, understanding, and interpretation of humanity's cultural heritage.

Why study blogs as infrastructure?

Bruno Latour (1991) said it best in the aphoristic title to his essay "Technology is Society Made Durable." The designs of technologies are deeply social (Bijker 1997) and provide a material mirror that we can use to reflexively decipher our own complexities. Infrastructural technologies are particularly interesting in this regard precisely because they encode and reify aspects of our social life into technical and material systems often invisible and dismissed as mundane. Infrastructures–as a configuration of people, technology, and practice–recursively shape subsequent configurations of people, technology, and practice. As Susan Leigh Star (1999a) said, "it is infrastructure all the way down." From this perspective the issue may not be *why study infrastructure,* but rather, *why study infrastructurally?* (Star 1999b) This minor pivot to the question shifts the focus away from considering infrastructure as an object to a focus on the infrastructural aspects of any phenomena, in this case, digital humanities blogs.

Research Questions

This project examines how members of a scholarly community use blogs as a mode of communication. The research brings together three distinct, but interconnected, areas of inquiry: scholarly communication, digital humanities, and infrastructure studies. Each has been productively studied separately; but never before have they been brought together into one analytical frame.

- *What roles do blogs and the Open Web play in facilitating scholarly communication in the digital humanities?*

This motivating question is addressed through an analysis of 396 digital humanities blogs sourced from an aggregation of blogs assembled by members of the digital humanities community, the *Compendium of Digital Humanities*.[2] 106,804 individual posts were scraped from the open web and prepared for quantitative analysis. Due to the amount of data, unsupervised topic modeling was used to scale down the data into representation suitable for qualitative

[2] Digital Humanities Now, a service for aggregating digital humanities blogs, assembles the Compendium. http://digitalhumanitiesnow.org/how-this-works/

analysis and interpretation. Topic modeling produces clusters of co-occurring words, topics, and connects documents based upon shared topic proportions. Trace ethnography informed a closer reading of the topic model, the scaling practices, and blog content.

The collection, analysis, and interpretation of blogs focuses on the following three questions:

- *What are the methodological dynamics and tensions in generating data from the Open Web?*
- *What themes are digital humanities scholars writing about on their blogs?*
- *How does scholarly blogging expand current understandings of informal scholarly communication?*

Chapter 2 reviews the relevant literatures in scholarly communication, digital humanities, and infrastructure studies. With respect to scholarly communication, the chapter focuses upon the distinction between formal and informal communication, the “invisible college” as a model of informal communication, and the new opportunities for studying the informal afforded by scholarly communication online. With respect to digital humanities, the chapter focuses the origins of the community in humanities computing, emergence of the “blogging humanities,” and the need for more social studies of digital humanities. With respect to infrastructure studies, the chapter focuses upon the relationship between technology and social order, and introduces the Open Web and scholarly blogs as phenomenon to be studied as infrastructure.

Chapter 3 introduces topic modeling and trace ethnography. There is a discussion of the blogs, and the *Compendium of Digital Humanities*, a review of topic modeling, what it is and how it has been used in social science and the digital humanities; and finally, the chapter addresses the problem of performing ethnographic analysis of traces *at scale*. The chapter argues trace ethnography of scaling must include a thick description of the traces of the subject and the traces of the scaling process in the ethnographic analysis.

Chapter 4 focuses on the sub-question: *What are the methodological dynamics and tensions in generating data from blogs/Open Web?* The chapter follows from the arguments in chapter three by providing a thick description of the data collection, preparation, and scaling. The chapter is an "ethnography of scaling" (Ribes 2014) and addresses benefits and challenges of how data driven social science can be transformed by reflexively engaging the documentary traces of its own research practice. The chapter concludes with a discussion about the current state of web archiving in support of computational and data driven research.

Chapter 5 focuses on the sub-question: *What themes are digital humanities scholars writing about on their blogs?* This chapter includes ethnographic description of the MALLET topic model; analyzes the degree of topic diversity on the blogs; visualizes the entire corpus as topic clusters; and introduces four analytical categories of scholarly discourse derived from the ethnographic analysis of traces. These categories, *quasi-academic*, *meta-academic*, *para-academic*, and *extra-academic,* seek to characterize *informal* scholarly communication.

The final chapter answers the question*: how does scholarly blogging expand current understandings of informal scholarly communication?* It introduces the concept of then (in)visible college as a holistic model of *informal* scholarly communication is *visible at scale*. The dynamics of an (in)visible college are explored through Menzel's six functions of informal scholarly communication. Blogs and the Open Web mean data driven methods can be used to *know* and *understand* scholarly communities. This exploratory analysis shows blogs serve a pivotal role in maintaining the digital humanities as a community and experimental methods like trace ethnography can scale while still maintaining their contextual and interpretive richness. The final chapter ends by considering some tensions around considering, and counting, blogs as a form of legitimated scholarship.

Chapter Two

Literature Review

Digital humanities scholars are important and interesting subjects to study because, as a community of practice, they tend to use information and communication technologies in novel and interesting ways. In particular, they use the web, blogs, and social media to circulate information and stay connected. At this point, a decade after the term “digital humanities” was popularized, the community has matured into something resembling a field or discipline, but with a particularly new relationship to technology (for a humanities community).

Communication is the “essence of science” (Garvey 1979). The more we understand how scholars communicate, the better we can understand the production and circulation of knowledge. The history of scholarly communication, and its progeny like bibliometrics, has mainly focused on scholar’s (mainly scientist’s) *formally* published artifacts: book and periodicals (Garfield 1955; Price 1963; Lievrouw 1989; Borgman 1989; Borgman and Furner 2002). Yet, informal channels of scholarly communication are vitally important to academic knowledge practice. Informal scholarly communication, characterized as the “invisible college,” has been difficult to study directly, so it has been theorized as a latent structure within formal citation networks (Crane 1972), or studied in in micro in fields like laboratory studies (Lievrouw 1989). New technologies, especially the Internet and the web, have transformed scholarly communication (Cronin et al. 1998). Preprint archives and networked platforms such as arXiv enable radical new modes of knowledge generation, circulation, and preservation. These platforms cloud the distinction between formal and informal scholarly communication by making previously *invisible* informal interactions not only visible, but also *visible at scale*. This creates new opportunities for research in informal scholarly communication and the digital humanities.

New platforms and new modes of communication exist as a function of the *Open Web* (Berners-Lee 2000). The Open Web has three technical dynamics, open standards, open access, and open publishing (Çelik 2010) and from these dynamics blogs have emerged as an infrastructure *built upon* the *installed base* of the Open Web. While blogs have been studied as a general phenomenon (Rodzvilla 2002; Bruns and Jacobs 2006; Rosenberg 2010; Dean 2010; Rettberg 2013), the scholarly use of blogs has only begun to be studied (Halavais 2006; Saper 2006; PuschmannI and MahrtII 2012; Hank 2013). Blogs are a technological substrate in and through which digital humanities constitutes itself as a scholarly community.

The Open Web and its associated communication technologies are the underlying technological infrastructure that powers nearly all information exchange today. As a vibrant field of inquiry, Infrastructure Studies provides a theoretical frame to guard against a technologically deterministic view of people, technology, and practice (Edwards et al. 2007). Studying blogs *infrastructurally* teases out the *sociotechnical* dynamics of people, technology, and practices (Star 1999b). However, current methodological approaches for studying phenomena infrastructurally don't scale; innovations mixing qualitative and quantitative methods are necessary to understand large infrastructural objects (Bowker et al. 2010).

This review has four sections, a brief introduction to the digital humanities, a review of research about informal scholarly communication, a description of the Open Web as a technical frame, and infrastructures as a theoretical lens. Each of these subjects, especially digital humanities and scholarly communication, has a depth of discourse far richer than the justice I can give it in this chapter. My focus is to highlight specific threads within each body of research to motivate a study of blogs as an infrastructure for informal scholarly communication in the digital humanities.

Digital Humanities

The practice of humanities scholarship has been informed by the use of computing technologies for at least sixty years. The early period of *humanities computing* (ca. 1949 to 1998) had an instrumentalized relationship to technology. Technology, and more specifically, the computer was (and is) a powerful, productive, and generative “telescope of the mind” (McCarty 2012) enabling new kind of research. A more complete early history of humanities computing has been thoroughly documented by Susan Hockey (2004) in her chapter of the *Compendium of Digital Humanities*.

Digital humanities has outgrown its humanities computing origins. And with that growth has emerged social and technical complexity. Computers are no longer simply an instrument; they are objects to be critically interrogated as systems with history, agency, and power. The “humanities computing” era has expanded into the “digital humanities” era with a much broader relationship to technology. Today, digital humanities use technology as a tool, a research object, an experimental laboratory, an activist project, and an expressive medium (Svensson 2010).

Humanities Computing: Technology = Tool

All histories of the digital humanities recognize the origins of the discipline in humanities computing. Cathy Davidson (2008) points out that computing has had a transformative impact, “humanities 1.0—computational humanities—has changed the way we do research, the kinds of questions we can ask, and the depth, breadth, and detail of the answers we can provide to those questions” (p." (710). Humanities computing has, for over six decades, slowly and methodologically transformed the foundations of humanities research practice. Humanities computing found its first foothold in the computer-assisted analysis of text (Hockey 2004). The path cleared by computing humanists created the opportunity space digital humanities now occupy.

When writing and talking about these origins, we cannot ignore the truly visionary work of Father Roberto Busa. In 1949, Busa began a lemmatization of the complete works of Saint Thomas Aquinas, a task previously considered impossible because the scale was too great.[3] Busa's work demonstrated a new horizon of praxis for the humanities (Hockey 2004; Klein 2014).

> Humanities computing is precisely the automation of every possible analysis of human expression (therefore, it is exquisitely a "humanistic" activity), in the widest sense of the word, from music to the theater, from design and painting to phonetics, but whose nucleus remains the discourse of written texts (Busa 2004).

By this sentiment, technology and computation enable scholars to overcome the ever-increasing scale of human expression (mainly written texts). This analytical relationship to human expression is what Franco Moretti calls *distant reading* (Moretti 2005). The term arose as a response to a methodological problem faced by literary historians. Traditional methods of humanist inquiry, such as the careful close reading of individual texts, do not scale with the increased volume of and access to a digitized and computable cultural record. While the close reading of the literary canon has proved extremely insightful, what new and important knowledge can we extract from an *entire corpus* of texts (Wilkens 2012)?

In the context of scale, humanities computing tests the ideological, methodological, and epistemological boundaries within and across the humanities. The practice of computationally analyzing text is alien, unfamiliar, incomprehensible, and possibly unconvincing to many traditional humanities scholars (Juola 2008; McPherson 2009). The practice of implementing the probabilistic models for document clustering is far from standard practice for literary historians (despite their truly elegant descriptions of how these models work). Furthermore, humanities computing tended to ignore issues of race, class,

[3] There is an untold history of Father Busa''s female assistantspunch card operators who encoded the writings of St. Thomas Aquinas onto punch cards. This work, which took thirty30 years, is typically left out of histories and stories of Busa's project. *In a blog post*, Melissa Terras posted photographs of these women and highlights the need for further research into the roles and responsibilities of these women whose work is the structural foundation of the digital humanities today. To what extent was the scale of St. Thomas Aquinas's works overcome by technology or the emergence of a new class of laborers after the war? http://melissaterras.blogspot.com/2013/10/for-ada-lovelace-day-father-busas.html

gender, and power (McPherson 2012), subjects of deep interest in the wake of the post-structural turn in the academy. The focus on computation, at the expense of other issues, set a tone for humanities computing (and subsequently Digital Humanities) where computers are an “instrumental tool” for the study of texts (Svensson 2009). Such an attitude ignores the many other ways that humanities computing scholars used technology in novel and innovative ways.

Early experiments in informal scholarly communication begin show how the humanities computing community sustained itself and, as we’ll see, eventually grow into the larger community called digital humanities.

> Humanities computing representatives were also early adopters of communication technologies such as email lists. The first message on the Humanist List was sent on May 13, 1987 by founding editor Willard McCarty, making it one of the first academic email lists to be established (Svensson 2009).

The HUMANIST mailing list has served–for decades–as a place for working out their collective scholarly identity far outside the existing structures of power and prestige in the academy.[4] The list has been an important safe space for establishing solidarity in the face of marginalization by the larger humanities community. Svensson (2009) quotes McCarty in an email to the list where he notes,

> We’re always worrying ourselves about whether humanities computing has made its mark in the world and on the world. It seems to me, however, that quiet change, though harder to detect, is sometimes much better and more powerful in its effects than the noisy, obviously mark-making, position-taking kind. If during these 17 years Humanist has contributed to the world, it has done so very quietly by nature, like conversation, leaving hardly a trace.

The notion of “quiet revolutions” is a powerful statement about how humanities computing relates to the academy. McCarty is wrong about one thing however; the Humanist mailing list *has* left a trace its archives, which are openly accessible on the web.[5] Very recently digital humanities software developer David McClure

[4] It may even be one of the first academic mailing list, although I am not certain about this fact.

[5]The Humanist discussion group has a complete archive hosted on their website http://dhhumanist.org/

posted a visualization of the Humanist mailing list on the web[6] and wrote a post about the project on his blog.[7]

For humanities computing, technology and computation was mainly a tool or instrument extending the scale and scope of research. Yet, as the existence of the HUMANIST list shows, information technology has played a role in how the community assembled and managed itself as a coherent social unit. Yet, their use of new modes of informal scholarly communication hasn't been seen as a defining characteristic of the community.

Digital Humanities: Technology > Tool

As the saying goes, ask ten scholars "what is digital humanities" and you'll get eleven different answers. Positioning and pontificating about the digital humanities is a genre of scholarship in and of itself (Kirschenbaum 2010; Gold 2012; Burdick et al. 2012; Berry 2012; Terras, Nyhan, and Vanhoutte 2013; Jones 2013; Klein 2014). These debates about the ontological status of the digital humanities have shown it to be a complex community of practice whose essential nature has yet to be determined. However, central to these debates is digital humanities' unique relationship to technology.

In surveying the "landscape of the digital humanities" Svensson found a diverse and evolving relationship between digital humanities and technology (2010). Technology is not a deterministic instrument, but instead as five distinct *modes of engagement*:

1. As a tool
2. As a study object

[6] The visualization is hosted on McClure's website http://humanist.dclure.org/ and the code is up on Github https://github.com/davidmcclure/humanist

[7] “"Visualizing 27 years, 12 million words of the Humanist list” The project was more a demonstration of a text mining and visualization technique than an analysis of the discussion group, but it shows what kinds of research are afforded by open communication on the web. http://dclure.org/essays/visualizing-the-humanist/

3. As an experimental laboratory
4. As an activist venue
5. As an expressive medium

Humanities computing embraced technology as a *tool*, digital humanities uses technology as much more. This is not to say tools are no longer a loci of research, scholars are still developing technologies for analyzing text such as Stéphan Sinclair and Geoffrey Rockwell with Voyant Tools.[8] Digital humanities also critically engage the technologies they use as *study object* in and of itself. Matthew Jocker's in *Macroanalysis* (2013), and Stephen Ramsay's *Reading Machines* (2011) situate computation in a humanistic context *and also* explore how algorithms and computation function in the production of knowledge. The practical and theoretical issues surrounding digital scholarship and it's epistemological standing are still being negotiated, hence Svensson's characterization of technology as an experimental laboratory. Within such a framing Digital Humanities can be though of as what Karin Knorr Cetina (1999) calls *epistemic cultures*, or "amalgams of arrangements and mechanisms…which, in a given field, make up how we know what we know. Epistemic cultures are cultures that create and warrant knowledge." (p. 1). Digital humanities, as both an *experimental laboratory* and *activist venue*, play a significant role in critically examining what counts as scholarship and advancing the acceptance of its digital forms.

Each mode of engagement is deeply intertwined, but most relevant to this project is how digital humanities scholars use technology as an *expressive medium,* particularly using technologies afforded by the Open Web for the purposes of informal scholarly communication. While the period of humanities computing, "Humanities 1.0" was innovative in their modes of communication, for example the Humanist mailing list, the digital humanities over the past decade can be

[8] http://voyant-tools.org

distinguished by a use of blogs at a size and scale unwarranted in their previous history.

Digital humanities scholars are as much about social media as they are about digital objects and methods. While we see important contributions to the self-actualization of digital humanities in formal publications (Berry 2012; Burdick et al. 2012; Gold 2012; Terras, Nyhan, and Vanhoutte 2013), there are equally important conversations about the digital humanities occurring in informal discursive spaces online. Focusing exclusively on formal publications misses a significant portion of digital humanities scholar's reflective discourse.

McPherson (2009) claims "blogging humanists" have bloomed by taking advantage not only of technology's computational abilities, but its communicative power as well. This category, which leverages Web 2.0's interactional capacity, has gained significantly more attention than it's computing humanist forebears. Blogging humanists, or "Humanities 2.0," (Davidson 2008) are novel not only for their enthusiastic adoption of blogs in scholarly communication, but for their provocative relationship with established academic institutions, hierarchies, and social orders.[9]

Blogging Humanities

With the advent of the web, scholarly blogging has become a crucially important communicative platform for long-form argument and discussion with unparalleled visibility and pace, no more so perhaps than in the digital humanities. According to Kirschenbaum (2010):

> Whatever else it might be then, the digital humanities today is about a scholarship (and a pedagogy) that is publicly visible in ways to which we are generally unaccustomed, a scholarship and pedagogy that are bound up with infrastructure in ways that are deeper and more explicit than we are generally accustomed to, a

[9] McPherson also posits the emergence of a third form of humanist, the "multimodal" humanist who "brings together databases, scholarly tools, networked writing, and peer-to-peer commentary." This vein of humanist scholar is now starting to emerge; this research project might potentially be considered part of this group.

> scholarship and pedagogy that are collaborative and depend on networks of people and that live an active 24/7 life online (p.6).

Unlike nineteenth-century "men of letters," the back-and-forth of these "women and men of blogs" become available to everyone, immediately. Leveraging the technical affordances of "web publishing;" the so-called "invisible college" has become substantially less invisible (Halavais 2006).

Blogging is a popular tool with digital humanities scholars. As of this writing there are 705 blogs listed in the Compendium of Digital Humanities (more on the Compendium in chapter 3).[10] Blogs invert traditional dynamics of social visibility and access in the academy. Those who choose to participate become visible and available with unparalleled ease. The network effects, where the value of a network increases in proportion to membership, grow as scholars join the conversation. In a discursive sphere where anyone can post their writing online, the challenges of scholarly communication flip from scarcity to surfeit. Given limited attention what should a scholar read? How does one find contributions of quality and value? Attending only to formally published and peer reviewed scholarly literature is one answer, but a digital humanist, as Kirschenbaum noted, ignores the blogosphere at his or her own peril. To ignore digital humanities blogs is to risk falling out of touch with the community.

Social Studies of Digital Humanities

While there have been efforts to understand the digital humanities holistically (Gold 2012; Berry 2012; Klein 2014), these efforts have been initiated and undertaken by practitioners themselves. As an interdisciplinary community of practice or epistemic culture, the books have been decidedly disciplinary and mainly written by English scholars. These studies are hermeneutic or even anecdotal in their rhetoric, which is not to say they are bad or wrong, but only that they do not necessarily bring the full scope of digital humanities to bear upon

[10] https://docs.google.com/spreadsheet/pub?hl=en_US&hl=en_US&key=0AucqXAIBhf_idGNlZzVjSGkxQU9XNU4ybow1clMxeXc&single=true&gid=3&output=html (Accessed March 14th, 2015)

digital humanities. There are very few large-scale, data drive, and systematic studies of the digital humanities as a phenomenal object. Bowman et al. (2013) observe a "thorough investigation and description of the communicative practices of DH is lacking" and began the only systematic, data driven, social scientific and bibliometric study of the digital humanities scholarly communication. Their study explicitly "examines informal and formal communication channels used by members of the DH community to diffuse information and build communities." Unfortunately, the results of their study have not yet been published either formally or informally.[11]

Given how active digital humanists are online, the various traces of their online activity provide substantial basis for analysis. To date, data driven analysis has not been applied to study the digital humanities. There have been a couple cursory investigations, but they are more methodological demonstrations or playful inquiries making no general claims about the field as a whole. Three that stand out are Melissa Terras' infographic charting the growth of the digital humanities,[12] Matthew Jocker's topic modeling the Day of Digital Humanities blogs,[13] and Elijah Meek's experiment with topic modeling a corpus of digital humanities definitions.[14]

[11] The study only exists as an extended abstract and presentation at the 2013 Digital Humanities conference.

[12] http://melissaterras.blogspot.com/2012/01/infographic-quanitifying-digital.html

[13] Jockers, Matthew. "Who's Your DH Blog Mate: Match-Making the Day of DH Bloggers with Top Modeling"" http://www.matthewjockers.net/2010/03/19/whos-your-dh-blog-mate-match-making-the-day-of-dh-bloggers-with-topic-modeling/

[14] https://dhs.stanford.edu/comprehending-the-digital-humanities/

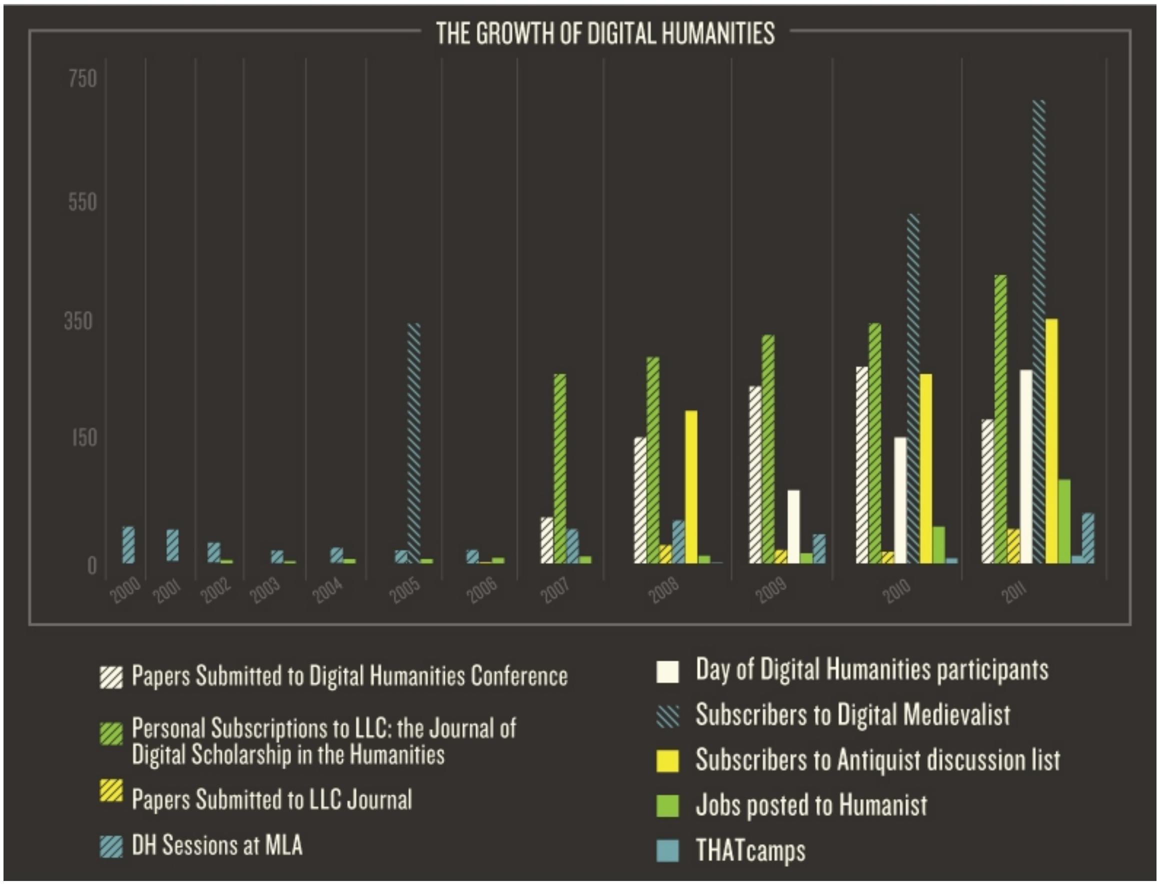


Figure (1). A section of the infographic produced by Melissa Terras charting the "Growth of the Digital Humanities."

Melissa Terras showed the growth of the digital humanities in a beautiful, data rich visualization. Figure (1) visualizes a portion of the infographic charting nine different growth indicators from 2000 to 2011. The chart illustrates the growth of the digital humanities over the past decade through submissions to journals, sessions at conferences, subscribers to mailing lists, and jobs posted. The graphic as a whole reveals the global distribution of research centers, participation in social media, readership of journals, and funding by agencies and foundations. All told, the infographic represents a compelling argument that the digital humanities is a quickly growing field.

Matthew Jockers analyzed a small corpus of posts from the annual "Day of DH" blogs. For the past four years, the DH community has collectively organized an effort to blog about "what digital humanists do." This results in an aggregated collection of blog posts documenting self-accounts of their activities for the day. From a corpus of 117 of these blogs, Jockers generated a set of ten “topics,” or clusters of words, that represented common things digital humanists wrote about. In the end, Jockers didn’t find much from his cursory analysis:

> Unfortunately, the Day of DH corpus isn’t truly big enough to get the sort of crystal clear topics that I have harvested from much larger collections, but still ... seen in aggregate, do give us a sense of what’s “hot” to talk about in the field.[15]

Further work distantly reading the digital humanities has been done by Elijah Meeks, a former “digital humanities specialist” at Stanford University.[16] Like Terras and Jockers, Meeks blogged (as opposed to formally published) somecursory quantitative work using digital humanities texts as “data.”

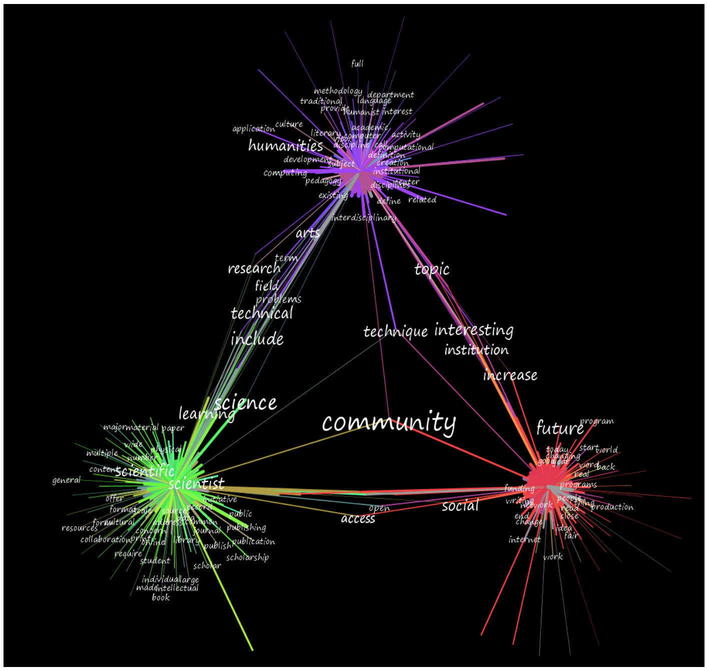


Figure (2). A network diagram of Eljiah Meeks’s topic model of Digital Humanities definitional works.

[15] Jockers, Matthew. “Who’s Your DH Blog Mate: Match-Making the Day of DH Bloggers with Top Modeling ”" http://www.matthewjockers.net/2010/03/19/whos-your-dh-blog-mate-match-making-the-day-of-dh-bloggers-with-topic-modeling/

[16] https://dhs.stanford.edu/comprehending-the-digital-humanities/

Drawing on a very small corpus of fifty texts defining digital humanities/humanities computing, he analyzed them, also with topic modeling, and created visualizations of the results. Meeks went one step further producing a series of interesting images representing the clustered topology of his small corpus. The results are quite beautiful, but, given the limitations of the corpus, do not necessarily contribute to a general understanding of the digital humanities. Meeks's underlying motivation was to explore the textual analysis and visualization tools, rather than generate substantive insights into the digital humanities.

The preliminary work in this area demonstrates a need for more systematic and rigorous research on digital humanities as a field of scholarship. The increasing growth and complexity of the digital humanities requires commensurate increase in complexity of analysis and reflection. Techniques like topic modeling provide powerful aids for new understandings of the field, but the size and constitution of a corpus are important factors when using such a method.

In her call to action, Borgman (2009) argues, if the digital humanities is to transition from a "specialty field" into a substantive and broadly recognized scholarly community they need to address several important questions: What are data? What are the infrastructure requirements? What is a humanities laboratory? What is the value proposition? And most importantly, *where are the social studies of digital humanities?*

> Why is no one following digital humanities scholars around to understand their practices, in the way that scientists have been studied for the last several decades? This body of research has informed the design of scholarly infrastructure for the sciences. … The humanities community should invite more social scientists as research partners and should make themselves available as objects of study. In doing so, the community can learn more about itself and apply the lessons to the design of tools, services, policies, and infrastructure (para. 76).

In the spirit of science studies, digital humanities studies could provide an important reflexive perspective to the ongoing maturation of the field. However the science studies approach should be careful not "scientize" the digital humanities. The methodological and theoretical frameworks borrowed from science studies will be crucial to advancing the field (Liu 2013). As Liu argues, “[science and technology studies] is another method for knowing meaning in the digital humanities,” The classic methods of science and laboratory studies were been extremely effective at surfacing the invisible work and labor of scientific knowledge production (Star and Strauss 1999), but as I’ll discuss below, they are not as effective at uncovering the digital knowledge practices (Bowker et al. 2010). Social scientists interested in studying the digital humanities should meet humanists on their own terms. This means meeting them online in the digital spaces they occupy in addition to the material world.

Meeting the digital humanities on their terms means meeting them on blogs. Both McPherson (2009) and Kirschenbaum (2010) have highlighted the innovative use of communications technology as important factors in the constitution of the community. Studying scholarly blogs present an opportunity to understand both the digital humanities and new modes of scholarly communication.

Scholarly Communication

Scholarly Communication is a broad category encompassing the forms of communication unique and distinctive to academics. Scholarly communication, broadly conceived, is the myriad of forms by which scholars interact, socialize, and disseminate knowledge (Borgman 2007). While scholars are often the central focus of this area of research, societies, publishers and libraries are also key facets in the full “lifecycle” of scholarly communication.

In 1964, Garvey and Griffith introduced a model of scholarly communication rooted in studies of psychology. Their model greatly influenced subsequent framings of scholarly communication:

> It outlined the process by which research is communicated and provided details of the various stages within a time frame encompassing from initial concept to integration of the research as an accepted component of scientific knowledge. Although the time scale varied from one discipline to another, the essential elements of the model appeared to be universal (Hurd 2000).

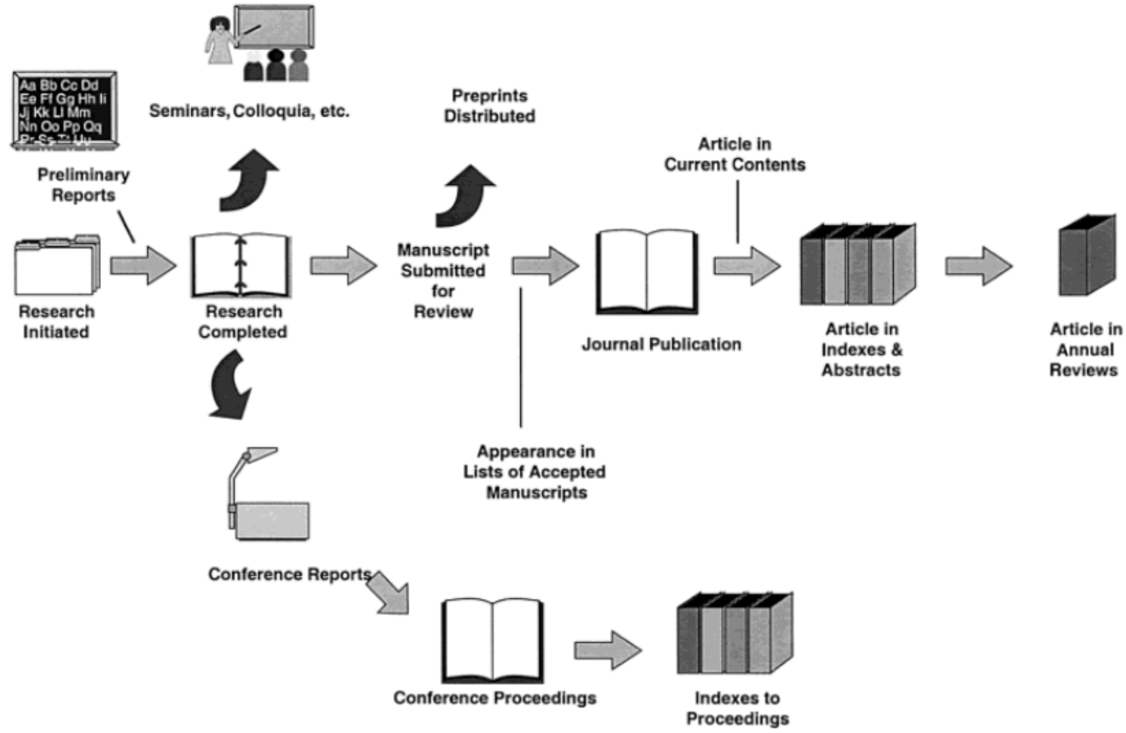


Figure (3). The Garvey Griffith model as represented by Hurd (2000)

The Garvey and Griffith model as represented in Figure (3) characterized a process by which ideas are transformed into knowledge through a series of stages and processes (Garvey and Griffith 1964; Borgman 2007; Hurd 2000). This model set a tone for a linearized characterization of scholarly communication that focused upon the *artifacts* or products of scholarly communication (Lievrouw 1988).

The focus on published artifacts and products lead to the study of scholarly communication in the form of *bibliometrics*, an approach whose goal was articulated in some of the field's earliest work by Pritchard as:

> To shed light on the processes of written communication and of the nature and course of development of a discipline (in so far as this is displayed through written communication), by means of counting and analyzing the various facets of written communication (Pritchard in Borgman 1989, 585).

The focus of bibliometrics is twofold. First, it examines "written communication," and implicitly the standard genres of scholarship, books and periodicals. Second, bibliometrics is a quantitative approach that prioritizes the countable and enumerable “facets” of scholar’s “written communication.” In the late 60s, before computer and text became available, the most countable facets of scholar’s written communication were citations.

Bibliometrics, informetrics and scientometrics are rich and complex fields of study too large to cover here. The practice of measuring scientific products began with Garfield’s (1955) seminal paper on the *science citation index*. The field exists because of Garfield’s metric, the *impact factor* and other measures like the h-Index (Hirsch 2005) and because of bibliographic databases like Scopus and Web of Science. In the last few decades the availability of digital data, in the form of peer reviewed citation indexes sold by scholarly publishers, has been the governing factor of growth and development in the field (Wouters 1999). This has lead to a focus on studying scholarly interactions via citations and references in *formal* publications (Garfield 1979; Borgman 2000; Borgman and Furner 2002).

The structure of the Internet affords access to scholarly communication as if the web was its own massive database of scholarly citations and references (Cronin 2001a; Thelwall and Wouters 2005; Wouters and Vries 2004). Yet, new genres and formats have their own dynamics (Cronin et al. 1998) and the relational assumptions of formal citations behavior don’t map to web linking behavior (Wilkinson et al. 2003). The web throws into sharp relief an important distinction between *formal* and *informal* scholarly communication.

Formal and Informal Scholarly Communication

Borgman, in her extensive synthesis of the field, argued formal scholarly communication has three functions, *legitimization*; *dissemination*; and *access, preservation, and curation* (2007). Legitimization is the function of review that establishes, to use a term common in the literature, the “trustworthiness” of the knowledge and information, typically via some form of peer review. Dissemination circulates that knowledge through channels of publications, as conferences, journals, or books. Finally, access, preservation, and curation, save the scholarly record and make it available in perpetuity. However, the *formal* system of scholarly communication is not the focus of this project. What has been given less attention in recent year are the functions of *informal scholarly communication*.

Informal scholarly communication has, by its very definition, not been the subject of bibliometrics, but it has been theorized and conceptualized within the various models of scholarly communication. It has always had a kind of implicit status in the literature as a kind of thing everyone knows about, but can't scientifically analyzed because it has been, at least historically, difficult to access and measure.

Garvey and Griffith (1968) distinguish between formal and informal channels of scholarly communication. In their characterization, the primary distinction between the two channels revolved around *temporality* and *visibility*:

> Formal and informal channels can be generally differentiated in terms of two gross characteristics of their products: *Formal channels* carry information which is public and remains in permanent storage; *informal channels* carry information to restricted audiences and its storage is relatively temporary (p. 131) .

Formal channels are "*public*"[17] and *permanent*; they have been published, in a journal or book, and preserved in a library or archive. Communication passing through formal channels becomes part of the scholarly record. Informal channels on the other hand are only available to a "*restricted audience*" for a short amount of time.

In his book, *Communicating Research*, Meadows also highlights these same distinctions between formal and informal. Formal scholarly communication is published, preserved, and archived to be "available over long periods of time to an extended audience" whereas informal scholarly communication is "often ephemeral and … made available to a restricted audience only" (Meadows 1997, 7). The formal have been archived and preserved in institutional repositories, in academic libraries, or across publisher databases. These materials are then

[17] To those with access to academic libraries.

subsequently made more available in space and time–albeit through academia's inherent gatekeeping.

Informal scholarly communication is, by these definitions, a short-lived and fleeting mode of discourse. It lives in the ephemerality of after-hours conversations or the restrictive access of private letters or email. Yet, the importance of informal scholarly communication should not be ignored. The visibility and temporality of informal scholarly communication enables other kinds of scholarly work.

Functions of Informal Scholarly Communication

This section introduces a framework for understanding the importance of informal communication in scholarly communities. Herbert Menzel (1968) compiled a set of six functions of informal scholarly communication: *promptness; selective switching; screening, evaluation, and synthesis; extraction of action implications; transmitting the ineffable; and instantaneous feedback*. These functions address the needs of scholars that are not currently met by formal scholarly communication. Scholars would not be able to function without informal scholarly communication.

Garvey and Griffith's 1968 classic model is too *linear* for capturing the complexities of informal communication, which is harder to track and trace. Menzel's model is significant for this project because it provides a *functional* model of scholarly communication. The functional model is useful for this project because it provides a framework for analyzing *content*, rather than structural relationships and processes.

Promptness speaks to the speed at which scientists can become aware of new discoveries via the interpersonal communications of their "invisible college" (discussed below). Formal publication cycles took months and years; hence the manifestation of preprint circulation networks to circulate papers within scholarly communities. Some disciplines, namely computer science, have adopted a culture of conference papers instead of journal articles or books. The "deadline" driven publication cycle has been criticized for incentivizing "least publishable unit" at the expense of quality scholarship (Fortnow 2009). The costs and benefits of a conference paper culture is a matter of much discussion (Korth et al. 2008; Vardi 2009; Vardi 2010).

Selective switching leverages interpersonal networks by "rout[ing] scientific news to the scientists to whom it is relevant" (156). This mechanism has two aspects. First, selective switching is more effective than formal classification and indexing because such systems are always more granular than the set of disciplinary peers who have a more dynamic sense of information needs. Second, peer-to-peer networks can circulation information that a scholar might not realize is relevant.

Screening, evaluation, and synthesis are, according to Manzel, the function of informal communication that delivers information, rather than documents, in response to a query. Peers can distill journal articles or books and tell you what they are "about," better than a librarian or information retrieval system. Menzel and the other scholars writing about scholarly communication in this period were heavily influenced by mid-century notions of information as an independent artifact (Lievrouw 1988) that circulates as a commodity through formal and informal channels. Contemporary understandings of information do not separate information from the social and technical contexts of its creation and use (Buckland 1991; Brown and Duguid 1996; Brown and Duguid 2002).

Extraction of action implication refers to the act of moving information across the epistemic boundaries away from research towards use, for example, theoretical research being translated into implications for design. In Menzel's conceptualization, this is a translation of *language* from "basic" into "applied" science. While this movement can occur through formal channels, the informal channels "can add … a sense of judgment as to its significance in giving practical situations" (159). Menzel's formulation here could be seen as a precursor to the richer and more nuanced concept of *trading zones* developed by Peter Galison (1997).

Transmitting the ineffable is an important function because it transmits tacit knowledge, “know how” or “the "minor details” that don’t make the cut for formal publication due to publication standards or disciplinary expectations. The ineffable may be “unduly lengthy and laborious ...[or simply] trivial” (160). Informal communication channels provide an avenue for the circulation of this kind of information. The qualitative and ethnographically oriented work of laboratories studies has uncovered the importance of tacit knowledge in scientific practice (Collins 1974; Latour and Woolgar 1986; Lievrouw 1989).

The final function of information scholarly communication is *instantaneous feedback,* which provides opportunities for criticism and clarification at a much faster pace than formal print publication cycles. The slow pace of formal publishing has, as noted above, resulted in the emergence of pre-print circulation networks to bring the dissemination of knowledge in line with the practices that produce it.

Unlike formal channels, informal scholarly communication has not historically left material or documentary traces behind for researchers to analyze. It is, at least in part, due to problems of visibility that prevents research beyond micro level ethnographic investigations. Historically, researchers could hypothesize about invisible communication or look for their latent patterns in citation networks (Lievrouw 1989).

Menzel’s six functions provide a framework for evaluating blogs as a unit of analysis. I use these six functions to provide a scaffolding for understanding how digital humanities scholars use blogs as channel for informal scholarly communication.

The Invisible College

The digital humanities, as represented by a collection of blogs and bloggers, constitute an “invisible college.” However, because of the technical affordances of blogs a portion of the community’s communication is *visible*, yet, it is more in common with the informal networks and discourses of the invisible college than those of formal scholarly communication (i.e., books and periodicals). The contemporary notion of the invisible college emerged during the same period as Menzel's six functions of informal scholarly communication, and similarly, provides a frame for thinking about blogs and scholar's discursive practice online (Halavais 2006).

The term invisible college, as every discussion of the term points out, has Rosicrucian origins in the mid-seventeeth century's loose assemblage of natural philosophers, the Royal Society of London (Kronick 2001). Today, the term has a slightly more modern, definition:

> For each [discipline] there exists a sort of commuting circuit of institutions, research centers, and summer schools giving them an opportunity to meet piecemeal, so that over an interval of a few years everybody who is anybody has worked with everybody else in the same category. Such groups constitute an invisible college (Price 1963).

Every scholar eventually learns about the invisible, peer-to-peer networks of communication that keeps the gears of their disciplines turning. A PhD is in many ways a process of conditioning individuals into the specific social and organization patterns of knowledge work within an academic field. Accessing these networks has, historically, been difficult and large-scale studies had to approximate their latent structure.

Diana Crane operationalized the idea of an invisible college as a social network of scientists in her book, *Invisible Colleges* (1972). Crane leverages sociometric techniques to identify latent social structures within networks of scientific authorship. Her work was an empirical test of Price's speculations on "invisible colleges." Crane empirically inferred the existence of latent social organization through questionnaires and co-citation analysis. As to whether this social organization constituted an "invisible college" as Price conceptualized, Crane had caveats:

> Price has popularized the term "invisible college" which refers to an elite of mutually interested and productive scientists. ... However, this conceptualization does not comprehend ... (a) the interaction between the most active and influential members of the [research] area and the "rank and file" and (b) the role of "outsiders" in the organization of the [research] area (Crane 1969).

While Crane's work and the idea of invisible colleges is old, its influence upon the study of scholarly communication cannot be denied. In 1989, Leah Lievrouw re-evaluated the concept in an effort to address the difficulty in operationalizing the content and an ambiguity between studies of the *structure* versus the *process* of scholarly communication.

> The invisible college construct reflects a recurrent problem in the social studies of science generally, which tend to examine the *products* of science (e.g., artifacts such as public documents) in order to understand the social *process* of science, which are essentially communicative in nature (e.g., interpersonal contact) (Lievrouw 1989).

This conflation of process and structure in invisible college highlights the tension between informal and formal modes of scholarly communication. Formal scholarly communication is composed of a fixed artifice of legitimized, disseminated, and preserved *inscriptions*, to use Latour and Woolgar's term (1986), of scholar's processes.

While the invisible college has been a useful conceptual model, the transformations occasioned by the web over the past three decades render visible the previously invisible, latent networks of scholar's informal communication. Scholars intellectual workings are now, empirically accessible as opposed to latent and theorized.

Scholarly Communication & the Web

> One of the by-products of the digital revolutions and the Internet has been the creation of huge ***informal*** [emphasis added] repositories of public access, easily discoverable information including the web and newsgroups...large sections of the information come from genres that have previously been inaccessible to researchers in any quantity, and may therefore help to address research questions that have been previously left unanswered (Thelwall and Wouters 2005, 188).

This quote highlights important points about *access* and *scale*. Formal bibliographic databases only provide researchers access to (some) peer reviewed scholarly communication. The web, from this perspective, is a vast database to be explored and analyzed (Blaise Cronin 2001a; Wouters and Vries 2004; Thelwall and Wouters 2005) replete with new genres and forms that are native to the web as a communications platform (Cronin et al. 1998).

Scholar’s use of the Internet and the Web challenges older conceptual models of scholarly communication. Activities that have previously been private and invisible are now public. While not all networks of informal communication are rendered visible by the web, preprint archives, blogs, and social media have opened up a “third place” of scholarly communication (Halavais 2006). Private and invisible communication hasn’t gone away, rather there has been a dramatic increase in public and visible communication on the web.

> What has changed most since the days of print and post is the balance between public and private communication. Conversations that previously were oral are now conducted by e-mail or online discussion lists, sometimes leaving a public record for a long period of time (Borgman 2007).

Information technology has thrown long held categories of distinction into disarray. One of the most prominent networks of scholarly communication, arXiv, challenges the distinction between formal and informal scholarly communication. arXiv began as an “e-print archive” for circulating preprint journal articles in high energy physics in response to “inadequacies” of formal journals (Ginsparg 1994). arXiv sits in an interstitial space between the formal and the informal.

> The site has never been a random UseNet newsgroup or blogspace-like free-for-all. From the outset, arXiv.org relied on a variety of heuristic screening mechanisms, including a filter on institutional affiliation of submitter, to ensure insofar as possible that submissions are at least “of refereeable quality” (Ginsparg 2007).

There is some measure of review, but it “operates at a factor of 100 to 1000” times cheaper than traditional formal peer review. The arXiv model of scholarly communication *delegates*, in the sense of Ribes et al. (2013), the work of legitimization and trust to individual scholarly communities and focuses instead on providing *access* to scholarly materials.

Providing easy access to the *full text* of scholarly materials is important because studies show a FUTON (FUll Text On Net) bias for materials that are easily accessible. As the systems of distribution and access have scaled up, the traditional systems of peer review have been left behind. This raises questions about new systems of authority (Jensen 2007; Fitzpatrick 2010) that take better advantage of new information technologies. Systems like *community-oriented, post-publication peer-to-peer review* transform the review "from a process focused on gatekeeping to one concerned with filtering the wealth of scholarly material made available via the Internet" (Fitzpatrick 2010, 161). Such a modality of review is built on *scalable* information systems and processes with few to no barriers for "publishing."

The model of scholarly communication embodied by arXiv reflects the "tribal customs" (Cronin 2003) of a particular scholarly community. It is a system whose norms, values, and practices are rooted in the discursive practices of high-energy physics. While arXiv has expanded and now supports communities outside of high-energy physics, the humanities are conspicuously absent from their taxonomy. The humanities, as Cronin points out, have a radically different communication culture centered on the monograph rather than periodicals. The circulation of preprint materials is much more difficult for scholars under restrictive book contracts and conservative University Presses (Waters 2004). So, what happens when an unstoppable force (digital humanities) reaches an immovable object (scholarly publishing)? Blogs!

The Open Web

The World Wide Web was originally designed as a platform for the "management of general information about accelerators and experiments at CERN," that is, a platform for managing the massive amounts of interconnected information, systems, and people in large, complicated physics experiment (Berners-Lee 1989). The Web was a system for managing the *informal* coordination and *communication* of *scholars* in high-energy physics. While working as a software consultant at CERN, Tim Berners-Lee created software that combined hypertext documents with computer networking. The rest, as they say, is a complicated history of people, technology, institutions, politics, and economics. The history of the Internet (Abbate 2000) and the web (Berners-Lee 2000; Gillies and Cailliau 2000) is far outside of the scope of this project. Instead, I want to draw attention to the *Open Web* as a technological frame for understanding the digital humanities, blogs, and scholarly communication.

The Open Web is not clearly defined. Dave Winer argues the Open Web means systems and infrastructure where information can be easily moved around.[18] Brad Neuberg also defines the Open Web as a place where “everyone can share information, integrate, and innovate without having to ask for permission.”[19]

These definitions focus on the ethical and ideological characteristics of the Open Web. This ethos of “openness” is rooted in the design and the architecture of the web. But, these definitions do not provide a suitable *technological* frame in and through which blogs can be understood.

The Technical Open Web

The technological aspects of the Open Web are often described as the *Open Web Platform*.[20] The Open Web platform refers to the technical and institutional structures that guide the design, operation, and maintenance of the web.[21]

The Open Web platform is designed and maintained with a set of technical principles to preserve its “openness.” The most widely used definition of these principles comes from a blog post by technologist and computer scientist Tantek Çelik titled “What is the Open Web?” (2010). Çelik claims the Open Web enables three sociotechnical practices:

[18]Winer, Dave. “What I mean by ‘the'the Open Web.’”"
http://scripting.com/stories/2011/01/04/whatIMeanByTheOpenWeb.html

[19] Neuberg, Brad. “Open Web Definition (Version 0.4)”
http://codinginparadise.org/weblog/2008/07/open-web-definition-version-04.html

[20]A list of the main Open Web standards is maintained by the World Wide Web Consortium here: http://www.w3.org/wiki/Open_Web_Platform

[21]Paul Ford, a writer and programmer, has written a very nice history of the social, technical, and organizational actors in “The Group that Rules the Web.” He is also working on a book about the history of the web. http://www.newyorker.com/tech/elements/group-rules-webFor our purposes the “open web” and the “open web platform” are synonymous.

1. **Code and implement** the web standards that that content/apps depend on
2. **Access and use** content / code / web-apps / implementations
3. **Publish** content and applications on the web in open standards

These practices are afforded by three infrastructural dynamics: *open standards*, *open access*, and *open publishing*.

Open Standards

Open standards are the sociotechnical underpinnings of the web. The most important standards of the Open Web are HyperText Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript (JS). While there are many standards involved in the web for a discussion about blogs the other important standard is Really Simple Syndication (RSS). These standards collectively make it possible to read, browse, comment upon, stay-up-to-date-with, and find blogs using different browsers, applications, and devices.

Interestingly, blogs are *not* a web standard but more of a convention of *expectation*, *structure* and *design* (Rettberg 2013). By convention, blogs are composed of individual posts written in reverse chronological order. The RSS standard defines a way of representing new content, like blog posts, as a machine-readable *feed*. RSS allow tools, namely *feed readers*, to “subscribe” to a blog allowing an individual to efficiently follow and read many blogs. RSS became the *de facto* standard for blogs because it defined a standard means of *syndicating* blog content across the web.

Open Access

Open access is a fundamental principle of the web’s network architecture. The idea of *open access* specifies, at its simplest, that resources (web pages) located at a particular Universal Resource Locator (URL) should be freely and openly accessible to any computer requesting that resource. From an infrastructural perspective, open access means that any endpoint can communicate with any other endpoint on the network without interference. This property is related to the hotly debated topic of *network neutrality* (Wu 2003).

Open access has another, more important, meaning with respect to scholarly publishing. In this context, open access is more than simply unfettered access to a resource located at a URL, it invokes complex social dynamics related to economics, policy, and intellectual property. According to Peter Suber, who literally wrote the book on open access, "Open access (OA) literature is digital, online, free of charge, and free of most copyright and licensing restrictions" (Suber 2012, 4). For Suber, open access is not merely the technical retrieval of content, but the legal freedom to use and share that content without the encumbrance of permissions from its owner.

Open access with respect to the Open Web does *not* have to be "free of most copyright and licensing restrictions." In the purely technical conceptualization of open access, content need only be "digital, and online, [and] free of charge.". While bracketing the copyright and intellectual property issues ignores some of the most contentious and important aspects of open access, the freedom and openness implicated in "free of most copyright and licensing restrictions" is not a technical requirement of the Open Web, nor is it a social expectation of content published on the web.[22]

Open Publishing

Where the premise of open access posits the ability to freely and programmatically request information from anywhere on the Open Web, open publishing posits the reciprocal. Open publishing means anyone, anywhere, can publish information on the web. However, unlike open access, open publishing doesn't mean anyone can put anything anywhere, rather it means anyone can claim their own piece of web real estate and publish whatever they want. Open publishing is also *distributed* publishing. It is a lot faster and cheaper to register a

[22] Most of the blogs I have scraped do not have free and open copyright licenses. Because my data are textual this makes sharing much more complicated than purely numeric data. For these reasons, I am not planning on posting my data publicly online.

domain and host website than it is to publish an op-ed in the New York Times online. Gatekeeping and editorial control, which do regulate what gets posted to certain portions of the web, are domain specific organizational functions, not structural features of the system itself.

This means the word *publish* has lost some of its semantic weight. Web publishing, in contrast to scholarly publishing, means making information accessible on the web and does not imply editing, peer review, preservation, and all the other functions of print-centric publishing. This is an important distinction because it delineates *formal* from *informal* scholarly communication.

The Open Web provides the infrastructural milieu in and through which the digital humanities, blogs, and informal scholarly communication can be understood. The technical properties of the Open Web specified the conditions of possibility for, but do not determine, how the web is used. Blogs are one realization of these possibilities.

Blogs

The term "blog" is a portmanteau of "web" and "logs" and those two words do a reasonable job describing the gist of blogs. Blogs are logs of a variety of stuff published on the web. Blogs are not *formally* defined, but are rather a collection of conventions that emerge out open publishing and open access's lack of control, and web designers composing documents with open standards. There is no W3C standard for blogs, there is only the ongoing practice of blogging and the platforms and products built to support that practice. Blogs have their origin in online, web diaries (Serfaty 2004; Rettberg 2013) and some of their best histories have been written not by academics, but by early bloggers themselves (Rodzvilla 2002; Rosenberg 2010).

In *Blogging*, Rettberg (2013) dedicates 30 pages to answer the question "What is a blog?" (2013). The answer, as is any academic treatment of a subject, is complicated. Blogs can be defined by their platform, i.e., websites published on Wordpress.com or Blogger.com. Blogs can be defined by conventions of design and format, i.e., the basic unit of content is a “post,” which are listed in in reverse chronological order, may or may not have comments, can have tags and categories, and are distributed by RSS feeds. Furthermore blogs can be defined by their content. According to Rettberg blog have three styles of content, personal blogs that provide narratives of an individual’s lived experience; filter blogs, which aggregate and curate links from across the web; or topic-driven blogs, which focus on sharing information, via links or narrative, about a particular topic of interest.

The question “what is a blog” was the subject of recent debate because 2014 is seen by many as the 20th anniversary of blogs. Justin Hall started the first blog in 1994 (Rosenberg 2010), so naturally the Guardian thought it appropriate to do an article interviewing early bloggers on the 20th anniversary.[23] On Twitter Ian Bogost raised the question, can we *really* be celebrating the birth of blogs? The blog, in his view, is deeply intertwined with the technical platforms like Blogger and WordPress and those infrastructures did not exist in 1994.[24] The definition, in Bogost view, is sociotechnical. However, as early blogger Anil Dash pointed out, the early design of blogging platforms accommodated the *existing* practices of early bloggers who had built blog-like structures by hand.[25] Peter Merholz, the web developer who coined the word blog, argued the Mosaic Communication Corporation’s “What’s New” page[26] from June of 1993, should be considered a blog, to which Bogost responded “no.”[27]

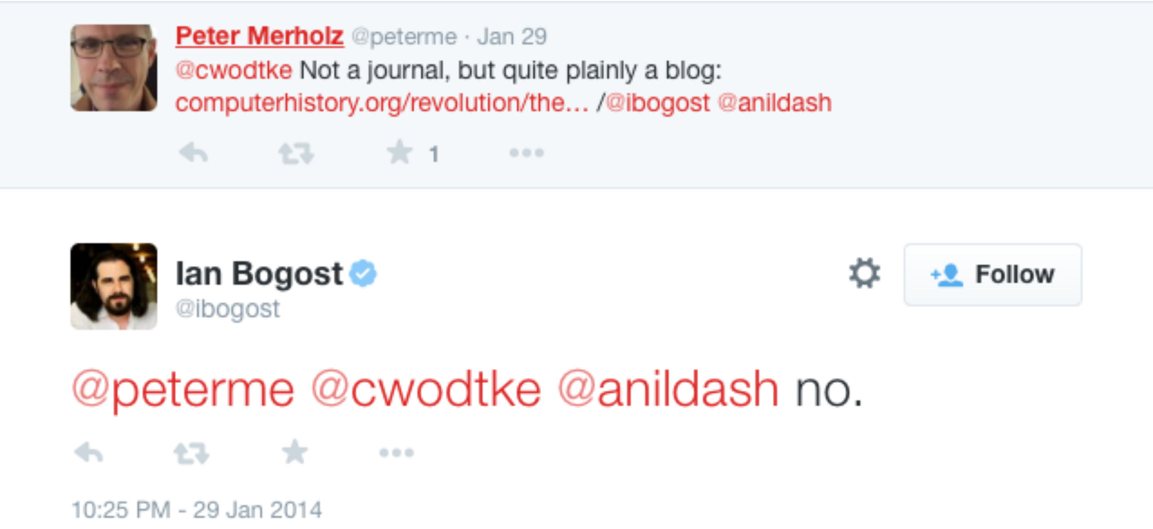
[23] Rogers, Katie & Spencer, Ruth. “The blog turns 20: a conversation with three internet pioneers” http://www.theguardian.com/media/2014/jan/29/blog-turns-twenty-conversation-internet-pioneers

[24] Bogost, Ian. “In'In what way is it accurate to claim that ‘blogging turns 20 this year?’”?" https://twitter.com/ibogost/status/428597798057959424 and “(**@ibogost**)" (**@anildash**) earlier I suggested that blogging names the specific kind of accessible CMS systems that made the form widely adoptable.” https://twitter.com/ibogost/status/428709391236214784

[25] Dash, Anil. “@anildash" (@ibogost) except I remember building the CMSes to match the format I was doing by hand in Notepad.” https://twitter.com/anildash/status/428709625689817088

[26] http://www.computerhistory.org/revolution/the-web/20/388/2129

[27] https://twitter.com/peterme/status/428716148553568257
https://twitter.com/ibogost/status/428730540326256640



Peter Merholz @peterme · Jan 29

@cwodtke Not a journal, but quite plainly a blog: computerhistory.org/revolution/the… /@ibogost @anildash

Ian Bogost @ibogost

@peterme @cwodtke @anildash no.

10:25 PM - 29 Jan 2014

Figure (4) Ian Bogost laying down the definitional law on blogs. I leave the reader to conclude who is right.

The Mosaic Communication Corporation's "*What's New*" page had the technical features we often associate with blogs, links, short entries, and posts listed in reverse chronological order. However, the site was not built on top of any of the infrastructure of modern blogs (Wordpress, blogger), it also existed long before the term Blog or Weblog came into being (but that doesn't necessarily mean it isn't a blog). The MMC *What's New* page doesn't have an author. The person behind the blog is a nameless faceless employee of a corporation. This is radically different from other early blogs like Justin Hall's, which were focused around an individual's lived experience.

Communications scholar Ignacio Siles has studied the stabilization and normalization of blogs from functionalist "filters" to "formats" with styles and structure. The early perceptions of blogs, as articulated by early and influential blogger, Rebecca Blood, stated the purpose of blogs is to filter and assemble links from across the web. Yet today blogs have come to be much more than so-called "linked lists." Siles explores the question, "How did blogs develop from filters into a flexible Web 'format'?" (Siles 2011).

Drawing on Pinch and Bijker's notion of *closure*, Siles, shows how blogs as a format have stabilized through "the mutual shaping of artifact and content" (Siles 2011, 739). The emergence of Blogger.com as a platform for blogging is given important significance in Siles's analysis. Early web publishers and online diarists wrote their posts directly in HTML, a high learning curve for online publishing. Software like Dave Winer's *Frontier* automated the publishing process, but also stabilized structure and format conventions like top posting, by putting new content at the top of the page where readers were more likely to see it.

With the stabilization of technical structure, genre expectations changed as well. The previous shape of blog's forms, as a curated collection of links from across the web, expanded to focus on forms writing like essays and journalist notebooks. Further stabilization occurred with the emergence of the Blogger software platform, which made publishing easier and semi-automated. Blogger's interface emphasizes text, a "blank box" where bloggers incorporated content above and beyond an annotated list of links.

> The expansion of the content of blogs was tied to their material stabilization, particularly by informing the emergence of blogging software. The creation of software crystallized the efforts of communities of Web users (such as online diarists, personal publishers, early and new weblog users) to communicate with others in the public sphere, expressed through the re-articulation of the blog's artifact and content (Siles 2011, 747).

Stabilization occurs through a mutually constitutive relation between content producers and software developers, between sociality and materiality. This is not an exclusively a technical or cultural explanation, it is both. Blogger's identities infused code and that code in turn shaped their conceptions of self (Siles 2012). Bowker and Star, in *Sorting Things Out* discuss the process of *normalization* as one in which expectations, assumptions, and decisions are sunk into the design of infrastructure (Bowker and Star 2000). As Siles demonstrated, blogging practice, and more importantly bloggers as a community of practice, stabilized blogging as a format, as a genre, through a sociotechnical shaping of publishing *infrastructure* like the Blogger software and hosting platform. This process was neither technologically deterministic, nor was it purely cultural, but a negotiation between the two.

Blogs, as a complex sociotechnical apparatus, afford distinct forms of discursive practice. These forms are, in turn, unique to the specific communities of practice that inhabit and shape them. Academic communities have also turned to blogs as a format for scholarly communication. This raises questions about the complex relationship between scholars and their blogs.

Scholarly Blogs

In the book *Uses of Blogs*, Alexander Halavais argues academic blogs constitute a "third place" supporting, but not replacing, traditional modes of scholarly communication (Halavais 2006). Blogs are a kind of electronic "coffee house" or conference that enable near strangers in a community of interest to engage in debate and discussion. The structure of blogs satisfies, in Halavais's mind, a kind of discursive ideal, a "virtual place for continual discussion with little cost or commitment from participants" (124).

Halavais's idealized notions aside, he notes that blogs are by their very definition visible and make available the latent "social circles" implicit in Crane's development of the *Invisible College*. Cronin et al. have called the web a "nutrient-rich space for scholars" (Cronin et al. 1998), and found it to be simultaneously a space for publishing and conversation. The web, they argue, invokes new forms of communication that are not fully understood and while they put forth a typology of web-based "genres of invocation," blogs were not included.

In her dissertation work, Carolyn Hank surveyed communities of scholars across the humanities, social, and natural sciences in an effort to better understand the characteristics, preferences, and perceptions of scholarly blogs (Hank 2011; Hank 2013). Her work focused specifically upon the archival implications of blogs as an important part of the scholarly record. Blogs leave a trace, but Hank asks the crucial question, "for how long?"

In her survey of 153 academic bloggers and analysis of 93 blogs, Hank found that 66 percent of her respondents considered their blogs to be part of the scholarly record and subject to critical review, although they did not significantly respond that their blogs improved their chances of promotion and tenure. In terms of preservation, 80 percent felt preservation and access to their blogs was important, but only 46 percent save their blogs. What this means is that while scholars are deriving personal benefit from their blogging practice, the social and technical infrastructures of promotion and preservation are lagging far behind the practice.

The scale and pace by which digital humanities scholars use blogs is a massive preservation problem, if we want to save these scholarly records. Even more generally, scholarly communication research has always lacked large-scale investigation in part because of a lack of citation data (Linmans 2009; Sula and Miller 2014). Yet, it is important to not take a technologically deterministic view that new technology shapes communicative possibilities and practice (Cronin 2003; Kling and Callahan 2003). The adoption and use of a particular technology, like blogs, intertwines social dynamics, like the digital humanities delight in new technologies, with the Open Web's technical affordances of open standards, open access, open publishing.

Infrastructure Studies

The study of infrastructure might seem to some as a rather idiosyncratic investigation of *boring things* (Lampland and Star 2009). But infrastructure can also reveal embedded assumptions and processes subtly shaping action and practice. The studies reveal infrastructure not as thick technical descriptions, but complex webs of social, technical, political, economic, and cultural elements intertwined and tucked away out of sight.

Infrastructure studies focuses on more than simply classic physical infrastructures like electrical grids, roads, and shipping containers (Hughes 1983; Goddard 1996; Levinson 2010). Information technology, from high performance computing grids (Avery and Avery 2007; Ribes and Bowker 2009; Ribes et al. 2013), to standards and classification systems (Bowker and Star 2000; Egyedi 2001; Zimmerman 2008) have been analyzed *as infrastructure*. Studying phenomena as infrastructure offers interesting practical and theoretical challenges. For this project, infrastructure offers a sociotechnical sentiment and emphasis that partially privileges the technical features of blogs, but does so without being technologically deterministic. Technology can never be removed from social contexts, but conversely, social context cannot be removed from technical contexts.

Infrastructure as Sociotechnical Order

Infrastructure studies' sentiment reflects a turn to the practical and material and the social constructivist turn towards actor's categories, distinctions, and understandings. We, as observers and analysts of infrastructure, must resist the temptation to import extant categories and theories to forge structure that is not actually there. It is a classic ethnomethodological move, look not for the existence of structuralist assumptions that specify a priori social orders, but rather seek the situated, interactional, and social work that *constitutes* social order's *recreation*, *regeneration*, and *maintenance*.

> Not because of our ability to 'use technology' which has so often been used to distinguish humans from animals, but because of our capacity to delegate the work of sustaining social order to objects, such as heavy keychains, or speed-bumps. These objects act with greater obduracy than humans, helping us to produce and reproduce order in the world: e.g., keys that return with the concierge or residential zones with slow driving (Ribes et al. 2013, 10).

Star argued most studies of sociotechnical systems follow "the traditional purview of field studies: talk, community, identity, and group processes, as now mediated by information technology" (Star 1999b, 378). This nod towards the technical means programmers, code, and database schemas can become the inadvertent guardians of social and moral orders. To exclusively focus analytical attention on either the social or the technical facets of these systems is to privilege only part of the story.

Attending to both the social and technical details of infrastructure presents non-trivial methodological challenges for the scholars interested in such phenomena. Given infrastructures are widely varied and distributed, Star asks, “How does one study action at a distance? How does one ever observe the interaction of keyboard, embodied groups and language? What are the ethics of studying people whose identity you may never know” (Star 1999b, 379)? These questions present not only practical challenges, but theoretic ones as well. Digging into the human and technical underpinnings of systems not only expands our horizons of understanding, it creates important connections to questions of social justice, power, and the ethnics and values *in* design. Such a framing reflects the moral imperative infused within studies of infrastructure. That is, infrastructure studies should seek to reveal the invisible and underserved by “surfacing invisible work” (Star 1999b). There is a potentially confounding tension that emerges around studies of invisible work from an empirically grounded, ethnographic perspective. How are we supposed to surface such invisible work if it is, by definition, invisible?

Invisible work really means invisible to *whom*. It is important to remember, one person’s infrastructure is another person’s job. No infrastructure is truly invisible; there are always traces of action and practice, of decisions and technical encodings. Remember, one of the significant properties of infrastructure revealed by Star & Ruhleder is that it becomes *visible upon breakdown* (Star and Ruhleder 1996). When the electrical grid fails or the wireless Internet stops working, a seemly invisible infrastructure becomes radically visible. Such moments render hidden infrastructures *practically observable* .[28]

[28] Careful readers might notice a similarity to infrastructural breakdown and Garfinkel’s breaching experiments as a method for revealing normally hidden social orders. The sociological investigation of norms can be difficult because norms are, by definition, tacit and implicit. Breaching experiments, or the purposeful violation of norms are powerful means for surfacing what might normally be taken for granted. I should note however that Infrastructure Studies does not usually advocate the purposeful violation of social and technical orders embedded within infrastructure. That is, we do not try and break the electrical grid so as to see what happens.

Holding true to the axiom, “infrastructure appears only as a relational property, not as a thing stripped of use” (Star and Ruhleder 1996, 113), means digital humanities blogs are deeply intertwined with the Open Web, the Internet, and the networks of fiber-optic cables and satellites constituting the backbone of information communications technology. This relational approach must, as Latour puts it, *follow the actors*. In infrastructural terms, this means following the traces of a litany of actors and objects in the context of their use. This study focuses on how digital humanities scholars use blogs and brackets out other adjacent networks, platforms, and infrastructures that are imbricated with blogs. Blogs invoke many layers of infrastructure, following *all* the traces leads to the origins of the Internet (Abbate 2000), which may be a trace too far. Because of the rationality, studying infrastructure and identifying the boundaries between them, is challenging.

Studying Infrastructures

Infrastructure studies do not ignore the ways in which instrumentation, code, weather, and physical spaces constrain and compel certain patterns of social practice. Yet, at the same time, infrastructure studies must be careful not to over-determine the influence of such technological and material conditions. We must recognize how the affordances of information communications technology make possible discursive interactions at a rapid pace, but WordPress as a technical artifact does not compel scholars to blog. Blogs are intertwined into a suite of normative obligations, avenues of access, and attentional economies situated between the habits and practices of academic life.

The study of blogs as infrastructure demands a kind of methodological flexibility. As Bowker et al. point out, the traditional ethnographic approach drawn from ethno/anthropological traditions of science studies, while extremely effective at uncovering the subtle nuance of particular social settings, produces awkward accounts of “someone typing on a keyboard” (2010, 113). Furthermore, the size and scale of web-based infrastructure far outpace the capacity of any individual scholar (Star 1999b). Qualitative analysis of four hundred blog with tens of thousands of posts is not possible. As infrastructures scale, the methods by which we study infrastructures must scale with them. The next chapter engages the theoretical and methodological challenges of studying blogs as infrastructure at scale.

Research Questions

A new and interesting scholarly community, the digital humanities, are communicating on new platforms, blogs and the Open Web. Because of the technical affordances of the Open Web as an infrastructure, that communication is visible to study, but at a scale that far outpaces the typical methods deployed to critically study infrastructural phenomena. All of the factors described in this chapter motivate the following questions:

- *What are the methodological dynamics and tensions in generating data from the Open Web?*
- *What themes are digital humanities scholars writing about on their blogs?*
- *How does scholarly blogging expand current understandings of informal scholarly communication?*

To answer these questions requires a new and innovative mode of inquiry that *scales* with the vastness by which the digital humanities community uses blogs. Fortunately, we can mix quantitative methods from the digital humanities and computer science, topic modeling, with qualitative methods from infrastructure studies, trace ethnography. The next chapter will introduce these two methodologies.

Chapter Three

Methodology

The previous chapter wound an intellectual path through three areas of scholarship. First, the chapter threaded the literature on scholarly communication, focusing upon the need for more work studying *informal* communication. The social and technical visibility afforded by blogs has radically transformed the conditions by which scholars informally communicate; the web's *visibility* affords access creating new opportunities for research. Second, the chapter ventured into a specific area where scholars are adapting to new technology, the digital humanities, and how this community has, in part, emerged from blogging as a mode of discourse. To understand these technical dynamics, the third section introduced the field of infrastructure studies, a sub-discipline of science and technology studies, as a framework to ground my inquiry into how blogs and the Open Web are used by scholars. Approaching blogs from an *infrastructural* perspective is useful because it symmetrically calls attention to sociotechnical dynamics, i.e., the interactions between human and non-human actors. However, complicated, large-scale phenomena such as Open Web present methodological difficulties for the kinds of inquiry typical of infrastructure studies, ethnography, participant observation, and qualitative interview. These methods do not scale.

To address these methodological quandaries, an unsupervised machine learning technique, topic modeling, was combined with a novel and experimental approach to understanding the *traces* of online interactions. Mixing these methodological approaches is new and untested, so this project, in addition to developing insights about blogs in the digital humanities, is also a contribution that explores how computational and qualitative methods of inquiry can be combined.

Methods Summary and Objects of Study

The research questions stated above are answered using multiple modes of inquiry: topic modeling and trace ethnography. Topic modeling is a quantitative technique from computer science and trace ethnography is a qualitative method from infrastructure studies. I discuss both methods in more detail below. I have also included an extensive appendix on topic modeling to explicitly foreground the assumptions embedded within its quantitative formalizations.[29]

Topic modeling finds patterns of co-occurring words within documents. These clusters of words can be interpreted to represent various “topics” within collections of documents. These “topics” are interpreted, with some qualitative effort, as high-level themes digital humanities scholars are writing about on their blogs. This means the topic model and the content of the blog posts are analyzed together as qualitative data. This qualitative work of reading and interpretation is informed by an emerging methodology called *trace ethnography* (Geiger & Ribes 2010, 2011). Trace ethnography is an approach to understanding the *traces* of online activity (edit logs, browser history, forum interactions, blog posts, etc.) by both *human* and *non-human* actors. This chapter establishes the *trace ethnography of scaling* as the means for analyzing and interpreting a computational representation, a topic model, of digital humanities blogs.

The *Compendium of Digital Humanities*

Digital Humanities Now (DHNow) is a filtering and discovery service dedicated to “discovering the best of digital humanities scholarship” being publishing across the web.[30] Editors monitor blog feeds and social media, looking for high

[29] Please see appendix A for a “thick description” of topic modeling.

[30] http://digitalhumanitiesnow.org/

quality blog entries from the digital humanities across the web. Editors review articles and make selections “editor’s choice,” every few days. The project has been running for the past five years at the Roy Rosenzweig Center for History and New Media (RRCHNM) at George Mason University.

To assist in finding digital humanities content, DHNow compiled an inventory of blogs, the *Compendium of Digital Humanities*. As of this writing, the Compendium contains the names, URLs, RSS feeds, and twitter handles of over seven hundred bloggers curated based upon the quality and relevance of blog’s content.[31] The blogs are actively monitored by DHNow editors to “highlight work from the Open Web that has gotten the attention of the digital humanities community or is worthy of such attention, based on critical editorial review.”[32]

Scholars at RRCHNM formally compiled the *Compendium of Digital Humanities* in 2009 (Interview). The *Compendium* is not a comprehensive list of every digital humanist who blogs, for there are surely digital humanities bloggers who are (or were) missing from the list. There are also scholars like danah boyd who would not self-identify as a digital humanist, yet who are nevertheless included because their posts are deeply relevant in discourses in and around digital humanities.[33]

When this project began, the *Compendium* listed five hundred domains, most of which, but not all, are blogs. Some domains are blog-like, but individually different in their blog-like quality. Each domain contains many types of pages, some of which are blog posts, but not all. They also contain general pages relevant to being a scholar like CVs, formal publications, research projects, and course information. This task of isolating the blog content from the other errata is one of the many tensions in transforming this list of domains into data.

[31] Interview with Joan Fragaszy Troyano

[32] http://digitalhumanitiesnow.org/about/

[33] danah boyd is also significant because she was both a very early blogger and one of the first academic bloggers.

The *Compendium* was the sample of digital humanities blogs selected for this project. It was created by members of the community and is comprehensive in its coverage of English language digital humanities blogs. The bias or skew of this sample is due to the selection and curation processes of digital humanities community members themselves. The *Compendium* exists as a performance of the community's own sense of self and membership. In this sense it is an effective sample of digital humanities blogs.[34]

Topic Modeling

Topic Modeling is a general term for a type of machine learning technique used to find patterns in high dimensional data. Over the past decade topic models have become very popular for analyzing and understanding large volumes of text.[35] These models are *unsupervised* meaning they do not need a pre-categorized set of training data to produce novel and useful insight into large amounts of data. Given the increasing amount of textual data, either from digitization projects or digitally native text from the web, it is no wonder that such a technique attained such popularity.

The literature on topic modeling can be divided into two categories, articles *about* topic modeling and articles *using* topic modeling. The first category is typically published in computer science journals or conferences. The main contribution in these publications is the *methods* and *models* developed by the authors. However, the second category is more diffuse, comprised of books, journal articles, conference papers, and blogs. These publications focus more upon the insights garnered from using topic modeling upon large volumes of texts.

[34] It is also important to note the Compendium is a dynamic construct. The list had five hundred domains when this project started, today it lists seven hundred. Futhermore, many of the domains listed were not included for practical or technical reasons. That rational is described in chapter 4.

[35] Topic models have also been applied to computer vision problems, but that is still a research frontier.

About Topic Models

Topic Modeling's origins go back to the early 90s when *latent semantic analysis,* a technique for clustering like-words from natural language processing and information retrieval research (Deerwester et al. 1990), was introduced. Contemporary topic models stem from probabilistic latent semantic analysis (PLSA) (Hofmann 1999). Topic modeling gained popularity with the introduction of Latent Dirichlet Allocation (LDA) in 2003. LDA is a generative model that considers documents to be composed of a mixture of "topics" (Blei, Ng, and Jordan 2003).

Latent Dirichlet Allocation expresses a process by which documents are composed by repeatedly selecting a word from a set of probability distributions over words. Creating a topic model means *estimating* the specific probability distributions that generated a corpus of interest. The estimated distributions, a set of topics and a set of document-topic mixtures, are useful for information retrieval or exploratory analysis. For a more in-depth discussion of the LDA model, see appendix A.

The model introduced by Blei et al. in 2003, sometimes referred to as "vanilla LDA," has been extended in multiple ways to accommodate particular structural features of text corpora. *Correlated topic models* infer relationships across topics, creating a network of topics across a corpus. This model was effective at identifying relationships across articles publishing in *Science* to support browsing and searching a collection of scientific articles (Blei and Lafferty 2007). In another study of *Science,* Blei and Lafferty developed *dynamic topic models,* to analyze the evolution of topics over long periods of time. This model analyzed journal articles from the 1880s through 2000; a period in which the words used to describe topics or themes would have changed over time (Blei and Lafferty 2006).

In another important paper on topic modeling, *Finding Scientific Topics,* Griffiths and Steyvers analyzed scientific journal articles and, more significantly, introduced the use of Gibbs Sampling as an alternative method for estimating models (Griffiths and Steyvers 2004). They discovered a set of topics in the abstracts of the Proceedings of the National Academy of Sciences and analyzed them in relation to major (biological, physical, and social science) and minor (math, economics, ecology, etc.) categories assigned the papers. They tested the model to see if the statistical topic model could identify these pre-existing categories based upon word choices in the abstracts. Using a document's extant categorical data and correlating it, either statistically or interpretively, is a common technique to externally validate a topic model (Templeton et al. 2011; Heuser and Le-Khac 2012; Binder and Jennings 2014).

Most of the articles about topic modeling in the machine learning, data mining, and knowledge discovery literature focus on the model, rather than the insights produced by the model, as the primary contribution to the field. We must look to other fields, like the social sciences and humanities, to see how scholars are using topic models to produce insight.

Using Topic Models

Outside of the digital humanities, topic modeling has been used to study scientific publications (Griffiths and Steyvers 2004; Blei and Lafferty 2007; Hall, Jurafsky, and Manning 2008), in authorship detection (Seroussi, Zukerman, and Bohnert 2011; Pearl and Steyvers 2012), differentiating language use (McFarland et al. 2013), to analyze political documents and newspapers (Grimmer 2010; Yang, Torget, and Mihalcea 2011; Bonilla and Grimmer 2013; DiMaggio, Nag, and Blei 2013; Mohr et al. 2013), in bibliometrics (Dietz, Bickel, and Scheffer 2007; Gerrish and Blei 2010), and most importantly to study blogs.[36]

The earliest work using topic models to study blogs was a spatiotemporal model developed by Mei, Liu, Su, and Zhain (2006) to detect spatiotemporal patterns of themes across blogs. In a follow up article Mei, Ling, Wondra, Su and Zhai, used

[36] For an even more extensive bibliography visit David Mimno"s "Topic Modeling Bibliography" http://mimno.infosci.cornell.edu/topics.html

topic sentiment mixtures to model opinions as topical facets of blog posts (2006). Additional early work using topic models to study blogs was a derivative model, Link-PLSA-LDA, developed by Nallapeati and Cohen (2008) to measure influence networks within and across blogs. In another early article Yano, Cohen, and Smith, used topic models to predict comments on political blog posts (2009). Both articles extend LDA to leverage unique features of blogs (linking and comments) to improve measurable and computable notions of performance. Neither article provides much insight into the nature of blogs or bloggers; their primary contributions were computational models.

Other articles studying blogs have combined topic models with entity extraction and sentiment analysis (Singh et al. 2013; Waila, Singh, and Singh 2013), with a focus on demonstrating how these various techniques interplay. In these studies, the subjects (blogs about discrimination India and Arab spring) were selected to demonstrate the effectiveness of an analytical framework; again, they do not provide significant insight in the nature of blogs as a discursive space.

Pagano and Maalej (2011) conducted a study of the software developer community, analyzing blogs and other communications infrastructure in an effort to better understand the discursive practices of programmers. They analyzed blog post content and metadata to understand how often developers posted blogs, what elements they included or referenced in their blogs, what topics were used, and if any patterns emerged in the relationship with development work. Their article is interesting because it used topic modeling to understand bloggers (a means to an end) rather than developing a better model.

The study examined four OSS blogging communities connected through a *blog aggregator*. Aggregators are infrastructural focal points for blogging communities as they automatically collect links to member's blog posts in a central website. They track members via RSS feeds, merge these feeds onto a single page/feed, and store a history of all posts.[37] Unfortunately, the specific details about how they extracted blog content are described in only one sentence, "We used regular expressions to extract [blog] elements (p.126)."[38]

Pagano and Maalej trained fifty -topic models on each of four blog corpora. They selected fifty because it "lead [*sic*] to the most meaningful results" (p.127). This isn't the most enlightening of selection criteria, but it emphasizes how the model was being used in the analysis. Their goal wasn't to find the statistically perfect model that could have generated the corpus, but to leverage the model to generate insights abut the content of a corpus that would otherwise be too large to read. The topics themselves were then interpreted as texts and annotated with descriptive terms. Pagano and Maalej used the topic annotations combined with aggregate document-topic proportions to identify the most popular topics by developer community.

The Pagano and Maalej study leverage topic modeling as a component in an interpretive process seeking to better understand how open source software developers used blogs. Their specific insights demonstrate a technique for analyzing the discursive practices of blogging communities. Their technique of labeling topics and segmenting the corpora along sociotechnical categories (OSS project) informed this project.

Several other studies have leverage topic modeling as part of a mixed methods approach to studying blogs (Mark et al. 2012; Al-Ani et al. 2012). In two papers presented at the 2012 ACM conference on *Computer Supported Cooperative Work,* topic models were used to study bloggers inhabiting conflict zones. Al-Ani

[37] The Compendium of *Digital Humanities*DH could be considered a blog aggregator, only in the fact that it tracks membership. Unlike a true blog aggregator, the Compendium does not (publically at least) store a history of posts. Where Pagano & Maalej had four sites to parse and discover links to posts, I had six hundred.

[38] RSS feeds often contain the full content (depending on the blogger) of a blog post in a normalized, highly structured format. This means the Pagano and Maalej may have had easily access to just the post content without the blog"s remaining structural elements (or comments). I did not have the same access and the diversity of my sample means the text extraction process was more difficult.

et al. (2012) used qualitative coding combined with domain understanding of real-world themes and discourses they examined the top-ranked words to determine the *theme* of each topic:

> We read each grouping of words and hypothesized about possible thematic associations, and then cross-referenced these with a random selection of documents that the topic modeler indicated as containing the topic in question, empirically verifying the validity of the themes.

Individual topics were qualitatively categorized by personal, political, or revolutionary themes and the proportion of these meta-themes were plotted over time. With the onset of the Arab spring uprising the authors observed a significant decrease in personal topics whereas revolutionary topics increased. The quantitative topic model was combined with qualitative reading and content analysis to develop an argument about how blogs formed a *counter-narrative*—a form of networked, communicative "counter-power" (Castells 2011)—in opposition to the narrative of the authoritarian Egyptian government.

While the insights about the Egyptian blogosphere and their counter-narrative are specific to their study, the methods by which Al-ali et al. came to that understanding are generalizable. The design of their study mirrors this project. They scraped the blogs to collect large amounts of data, which were explored using topic models. Topics became loci of an interpretive lens and combined with visualization of the metadata and qualitative readings of the blog content the authors located what they observed within a theoretical and conceptual framework. Al-ali argued that blogs distributed architecture, what Castells calls *mass self-communication* can be a tool of semi-organized resistance to the narratives of institutionalized power like government or corporate controlled mass media. The paper's contribution is a theoretical and conceptual argument built upon data, quantitative analysis, and interpretation of computational models.

Nardi (2006) characterized blogs as diaries; Mark et al. (2012) extended this conceptual framing to consider wartime blogs as a kind of conflict diary. They wondered is the blogosphere could be used to understand how a society responds to war over time. The challenge, they rightfully identify, is how to manage the sheer volume of posts a society produces over a long period of time.

> This leads to our concomitant research objective: to investigate how data analysis techniques yield insight into the collective expression of large numbers of people about a significant event over time. (Mark et al. 2012, p. 38)

At the onset of the study, the authors did not know the extent to which Iraqi bloggers would discuss war in such a public forum. Consistent with the study of Egyptian bloggers, the Iraqi study found a personal blogging decreased at the height of the crisis, to increase again as the conflict reduced in intensity. Theses findings were then situated alongside sociological theories of disaster and crisis coping as well as Giddens's classical notions of social structure and normalcy.

Topic Models in Digital Humanities

Newman and Block's "Probabilistic topic decomposition of an eighteenth-century American newspaper" is the very first article in the humanities to use topic modeling like approaches to study historical documents. Their study tested three methods, Latent Semantic Analysis (Deerwester et al. 1990), k-means clustering (Duda and Hart 1973), and probabilistic Latent Semantic Indexing (Hofmann 1999) on the *Pennsylvania Gazette* from 1728–1800. Of the three techniques, pLSA (a close relative of LDA) was the most effective because of its support for mixtures of topics in documents. The article served as an introduction and early demonstration to how computational clustering techniques could be used to identify historical trends.

An early contribution to use topic modeling was never formally published in the print-centric sense of the word. Robert K. Nelson's project *Mining the Dispatch* used topic modeling to analyze a digitized nineteenth-century newspaper, the *Richmond Daily Dispatch*.[39] According to Nelson "the real potential of topic modeling, however, isn't at the level of the individual document. Topic modeling, instead, allows us to step back from individual documents and look at larger

[39] http://dsl.richmond.edu/dispatch/

patterns among all the documents." Given the size of the Dispatch, "over 112,000 pieces amounting to nearly 24 million words," traditional historiographic methods of skimming and sampling might miss crucial insights. Topic modeling, Nelson argues, "provides historians an additional method that allows us to examine and detect patterns within not a sampling but in the entirety of an archive."[40]

Using the MALLET toolkit,[41] Nelson extracted a set of topics from the collection of articles over the five-year run of the paper. Then, by counting the number of articles associated (by percentage) to each topic over time, he was able to identify trends and patterns in the paper that mapped to broader historical and contextual understandings of the period. For example, when the Union army occupied areas near Richmond, the number of articles related to the topic "fugitive slave ads" increased. According to Nelson, this can be explained by how the Union army provided opportunities for slaves to escape to freedom. However, not all patterns could be easily explained by existing understandings of the time. He observed spikes in "fugitive slave ads" with no readily apparent historical explanation. "Topic modeling and other distant reading methods are most valuable not when they allow us to see patterns that we can easily explain but when they reveal patterns that we can't, patterns that surprise us and that prompt interesting and useful research questions." Distant reading techniques are most powerful when they motivate subsequent close readings of interesting and inexplicable phenomena not otherwise discoverable by skimming and sampling.

Another early exploration of topic modeling was Cameron Blevins's topic model of Martha Ballard's diary. The study, published as a blog post, described Bleven's experiences using MALLET:

> When I first ran the topic modeler, I was floored. A human being would intuitively lump words like attended, reverend, and worship together based on their meanings. But MALLET is completely unconcerned with the meaning of a word (which is fortunate, given the difficulty of teaching a computer that, in this text,

[40] http://dsl.richmond.edu/dispatch/pages/intro

[41] http://MALLET.cs.umass.edu/

> discoarst actually means discoursed). Instead, the program is only concerned with how the words are used in the text, and specifically what words tend to be used similarly (Blevins 2010).

Blevins's reaction is common, topic modeling produces sometimes spooky clusters of meaningful words, but once you understand the statistical underpinnings of the model the magic begins to fade (but the usefulness does not). The study of Ballard's diary was never formally published, nor were there any specific "findings." The model, annotated with metadata, only confirmed things he already knew. Entries that occurred in the summer months had a higher proportion of the "gardening" topic, topics about winter had a higher proportion in the winter months, and days when she attended childbirths had a higher proportion of the "midwifery" topic.

In a recent article in *Literary and Linguistic Computing* Binder and Jennings address what Alan Liu called the "meaning problem" in digital humanities (Liu 2013). The challenge, they reflect, is how to bridge the gap between the output of computational models and the "humanistic discourse in which historical questions are posted" (Binder and Jennings 2014, 1). They propose a visualization method, the Networked Corpus, which situates the results of topic modeling alongside an original index and the text itself. Such a mode of visualization literally links computational and historic indexes with the text itself. It is an *enriched* close reading of the texts and their metadata.

In his book *Macroanalysis: Digital Methods and Literary History* Matthew Jockers argues for topic modeling as a method for "discovering" the *themes* within a collection of 3,346 books (segmented into 1,000-word chunks). The analysis uses metadata like publication date, author's nationality and gender to annotate a five-hundred-topic model. Jocker's analysis of topic models as themes is part of *Macroanalysis's* larger project of initializing a methodological discussion within studies of British and American literature.

The *Journal of Digital Humanities* dedicated an entire issue to topic modeling as a method of inquiry.[42] The issue consolidated some of the extensive discussions occurring in the digital humanities blogosphere about topic modeling, as well as solicited commentary from David Blei about the relationship between topic modeling and the Digital Humanities. Blei (2013) argued that probabilistic modeling provides a

> statistical lens that encodes her specific knowledge, theories, and assumptions about texts. … Traditionally, statistics and machine learning gives a "cookbook" of methods, and users of these tools are required to match their specific problems to general solutions. In probabilistic modeling, we provide a language for expressing assumptions about data and generic methods for computing with those assumptions.

Blei's dream is that humanities scholars could use the language of probabilistic modeling to create their own generative models that encode assumptions, ideas, and theories about texts. Humanists with these skills could create more complex and contextualized models than those created by computer scientists with litter domain expertise.

The special issue featured two important critical reflections on topic modeling, Lisa M. Rhody's *Topic Modeling and Figurative Language* and Benjamin M. Schmidt's *Words Alone: Dismantling Topic models in the Humanities*. In *Figurative Language* Rhody discussed how the "beautiful" failures of topic modeling poetry actually opened up a generative interpretive space.

> Searching for thematic coherence in topics formed from poetic corpora would prove disappointing since topic keyword distributions in a thematic light appear at first glance to be riddled with "intrusions." However, by understanding topics as forms of discourse that must be accompanied by close readings of poems in each topic, researchers can make use of a powerful tool with which to explore latent patterns in poetic texts (Rhody 2013).

Rhody emphasizes scholars should return to the texts (in her case Ekphrastic poetry) and not rely exclusively on the meanings derived from the model (i.e., topic word distributions) because, especially in the case of poetry, the demolition of document structure can lead one's analysis astray (Rhody 2013).

[42] Journal of Digital Humanities Volume 2, Number 1. http://journalofdigitalhumanities.org/2-1/

Schmidt's *Words Alone* argues topic models create a potentially dangerous gap between the humanist and individual words of a text. This gap can create a false assurance of coherence within the topics. Schmidt demonstrates the interpretive pitfalls of this gap by modeling geographic data, nineteenth-century ship's logs, and plotting the "topic" clusters on a world map. Geographic meaning, unlike linguistic meaning, can be quickly and easily scrutinized by plotting on a map.

> With geodata, it is much easier to see how meaning can be constructed out of low-frequency sets of points (points that might available in another vocabulary as well); but in language as well, the most frequent words are not necessarily those that create the meaning. A textual scholar relying on top ten lists to determine what a topic represents might be as misled as a geography scholar mapping routes based on the red above, rather than the black (Schmidt 2013).

Using the geographic ship-route data, Schmidt argues that attending only to the high proportion words in a topic can mislead a textual scholar trying to interpret the meaning of a particular topic distribution. There is potentially important meaning at the tail end of the distribution; i.e., the low ranking words (Schmidt 2013).

Scholarly literature, especially in the humanities, has been a fruitful area of study for topic modeling. Five studies have used topic modeling to explore scholarly discourse in formally published journals. Mimno analyzed a century of classics journals (Mimno 2012), Laudun and Goodwin looked at folklore studies (Laudun and Goodwin 2013), Riddell has looked at German studies (Riddell 2014), and Goldstone & Underwood have studied both the PMLA specifically (Goldstone 2013) and literary studies more broadly (Goldstone and Underwood 2014).[43]

Each of these articles are responding to recent changes, and challenges, in what is analytically possible with digitized historical texts and computational models. The digitization of journals in their respective disciplines (and easier access in terms of search and APIs), these authors see the potential for better insight into disciplinary history. However, they all point to the problem of abundance; there

[43] Goldstone has also posted a digital appendix to their forthcoming literary studies article: http://www.rci.rutgers.edu/~ag978/quiet/#/about

is simply too much for any one scholar to read. Enter the familiar turn towards computation and distant reading.

Laudun and Goodwin's (2013) analysis of folklore studies, as well as Riddell's analysis of German studies are both, what I might call, classical examinations of a large, diachronic corpora using topic models. Each study trained a topic model on journal articles and then used document level metadata, publishing date, to generate a visual representation of topic proportion over time. The analysis was driven by the question of whether they could see folklore studies'' *turn to performance* and the rise of post-structural discourse in the model. They identified specific topics relating to "performance" and plotted the topic's density.

> While we predicted that topic modeling would reveal an increase at the time corresponding to the performative shift in folk-loristics, the five-year means of the cultural performance discourse topic's occurrence from 1888–2012 … shows a clear rise in the 1970s (463).

Similar to the previous studies in the social sciences, Laudun and Goodwin interpret the model's top words to assign more meaningful labels to topics. However, unlike the social scientists their labels do not reflect any kind of inter-code reliability/alpha rating. There are multiple ways to label a topic, from simply taking the three or four most probable words, to performing a qualitative coding exercise, to a more hermeneutic approach. Disciplinary conventions and expectations should hopefully govern this practice, but given topic modeling is a relatively recent methodological practice, these conventions and expectations are still forming.

Mimno's *Computational Historiography* (2012) explored the use of computational techniques to study classics journals. Using an assortment of methods, Mimno first represented documents as word distributions and compared them with Jensen-Shannon divergence; a well recognized technique for computing similarity of probability distributions. This technique produced an NxN similarity matrix that he scaled down to two dimensions for visualization.

Mimno was able to use this plot to show the divergence of two well known and distinctive disciplinary traditions, philology, and archeology. It is notable the transformations to the data to produce this model and representation reflected this trend. Using principal component analysis or multidimensional scaling to render visualizations in this way does not always yield meaningful results.[44]

Computational Historiography uses a topic model to examine variation within journal vocabularies. In a case study of Roman Studies Journal, Mimno segmented topic distributions ranking most probable words by decade. This is a novel diachronic analysis of a topic, vs. the typical diachronic analysis of documents, to see how particular sets of words move in relative proportion *within the topic*.

In *Quiet Transformation* Goldstone and Underwood (2014) examined 21,000 articles from the past 120 years of literary history. The main argument of the article is a methodological demonstration to show new ways of analyzing and understanding both literary history and the textual subjects of the field. Their model was validated, similar to Laudun and Goodwin, by tracking disciplinary discourses (i.e., New Criticism, New Historicism, Marxism, post-structuralism) via specific topics. All of these analyses of scholarly disciplines (through their scholarly communication) are predicated upon a reflexive, collective, a priori disciplinary self-understanding. As Laudun and Goodwin (2013) put it,

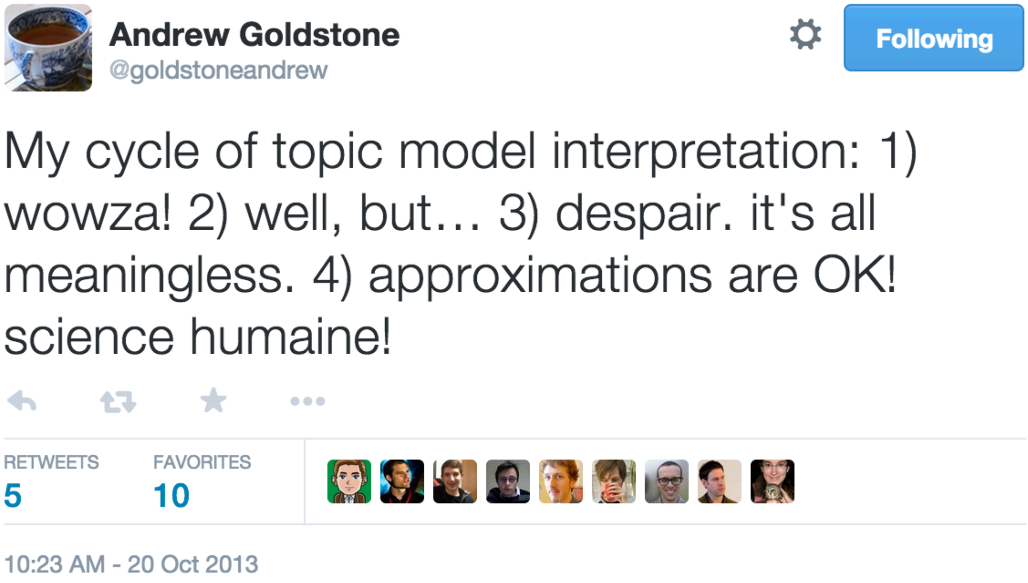
> We would like to suggest that folklorists, like most practitioners in a field, understand the history of their discipline through a combination of their own reading and the consensus inherited from their graduate training and professional interactions. Disciplinary history, an effectively oral form of communication, codifies quickly. Highly contingent and random processes become widely understood as historically inevitable. (457)

This highlights a few distinctions between this study of digital humanities scholarly communication and these field-specific studies. First, these studies are being conducted by practitioners who understand their discipline's history "inherited from their graduate training and professional interactions." Second, these histories are long (more than a hundred years in some cases) and rich with well-known meta-narratives (the rise of post-structuralism, the performative

[44] Ted Underwood has discussed the use of compressing topic model distributions into two dimensions with mixed results. http://tedunderwood.com/2012/11/11/visualizing-topic-models/

shift, etc.). While the digital humanities has a surprisingly long and rich tradition, it is a sparse history in comparison. Finally, the informal discourse on blogs has no preexisting self-understanding. Digital humanists are still in the processes of finding their "performative shift" or new criticism. The most significant difference and arises from the formal vs. informal mode of publishing. Disciplinary histories, their conflicts and paradigm shifts, are materially rooted in their formal publications and socialized through graduate training.

Interpretation at Scale


Andrew Goldstone
@goldstoneandrew

My cycle of topic model interpretation: 1) wowza! 2) well, but… 3) despair. it's all meaningless. 4) approximations are OK! science humaine!

RETWEETS 5
FAVORITES 10

10:23 AM - 20 Oct 2013

Figure (5). Andrew Goldstone's "hermeneutic cycle" of topic model interpretation.

The question of how to interpret a model produced via distant reading in a way that is congruent with humanistic interpretation is an open and ongoing area of research. In practice, distant reading produces a pile of numbers or synthetic representations that require interpretation in their own right. Topic modeling produces a probabilistic distribution over words, which in practice is treated as a finite number of "top keywords" with the most probability for that topic. These keywords are generally clusters of words that co-occur within documents in the corpus.

The larger epistemological questions about what it means to *know* using modes of distant reading are well outside of the scope of this project. However, such problems cannot be avoided. *How do we make sense and meaning from a topic model?* This question is posed from a point of origin in *interpretive social science*. The goal is not to arrive at generalizable claims about the digital humanities or scholarly blogging; this project seeks to *explore* a discursive space to understand *what* digital humanities scholars are writing on their blogs.

As discussed above, using topic models as a mode of inquiry in the humanities and interpretive social sciences is still experimental. Interpreting topic models, and quantitative or computational models more generally, is an active research topic (rimshot) in the digital humanities (Mohr and Bogdanov 2013; Moretti 2013; Underwood 2014).

With respect to methodology, Goldstone and Underwood's *Quiet Transformations* explicitly tried to formalize their interpretive practice. Their analysis of a topic model draws on literary hermeneutics *and* content analysis (Krippendorff 2012) from the social sciences. The subjects of their content analysis are the original texts, the model, and their intersections. The number of topics selected and the model produced constitute an *interpretive frame* supported by the pluralistic notion of "multiply determined discourses."

> Discourses are always multiply determined, and lend themselves to multiple valid interpretations. This multiplicity isn't only produced at the margin—because we could change a corpus, for instance, by including or excluding authors. It's equally true at the center of the interpretive act, since the very same corpus can be divided in more than one persuasive way. We're always constituting some figure by excluding some ground, and there is usually more than one interesting pattern that could be produced. (Goldstone and Underwood 2014, 14)

In this argument, the computational model, graphs and numbers are not being deployed to support a predetermined thesis rather they are enrolled in an interconnected human/nonhuman hermeneutic. The promise is a generative framework where a new discussion around interpretation opens up within the discipline(s). These are the first of what Goldstone and Underwood hope are additional investigations of the same corpora, but through different interpretive frames.

A less explicit, but no less crucial part of their interpretive frame are the technical details of their text preparation and modeling. Their online appendix includes the number of documents, the number of topics, OCR issues, and links to the code used to generate the model. The appendix, the code, and the online addendum a part of methodological transparency necessary when conducting an interpretive analysis of computationally inflected research. If other researchers want to foreground other patterns from the data, the process, or the interpretation, they need to know, in greater detail than the typical methodology section, how Goldstone and Underwood executed their research.

This level of transparency leverages is an absolute necessity when considering these new modes of scaling digital and computational methodology. How to make sense of and interpret the model, in conjunction with the data, is part of the fundamentally new dynamic introduced by topic modeling as a method of inquiry. Topic modeling requires a transformed representation of the research subject and the process of training a model can be distilled into a sequence of discrete steps *that can be documented and recorded in code.* Digital methods afford richer *traces* of scholarly method and practice, at a much lower cost of work and effort. Considering the code as a *trace* of research practice creates an opportunity for a reflexive methodological practice.

Trace Ethnography

Trace ethnography provides an analytical frame and theoretical basis for interpreting the blogs and topic model *together*. Trace ethnography provides a way of reading blogs, the topic model, and the data collection and preparation practices as a set of documentary traces. Trace ethnography is novel and different from classical ethnography because it provides a theoretical and methodological means for drawing out thick descriptions from thin documentary traces. Trace ethnography does this by analyzing the documentary traces in the context of the social and technical context. It is an approach deeply informed by and contributing to the study of infrastructures.

Considering traces from an ethnographic perspective builds upon the direction of Goldstone and Underwood (describe above) whereby interpretive social science methods can be used to analyze a phenomena (DH blogs) with a computational representation of that phenomena (a topic model). Trace ethnography builds upon on their effort by drawing upon the ethnographic tradition for interpretation. "Traces" provides a richer frame than "content" because it can include both human and nonhuman documents, records, and data.

Using trace ethnography as a means to understand and interpret the topic model of digital humanities blogs follows from three approaches to ethnography. First, we must accept the premise that ethnography can be *mediated*. While face-to-face interaction is the gold standard for ethnography inquiry, daily live is increasingly lived online and ethnography methods (Hine 2000; Beaulieu 2004; Boellstorff et al. 2012) can accommodate mediated lived experience. Second, mediated environments are constituted through documentary *traces* and through these traces people interact with their communities. Third and finally, within some online communities, like the Open Web, the universes of documentary traces are *visible at scale*. Ethnographic inquiry does not scale, but through an ethnographic examination of *scalar devices* a way of knowing large-scale phenomena can be achieved.

Mediated Ethnography

Beaulieu (2004) finds there are multiple ways in which an ethnographer can engage the "field," constitute an ethnographic object, or reconcile their intersubjectivity within online environments. Mediated environments challenge classical notions of going to and returning from "the field." Beaulieu contrasts the classical ethnography model of ethnography "fetishized" in American cultural anthropology, which privileges *colocation* as the only way to engage the field (Beaulieu 2010).

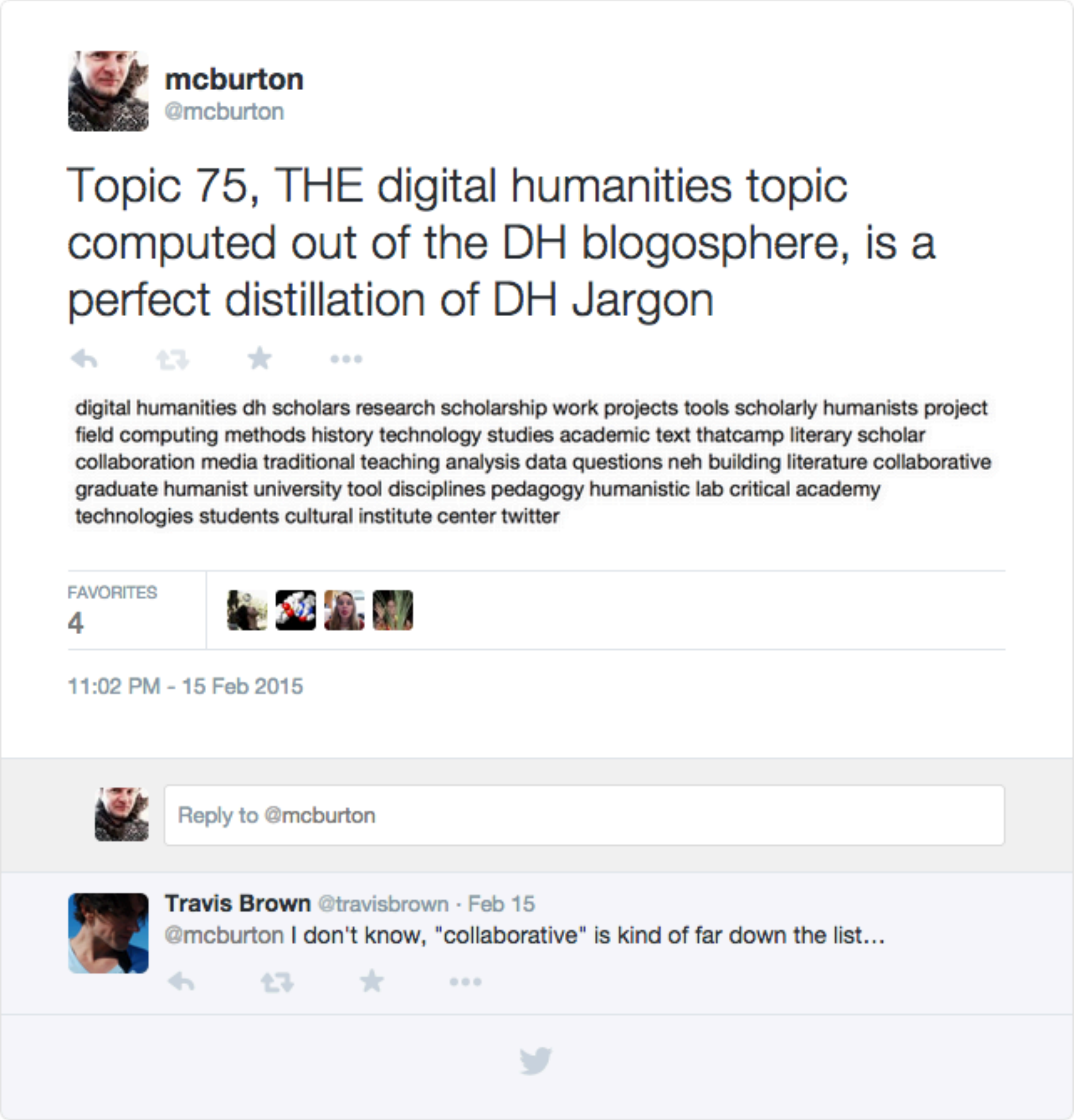
> This [colocation] model ... implies a process of face to face interaction leading to transcription and writing up of notes, then upon return to the home territory, writing of the ethnography (Beaulieu 2004, 154).

In online environments "there is no there there" in the classical sense. However, this is not to say that ethnographic inquiry is impossible. What is required is a shift in emphasis away from *co-location* to one of *co-presence*, whereby mediation and inscriptions can also serve as a substrate for relations and interactions.

Co-presence rather than colocation "opens up the possibility that co-presence might be established through a variety of modes, physical co- location being one among others" (Beaulieu 2010). This theoretical orientation opens up the possibility for mediated ethnography (Beaulieu 2004). Once we accept such possibilities, we must address the temporality of the relationship between the researcher and the research objects. Ethnographic research can be historic, as evidenced by Diane Vaughan's classic study of the Challenger launch. This can mean than documentary evidence, rather than just online interaction, can serve as the "data" for ethnographic inquiry.

According to Beaulieu, this shifts the question of entering the ethnographic field away from "Where do I go" to "How do I establish co-presence?" Establishing co-presence means to enter into a phenomenological space whereby shared meaning and mutual comprehensibility is achieved. There are many ways to achieve mutual comprehension, and I defer to the fields of Ethnomethodology and Conversation Analysis for more complete discussions around that topic. For the purposes of this study, co-presence was attained by *becoming a member of the community.*

For the past several years, I have attended digital humanities workshops, I have presented at conferences, and I have engaged with the community online via blogs and Twitter. My main avenue has been through Twitter, which has been widely recognized as an important "place" where digital humanities happens online.



mcburton
@mcburton

Topic 75, THE digital humanities topic computed out of the DH blogosphere, is a perfect distillation of DH Jargon

digital humanities dh scholars research scholarship work projects tools scholarly humanists project field computing methods history technology studies academic text thatcamp literary scholar collaboration media traditional teaching analysis data questions neh building literature collaborative graduate humanist university tool disciplines pedagogy humanistic lab critical academy technologies students cultural institute center twitter

FAVORITES
4

11:02 PM - 15 Feb 2015

Reply to @mcburton

Travis Brown @travisbrown · Feb 15
@mcburton I don't know, "collaborative" is kind of far down the list...

Figure (6). Establishing co-presence on Twitter through humor

Not only do I many followers and interact regularly with other digital humanists on Twitter. I want to argue that humor, jokes, and snark can be powerful indicators of shared meaning and subsequently co-presence. Understanding humor signals a nuanced and contextual understanding of what something means within a community.

Traces as Ethnographic Data

Trace ethnography is oriented around two fundamental principles:

- "D*ocumentary traces abound in today's technological systems*"
- "D*ocumentary traces are the primary mechanism in which users themselves know their distributed communities and act within them*" (Geiger and Ribes 2011)

Trace ethnography is a term and methodological technique introduced by Geiger & Ribes in their article, "The Work of Sustaining Order in Wikipedia: The Banning of a Vandal" (2010). In the article, Geiger & Ribes investigated the role of software tools in anti-vandalism activities on Wikipedia. Their qualitative research unpacked an intricate network of human and non-human actors who work in concert to battle the onslaught of vandals who continually threaten the quality of Wikipedia.

In online communities, the historical records of member's activities, their traces, are the only discursive formation available. Members of highly distributed online communities assemble by and through the mediation of digital communicative infrastructure. In many cases the documentary traces are, in essence, "all they have to go on." For example, as Geiger & Ribes showed Wikipedia editors used the edit histories of Wikipedia entries in the course of a disciplinary hearing about editorial decisions.

Human and Non-human Traces

Once we begin thinking along the lines of mediated ethnography and "following the actors" it is only a small tweak in our ontological assumptions to consider the possibility that an "actor" might not be a human. Actor assembled documentary traces reconstitute locally meaningful historical activities, such as a disciplinary hearing on Wikipedia. As Geiger & Ribes put it:

> One reason why [preassembled] documentary traces are so useful is that they are produced and circulated in a specific sociotechnical environment, embedded with local meaning. While it is tempting to think of such data as ancillary, kept simply

> because computer systems log data, they are often used by members themselves to render accountable a number of social and organizational practices (Geiger and Ribes 2011, 8).

At a theoretical level, they draw inspiration from Actor Network Theory, specifically the concept of "delegation" as a "heuristic for symmetrically analyzing the way in which both humans and machines contribute to the production of social order" (120). The analytical move here, following from Latour's *principle of generalized symmetry,* is not to think about how non-human actors (bots) inhibit the agency of human actors (editors), but rather to treat both equally and observe how agency flows through their relations.

Practically speaking, this means there is no immediate analytical distinction to be made between traces created by human beings or by computational processes. The activity and edit log is not inherently more meaningful than the content of blog post or chat log. Now, this is not to say that distinctions won't emerge from the analysis, but the goal is to not close a door and not follow a path the path of actors (both human and non-human) through the systems by which they are constituted.

The primary traces of this project are blog posts. An obvious form of analysis would be to read blog posts and follow their comments and tracebacks to reconstruct a space of stories and meaning. If the sample of blogs or the motivating question is small, this would be a perfectly adequate trace analysis. However, what happens in situations when the volume of trace data outpaces the traditional ethnographic modes of inquiry?

Trace Ethnography of Scaling

As the Open Web and digital technologies increasingly make phenomena *observable,* that is, "visible (or audible, as with a conference call) to the ethnographer, whether through unassisted observation or through the use of digital traces, field-notes, audio and video recordings or other documentary materials." More than simply becoming visible and observable, they are observable at scale. Computational and quantitative techniques are typically the sole solution to the problems of scale; however, sometimes the richer narrative and contextual insights from interpretive and qualitative techniques of ethnographic research are desirable. However, ethnography does not scale, yet ethnography is being used to study a variety of large-scale phenomena (Star 1999b; Howard 2002; Velden, Haque, and Lagoze 2010; Williams and Pollock 2011). How can ethnography scale?

Sampling is a common and effective technique for managing a large-scale phenomenon. Sampling in ethnography can be done by scaling up the number of ethnographers or the number of field sites (Williams and Pollock 2011). Multi-sited ethnography is effective up to a point because it is constrained by the transactional overhead of managing a team of ethnographers and by physical travel. Another popular approach utilizes mixed-methods research to bridge the scalability of quantitative methods with the interpretive richness of qualitative methods (Johnson, Onwuegbuzie, and Turner 2007; Clark and Creswell 2008; Creswell 2009; Teddlie 2009; Creswell and Clark 2010). Mixed method ethnographers can assemble a networked representation of the phenomena of interest and leverage a quantitative techniques like network analysis to identify potential sites for richer investigation with ethnography (Howard 2002; Velden and Lagoze 2008; Velden, Haque, and Lagoze 2010; Velden and Lagoze 2013).

Building upon these previous approaches to scaling ethnography, Ribes (2014) introduces the notion of an *ethnography of scaling* that frames the actual practice of scaling *as an ethnographic subject*. Rather than attempting to scale ethnography up, Ribes realigns the focus of inquiry to investigate how actors understand large-scale phenomena themselves. Individual members of large scale enterprises, like scientists on massive, multi-year, multi-site research projects, have particular ways of knowing and understanding the whole, be it through surveys, descriptive statistics, or All-Hands Meetings. Actors participating in large-scale endeavors use what Ribes calls *scalar devices* as techniques or tools for managing scale.

A scalar device is the "assembly of techniques, tools and representational conventions that are used to know and manage scale" by actors engaged in their everyday practice. A PowerPoint presentation about an actors work product or a data-driven dashboard of an enterprise infrastructure can both perform as scalar devices. An ethnography of scale opens up the black box of these indicators and find out how they work. Any scalar device, like descriptive statistics, reduces and eliminates the richness and granularity of a phenomenon. As Ribes points out, every "scalar device only preserves particular relations as it re-describes an object in its distinct representational form" (Ribes 2014, 166). Furthermore, every scalar device is distinct in the particular sets of relations it preserves.

> No single scalar device can service to represent or manage all forms of scale ... [any approach] to scale provides a unique view and suggests different kinds of solutions to manage oncoming problems. But each device also generates exclusions and invisibilities (Ribes 2014, 168).

There is a tendency, according to Ribes, in quantitative statistics, social network analysis, and now big data, for privileged and totalized narratives about a large-scale phenomenon as if network analysis were the only way to "know" a social network. Any scalar device and the interpretation of the phenomena it represents is merely one of a multitude of interpretations. Other scalar devices, like principle component analysis or more comprehensive machine learning techniques, produce *different* representations and different interpretations of a large-scale phenomenon.

Topic Models as a Scalar Device

I use topic modeling *as a scalar device* to "scale down" the incredibly rich discursive space of digital humanities blog into an analytically smaller object more digestible by myself as a solitary scholar.[45] My use of topic models, descriptive statistics, charts, and other computational techniques, could be simply framed as classic social scientific research methods producing objective claims about a phenomenon (digital humanities blogs). Such an interpretation of

[45] This section is purposefully written in the first person because it is important to explicitly recognize myself as the instrument of interpretation.

my work is theoretically and epistemologically unsatisfying. I am not a social scientist attempting to produce a generalizable model of digital humanities blogging. Rather, I am more akin to computational ethnographer trying to understand the process of collecting, preparing, and scaling data.

Topic modeling is only one of many scalar devices I could use to manage scale. Social network analysis is another example of a scalar device I could use to analyze blogs. Each scalar device foregrounds certain features and occludes other features of the thing being scaled. Network analysis is excellent at identifying structures and substructures of the networks within blogs, but it doesn't tell us much about the content. Because I am interested in understanding the different themes digital humanities scholars are writing about on their blogs, topic modeling is a better scalar device because it analyzes content (at the expense of structure).

In considering topic models, specifically Latent Dirichlet Allocation, as a scalar device, I must perform an ethnography of scaling. This means it is important to extensively document *how* to collapse tens of thousands of individual blog posts into a comprehensively smaller representation. All of the scaling work, the collection and re-shaping of data, the use of topic modeling, and the synthetic representation of the blogs is part of the scalar device. The algorithms, processes, and documentary traces are the subject of *trace ethnography of scaling*.

In computer science and machine learning terms, I have a large multi-dimensional space and I want to reduce the dimensionality of the space. Topic modeling is one such technique for dimensionality reduction, but a phenomenon must be quantified into a multidimensional space in the first place. For my project, this means transforming blogs on the Open Web into an n-dimensional document/term matrix. Digital humanities scholars, for all of their computational techniques and know-how, do not write their blog posts in the form of term/frequency lists. This mean I have to conduct the transformation. This process of transformation has important implications for what is preserved or left out of my reduced representation of digital humanities blogs. For example, did I include comments as part of the content of a blog post or not? How did I separate the text content of the blog post from the structural or design elements of the web page?

The process of data cleaning and normalization is filled with a series of small decisions that are carried through, perhaps even amplified, during the research process. These decisions, or *simplifying assumptions*, must be documented as part of the interpretive process. Documenting the computational and reductive work is part of the ethnographic process, like taking field notes. This produces a set of documentary traces *of my own research practice*, which, along side the model and "raw" data, are the substance of my trace ethnography.

Like Ribes, I am not scaling the ethnography, rather, I am engaging is a deeply reflexive self-ethnography of my own scalar devices. This is a very subtle point so I want to make it very clear. Like the actors in Ribes study who deploy scalar devices to help them know a thing at scale, I am deploying topic modeling to re-know digital humanities blogs *at scale*. Such a way and form of knowing is deeply intertwined with my choice of computational practice, topic modeling, and the knowledge produced is synthetic. Ethnographically, I am not re-constituting the lived experience of digital humanities scholars as they blog, rather I am constituting the *lived experience of a algorithm*.

In practice, this means I documented *everything* involved in the creation of the topic model. Using interactive computation in the form of "notebooks" I maintained a set of *traces* of my research practice. These traces become the basis of a self-ethnographic inquiry into the methods by which I "scaled" 106,804 individual blog posts "down" into a representation I could practically interpret. The next chapter is an ethnography of scaling, that is, a thick descriptions of what I call the *simplifying assumptions* involved in transforming blogs from the open web into data.

Chapter Four

Turning Blogs Into Data

> 50 percent to 80 percent of time in data analysis is spent on data cleaning (Dasu and Johnson 2003).
>
> Data preparation is not just a first step, but must be repeated many times over the course of analysis as new problems come to light or new data is collected. Despite the amount of time it takes, there has been surprisingly little research on how to clean data well (Wickham 2014).

Data collection, cleaning, and preparation for computation analysis are, according to Wickham, not popular research topics despite their prevalence across natural science, social science, and digital humanities disciplines. The situation is more complicated, but a full discussion is outside the scope of this project. The important point is data preparation has become ordinary and by being so less interesting and sexy as a research topic. However, as science studies have shown, the ordinary is not as banal as it seems (Lynch 1997).

The first year and a half of this project was spent in collecting, cleaning, and re-shaping data in preparation to generate the topic model. In many computational traditions, this effort is not adequately documented and the knowledge remains apart of the tacit and informal networks of computational researchers. Stopword lists and text processing Perl scripts are emailed between collaborators. These casual interactions are, from an ethnographic perspective, significant modes of circulating information and knowledge.

Documenting and describing this work is crucial to consider when scaling trace ethnography. Data preparation is part of topic modeling as a scalar device, and thus subject to ethnographic description. This chapter addresses the question *what are the methodological dynamics and tensions in generating data from the Open Web?* The first section traces the process of transforming the indeterminate and complex phenomena of blogs into a pile of numbers suitable for training a topic model. The seven steps outlines are the *simplifying assumptions* of the scalar device. The second section presents some brief summary statistics about the data. The final section addresses the current state of web archiving and the documentary practices of studying the web.[46]

What Did I Do?

This first section of the chapter outlines my work to scrape, clean, and format a dataset so it can be consumed by the MALLET topic modeling toolkit. This section is explicit about my work because data cleaning is typically invisible in after-the-fact accounts of academic practice. However, as studies of scientific practice have shown (Garfinkel, Lynch, and Livingston 1981; Latour and Woolgar 1986l Latour 1988; Lynch 1997) significant knowledge work is accomplished behind the scenes. This tendency to focus on findings and knowledge contributions, at the expense of process and practice, has also been the case in the publications of the digital humanities. Matthew Jocker's *Macroanalysis,* for example leaves much to be desired when it comes to descriptions of method and work sufficient to either reproduce findings or apply such methods elsewhere (Jockers 2013).[47]

[46] This chapter is purposefully written to foreground the first person because it is written as a reflexive self-ethnographic narrative. The author cannot be disassociated from the descriptions of practice. To reframe this chapter in a more academic tone and style does injustice to the theoretical and methodological work of ethnographically documenting the simplifying assumptions of the data preparation and topic modeling as a scalar device.

[47] What is a "methods section" and how to write one is a problem for digital humanities. Unlike the natural and social sciences, the concept of a "methods section" is foreign to humanists. The purpose of methods sections aspire to enable reproduction and replication, cornerstones of the scientific method's claim to being able to produce generalizable knowledge. In the digital humanities, reproducing experimental conditions may be as "simple" as assembling a digitized text and algorithms. I put scare quotes around "simple" because such assembly is actually much more complicated than many scholars realize. There are many choices a scholar must make in the process of assembling texts for quantitative analysis. For example how text is tokenized, what stop-words are removed, if infrequent or too frequent words are filtered, can impact an algorithmic analysis.

What is missing from the digital humanities is a tradition of documenting scholarly practice in sufficient detail to not only reproduce the findings or leverage a method, but understand how the accumulation of small decisions in the preparation of a corpus (or collection of images or sound)—the processing of data—impacts a theoretical and epistemological claim a quantitative, algorithmic analysis. That is, not only must we be clear about what an algorithm is doing, we need to be explicit about the data upon which an algorithm is acting. The preparatory work needs to be brought forth into the scholarly discourse, if only as an online appendix, it needs to be available for the evaluation of digital scholarship and for a richer understanding of knowledge production. How can the digital humanities as a field begin to establish a conversation around methods if they don't reveal the accumulation of "whats in/whats out" decisions?[48]

This discussion extends the intellectual justification for this chapter. Thick descriptions of the data practice, both the simplifying assumptions of the data collection, cleaning, and preparation, as well as the assumptions built into the generative model, need to be explicitly articulated as part of the interpretation and analysis work covered in subsequent chapters. Without such justification the following would appear as an exhaustive, if not dull, account of all my work to build a single file containing a term/document matrix of digital humanities blogs. To gloss the production of this file is to ignore a year of trials and tribulations, of hard and soft decisions, and mindless data cleaning. The remainder of the chapter retells a story reconstructed from digital traces of my own research practice transforming a list of domains into a term/document frequency matrix ready for MALLET.

Step One: Enumerating URLs

The first step in transforming blog posts into a topic model was to assemble the individual posts from the blogs listed in the *Compendium*. While the

[48] Methods sections could be a temporary trend in DH, as best practices are established and a clear set of methodologies are settled upon (through something like a textbook).

Compendium lists the URLs of blogs, it does not contain the URLs of the posts belonging to those blogs.

I started with a CSV copy of the *Compendium of Digital Humanities* downloaded from the Digital Humanities Now website on April 13th, 2013.[49] At the time the file contained 615 entries. Each entry had the following pieces of information:

- A blogger's name
- A URL for the blog, personal website, or professional profile
- A Twitter Username
- A link to an RSS feed

The two most important pieces of information are the URL for the blog and the link to their blog's RSS feed. An RSS feeds are useful for discovering new posts, but don't normally link to historical content. My interest was in the history of posts, a list not always available with RSS feeds or other standardized format. This meant I needed to programmatically crawl each domain in the *Compendium* and generate a list of URLs. Not all domains were scraped. I removed entries from the list based on the following criteria:

- Sites in non-English languages
- Sites without a blog or no blog-like content
- Sites that are too large or structurally complex to scrape
- Sites whose domain is dead or return an error message
- Sites that disallow scraping via `robots.txt`

My decision to only scrape English blogs was primarily pragmatic. I can't read Non-English posts they wouldn't be included in my analysis. Topic modeling does a fairly good job of clustering documents in other languages, this won't affect my model, but it will produce unprofitable topics. I haven't eliminated all non-

[49] The DHNow "About" page has a link to a Google Spreadsheet. By changing the `output` HTTP parameter from `html` to `csv` you can download the sheet as a CSV file. https://docs.google.com/spreadsheet/pub?hl=en_US&hl=en_US&key=0AucqXAIBhf_idGNlZzVjSGkxQU9XNU4yb0w1clMxeXc&single=true&gid=3&output=csv

English blog posts, only those blogs I know have very little English content. There are blogs in my dataset that have a small number of posts in non-English languages, those remained in the dataset and were segregated by the topic model.

Not all sites listed in the *Compendium* are blogs. I sometimes came across scholar's Institutional websites and no RSS feed. Other times I encountered a scholar's personal domain and website, but it contained no blog-like content. These sites may have had a blog at an earlier time, but didn't have anything blog-like when I commenced scraping.

Some sites are simply too large to be scraped for their blog content. Sites like HASTAC are very large and complex for my study; they would require significant time and effort to crawl and scrape. These large entries, like HASTAC, were removed because including they are massive ecosystems in their own right.[50]

The web is not an archive. Websites, pages, and domains blink in and out of existence all the time. Several blogs in the *Compendium* used the commercial blog host Postereous. Unfortunately, Postereous was purchased by Twitter and shut down in April of 2013. All of these blogs vanished. This is the danger of delegating such infrastructural work to a commercial third party. Some scholars set up new domains, but some blogs were completely lost. Sometimes this task is too daunting, priorities shift, or who knows what. What results is a dead domain or a website that returns an error. In both cases, commercially or self-hosted, a handful of sites were completely gone. While some of these have been preserved in the Internet Archive, accessing that content is difficult.

Finally, I did not scrape sites that requested no automated scraping or spidering in their `robots.txt`. A robots file is a text file stored at the root of a domain, i.e., http://dancohen.org/robots.txt, that informs an automated web crawler of the permissions it has to crawl a site (or portions of a site). A `robots.txt` file is

[50] HASTAC has received a large (for the humanities) NSF grant to study the social dynamics of the site, I'm not concerned with leaving them out of my study. http://www.hastac.org/blogs/superadmin/2012/08/16/hastac-wins-nsf-grant-study-its-own-social-network

merely a suggestion and a web spider could choose to ignore the convention. Being a polite web crawler, I abided by a domain's `robots.txt (`if it existed) and didn't crawl sites that disallowed automated scraping.

The infrastructure needed to index a domain, a web crawler or spider, is non-trivial to set up and configure. There are some open-source web crawlers like the Heritrix[51] crawler or the wget Unix utility, there are also libraries for the Python programming language, like Scrapy,[52] for developing your own crawlers. While all of these are excellent choices for crawling the web, as a scholar whose primary goal is to extract information from crawled web pages configuring or developing a scraper is a distraction from the essential task of isolating specific features from the crawled and analyzing the extracted content. So, I used a commercial service, Diffbot,[53] to crawl my selected domains and enumerate all of the URLs, do classification, and information extraction (more on this below).

Diffbot provides a combination of crawling and feature extraction via an easy to use web-service API. They manage the crawling and feature extraction on their own servers, relieving me the administration and maintenance of the infrastructure needed to perform these tasks. The price I pay for this convenience is dealing with data quality issues, artifacts of their machine-learning heuristics. I'll discuss the specific problems I faced and my fixes below, but I should emphasize how Diffbot saved me from getting too mired in the excruciating task crafting a web scraper and information extractor for each and every blog in the *Compendium* (a task that would have added at least a year to the data collection).

Using the subset of blog domains I generated from the *Compendium*, I used these as *seed* urls for Diffbot's Crawlbot API.[54] The Crawlbot API indexes a domain generating a list of every single URL belonging to the domain. Diffbot also

[51] https://webarchive.jira.com/wiki/plugins/servlet/mobile#content/view/2800

[52] http://scrapy.org/

[53] http://diffbot.com

[54] http://www.diffbot.com/products/crawlbot/

provides a service for classifying these URLs into different categories and initiating feature extraction on certain URLs. At first I was delighted that Diffbot handled the combination of enumeration, classification, and feature extraction of blog posts, but I discovered it wasn't as useful as I had hoped.

Step Two: Classifying URLs as Blog Posts

Step one produced big lists of URLs for the domains. Unfortunately, not all of those URLs belonged to blog posts. For example, look at this short list of URLs from the blog 4humanities:

```
http://4humanities.org/2013/10/interview-with-sue-gollifer/
http://4humanities.org/2010/12/wordpress-blog-platform-and-content-mana
gement-system/
http://4humanities.org/2013/page/2/
http://4humanities.org/2012/12/challenges-in-humanities-advocacy/
http://4humanities.org/wp-login.php
http://4humanities.org/2013/03/interview-with-cara-ann-simpson/
```

There are several different types of pages identified by these URLs, only some of them are blogs. URLs like the first one, `http://4humanities.org/2013/10/interview-with-sue-gollifer/,` I recognize as blog posts, with the standard WordPress URL structure of `<domain>/<4 digit year>/<2 digit month>/<post title slug>/`. Blog posts on platforms like WordPress with consistent URL patterns are the URLs I want. However there are also URLs that clearly don't belong to blog posts like `http://4humanities.org/2013/page/2/` or `http://4humanities.org/wp-login.php`. These are part of the WordPress infrastructure. The first being pagination URL for posts from 2013 and the second being the standard administrator login URL for WordPress. I don’t want to include these URLs in my dataset.[55]

In total, my first crawl of the *Compendium* sites returned a total of 1,471,471 URLs. This number includes every URL found, including outbound links, for each page of each domain. To isolate just the blog posts I needed a means of disambiguating URLs for blog posts from other URLs so I could extract only blog content. Furthermore, I needed a way to do this *at scale;* I don’t have the time to classify each URL by hand that could take years. To speed up this process I used two approaches, I tried Diffbot’s fully automated Page Classifier, and when that didn’t perform adequately, I turned to the most powerful and infamous tool in data manipulation, regular expressions.

Classifying URLs with Machine Learning

Diffbot’s Page Classifier[56] integrates with its Crawlbot API and to assign a confidence score to each URL for a set of categories such as *article*, *image*, *product*, *anddiscussion*. Diffbot’s classifier examines the rendered look of the page, its the structure and content of the text, and the structure of the page itself. Blog posts, and more generally “articles“ have a *look,* a pattern that the developers at Diffbot have used to train their classifier. Being a heuristic, and a black box, I don’t have the ability to tune the classifier to my data. This means I had to rely upon the accuracy of Diffbot’s infrastructure and that created data cleaning headaches.

Diffbot’s classifier gave me over two hundred fifty thousand blog posts from the larger set of 1.4 million. I knew Digital Humanities scholars loved to blog, but this was a still an astonishing number. When I began to dig into the results I found

[55] Every blogging platform has its own URL design which acts like a signature when enumerating URLs. There are URLs designed for humans and there are URLs design for the technical plumbing of the platform. When analyzing, curating, and archiving the Open Web, the URL is the atomic unit. Good URL design practice is best summed up as *good URLs don’t change.* http://www.w3.org/Provider/Style/URI.html

[56] http://www.diffbot.com/products/automatic/analyze/

that the computer was doing *exactly what it was told*. When enumerating and classifying URLs, it finds both blog posts *and* any permalinks URLs associated with comments on that post.

```
http://4humanities.org/2012/10/distance-learning-in-india/
http://4humanities.org/2012/10/distance-learning-in-india/?replytocom=2
2048
http://4humanities.org/2012/10/distance-learning-in-india/?replytocom=2
1192
```

Diffbot treats these as separate and distinct URLs and classifies them independent of each other, despite the fact they resolve to the same blog post. So for blog posts that are really popular and have a lot of comments, each of the comments gets their own URL and each of those URLs are classified as an “article“ resulting in a large number of duplicated posts. Eliminating duplicated from data is not that complicated, but if I wanted approximate counts of blog-posts-per-domain, these URLs will skew the post count. Blogs whose posts garner few comments (or blogs that don’t use the permalinks system) would report an accurate number of posts whereas popular blogs would have a skewed number of blog URLs, one for each comment on the post.

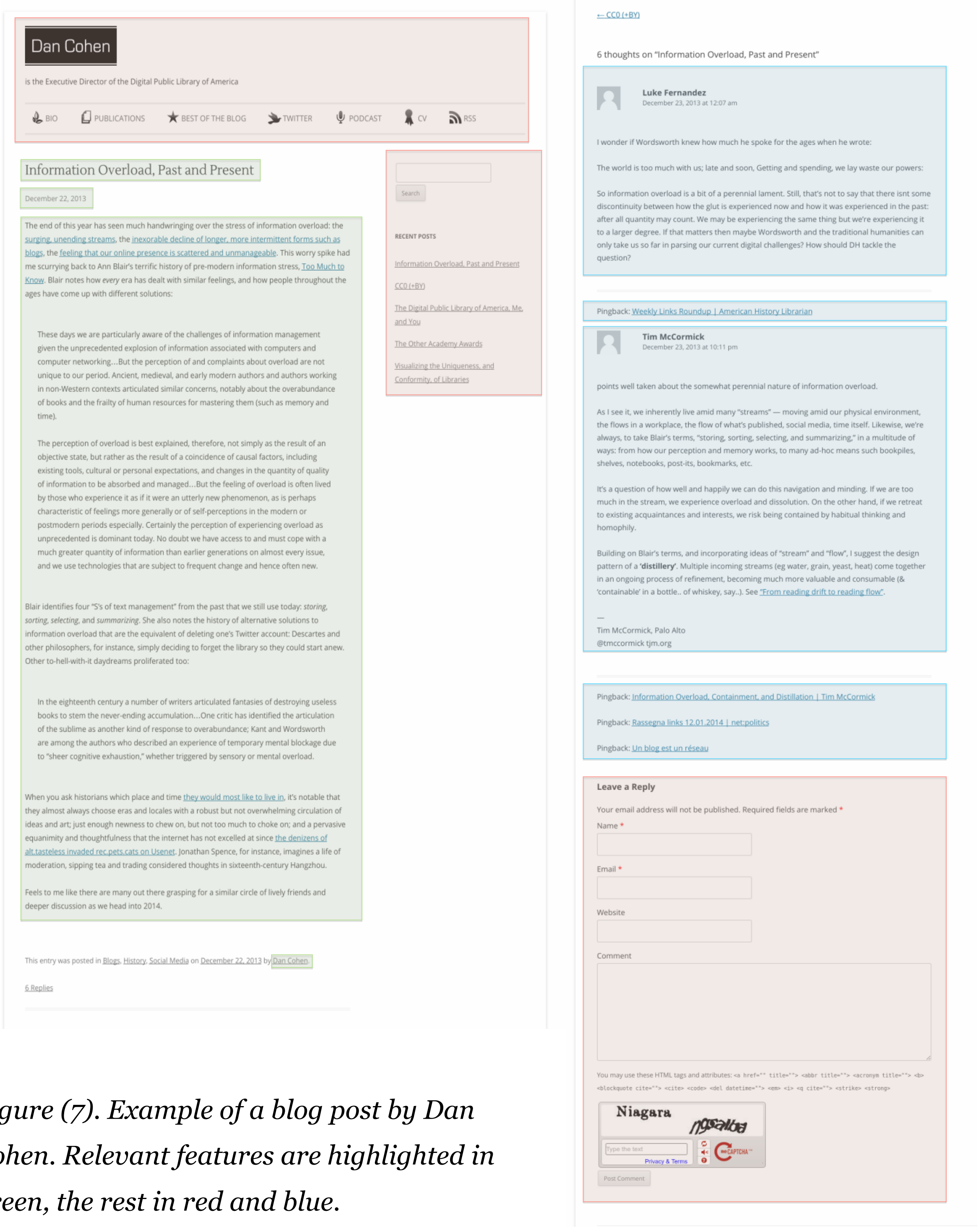
Identifying Blog Posts with Regular Expressions

There are many ways to eliminate duplicate documents from the dataset and I employ some of the more automated and computational techniques below, but I also employed a labor-intensive technique so that I could be sure to specifically isolate URLs for a blog post. At this point in the analysis, URLs are my fundamental unit; I am not yet working with the text content of posts. In a perfect dataset a single URL would identify each post. However, as is the case with the WordPress comment system, I could end up with URLs representing comments on a popular blog post. I don’t want comment URLs in my dataset. Scraping those URLs would result in duplicated text and skew the number of posts per blog and the distribution of blog posts over time.

For each of the blogs in my *Compendium* list I have a list of all the URLs Diffbot crawled on that domain. Some of those URLs belong to blog posts and some of them do not. Other URLs are "About Me" pages, course pages, or the plumbing of blog platforms. I want to remove those unnecessary URLs and focus specifically on those representing blog posts. To do this I decided to craft a custom regular expression for each blog. This regular expression would target URLs that represented blog posts, and ignore everything else. There were seven general patterns I was able to replicate across the dataset, so in practice I wasn't actually crafting a unique regular expression for each blog as much as selecting one of the general regular expression patterns and tailoring them for the specific domain.

Step Three: Extracting Text Content

Each blog post contains a variety of information, only some of which is relevant to my study. The challenge in transforming blog posts into data is distinguishing and extracting the *features* relevant to my study and ignoring the rest. The example image in figure (7), a blog post from Dan Cohen's blog, highlights a set of features of a blog post.



Dan Cohen

is the Executive Director of the Digital Public Library of America

Information Overload, Past and Present

December 22, 2013

The end of this year has seen much handwringing over the stress of information overload: the surging, unending streams, the inexorable decline of longer, more intermittent forms such as blogs, the feeling that our online presence is scattered and unmanageable. This worry spike had me scurrying back to Ann Blair's terrific history of pre-modern information stress, Too Much to Know. Blair notes how *every* era has dealt with similar feelings, and how people throughout the ages have come up with different solutions:

> These days we are particularly aware of the challenges of information management given the unprecedented explosion of information associated with computers and computer networking…But the perception of and complaints about overload are not unique to our period. Ancient, medieval, and early modern authors and authors working in non-Western contexts articulated similar concerns, notably about the overabundance of books and the frailty of human resources for mastering them (such as memory and time).
>
> The perception of overload is best explained, therefore, not simply as the result of an objective state, but rather as the result of a coincidence of causal factors, including existing tools, cultural or personal expectations, and changes in the quantity of quality of information to be absorbed and managed…But the feeling of overload is often lived by those who experience it as if it were an utterly new phenomenon, as is perhaps characteristic of feelings more generally or of self-perceptions in the modern or postmodern periods especially. Certainly the perception of experiencing overload as unprecedented is dominant today. No doubt we have access to and must cope with a much greater quantity of information than earlier generations on almost every issue, and we use technologies that are subject to frequent change and hence often new.

Blair identifies four "S's of text management" from the past that we still use today: *storing*, *sorting*, *selecting*, and *summarizing*. She also notes the history of alternative solutions to information overload that are the equivalent of deleting one's Twitter account: Descartes and other philosophers, for instance, simply deciding to forget the library so they could start anew. Other to-hell-with-it daydreams proliferated too:

> In the eighteenth century a number of writers articulated fantasies of destroying useless books to stem the never-ending accumulation…One critic has identified the articulation of the sublime as another kind of response to overabundance; Kant and Wordsworth are among the authors who described an experience of temporary mental blockage due to "sheer cognitive exhaustion," whether triggered by sensory or mental overload.

When you ask historians which place and time they would most like to live in, it's notable that they almost always choose eras and locales with a robust but not overwhelming circulation of ideas and art; just enough newness to chew on, but not too much to choke on; and a pervasive equanimity and thoughtfulness that the internet has not excelled at since the denizens of alt.tasteless invaded rec.pets.cats on Usenet. Jonathan Spence, for instance, imagines a life of moderation, sipping tea and trading considered thoughts in sixteenth-century Hangzhou.

Feels to me like there are many out there grasping for a similar circle of lively friends and deeper discussion as we head into 2014.

This entry was posted in Blogs, History, Social Media on December 22, 2013 by Dan Cohen.

6 Replies

Information Overload, Past and Present

CC0 (+BY)

The Digital Public Library of America, Me, and You

The Other Academy Awards

Visualizing the Uniqueness, and Conformity, of Libraries

← CC0 (+BY)

6 thoughts on "Information Overload, Past and Present"

Luke Fernandez
December 23, 2013 at 12:07 am

I wonder if Wordsworth knew how much he spoke for the ages when he wrote:

The world is too much with us; late and soon, Getting and spending, we lay waste our powers:

So information overload is a bit of a perennial lament. Still, that's not to say that there isnt some discontinuity between how the glut is experienced now and how it was experienced in the past: after all quantity may count. We may be experiencing the same thing but we're experiencing it to a larger degree. If that matters then maybe Wordsworth and the traditional humanities can only take us so far in parsing our current digital challenges? How should DH tackle the question?

Pingback: Weekly Links Roundup | American History Librarian

Tim McCormick
December 23, 2013 at 10:11 pm

points well taken about the somewhat perennial nature of information overload.

As I see it, we inherently live amid many "streams" — moving amid our physical environment, the flows in a workplace, the flow of what's published, social media, time itself. Likewise, we're always, to take Blair's terms, "storing, sorting, selecting, and summarizing," in a multitude of ways: from how our perception and memory works, to many ad-hoc means such bookpiles, shelves, notebooks, post-its, bookmarks, etc.

It's a question of how well and happily we can do this navigation and minding. If we are too much in the stream, we experience overload and dissolution. On the other hand, if we retreat to existing acquaintances and interests, we risk being contained by habitual thinking and homophily.

Building on Blair's terms, and incorporating ideas of "stream" and "flow", I suggest the design pattern of a **'distillery'**. Multiple incoming streams (eg water, grain, yeast, heat) come together in an ongoing process of refinement, becoming much more valuable and consumable (& 'containable' in a bottle.. of whiskey, say..). See "From reading drift to reading flow".

—
Tim McCormick, Palo Alto
@tmccormick tjm.org

Pingback: Information Overload, Containment, and Distillation | Tim McCormick

Pingback: Rassegna links 12.01.2014 | net:politics

Pingback: Un blog est un réseau

Leave a Reply

Your email address will not be published. Required fields are marked *

Name *

Email *

Website

Comment

You may use these HTML tags and attributes: `<a href="" title=""> <abbr title=""> <acronym title=""> <b> <blockquote cite=""> <cite> <code> <del datetime=""> <em> <i> <q cite=""> <strike> <strong>`

Niagara ngsalba

Figure (7). Example of a blog post by Dan Cohen. Relevant features are highlighted in green, the rest in red and blue.

The green boxes are the features of the blog post I want. They represent the title of the post, the date, the text content, and the author. The blue boxes highlight blog comments. Comments are an important sociotechnical feature of blogs and should play a role in any analysis of scholarly discourse on blogs. However, comments are technically difficult to isolate and extract at scale. Because of the technical and logistical issue of extracting comments, I chose not to include them in the quantitative analysis.

In addition to crawling domains, Diffbot can scrape the semi-structured contents of a web page, parse the contents, and return structured and classified data. The main function is extracting the text-body of an "article-like" web page, but it will also try and extract metadata like author, date, and post title. In an evaluation of text extraction algorithms, Diffbot performed well on precision and recall tests.[57] Diffbot also has a well-designed and easy to use REST API that I found appealing when I was making decisions about web crawling and extraction infrastructure.

Diffbot is a black box; I do not fully know *how* it identifies and extracts certain features from that page.[58] Diffbot uses algorithms from machine learning and computer vision to identify page elements, but its documentation does not reveal the details. What this means in practice is the quality of structured data and metadata is not 100 percent perfect. The quality of information extracted using the Diffbot API is good, but there are still many posts where the heuristic triggered on non-content page elements. Short posts, or posts containing just a few links, had a tendency to trip up the extraction function, although the algorithms improved over time (which can be problematic for algorithmic transparency reasons). The biggest problem with the text data was a recurring issue where "About Me" pages were concatenated to each post's text. This was a

[57] The original post, "Evaluating Text Extraction Algorithms," by Tomaž Kovačič has been taken down.
https://web.archive.org/web/20111219131334/http://tomazkovacic.com/blog/122/evaluating-text-extraction-algorithms Kovačič also wrote an unpublished thesis on the subject
http://eprints.fri.uni-lj.si/1718/1/Kovacic-1.pdf

[58] While Diffbot remains a black box from a technical perspective, this does not mean Diffbot is an impenetrable Borg cube. Its algorithms might be a trade secret (or at the vary least a competitive advantage), but the folks who work at Diffbot have been incredibly gracious in allowing me to use their APIs in the service of scholarship (for free). Furthermore, they have been responsive and accommodating whenever I had a question about the service or encountered bugs.

per-blog, not a per-post, issue meaning all posts from a scholar’s particular configuration on their blogging platform caused Diffbot to trip up.

The topic model was most effective at identifying blogs with the Diffbot concatenation issue. The extracted text content from these blogs began with unique content, but ended with the same collection of words (the “About Me“ page concatenated to the end of each post). This is just another of the many blackbox issues in using Diffbot that made data cleaning, analysis, and interpretation of the model more nuanced. Only 21 of the 396 blogs were affected.

I should emphasize the quality of most of the data I received from Diffbot was very good. Automatic and unsupervised feature extraction to get data and metadata from web pages is extremely challenging. While web pages, like blog posts, are structured entitities, the underlying HTML structure does not always express sufficient semantics to explicitly signify the intent and purpose of a document’s features. While there are specific tags in the HTML5 standard for signaling the semantics of the content they encapsulate (the `<article/>` tag, for example), more frequently web designers and developers use generic structural tags, such as the `<div/>` and `<span/>` elements, when composing HTML pages. These tags are “meaningless“ structural elements for styling or providing interaction hooks that don’t indicate document semantics. This makes the process of separating the meaningful content from the structural scaffolding a challenge. Web developers and designers, and academic bloggers, each have their own bespoke designs for their blogs, which don’t generally follow guidelines for semantic HTML.

This highlights an infrastructural tension when trying to computationally analyze blogs and the Open Web. While the underlying HTML documents have a definitive structure (expressed in HTML tags) and are programmatically rendered using CSS to specify the visual style, there are no explicit requirements for how to compose those elements into documents.[59] This means extracting information from semi-structured web pages is more art than science. Furthermore, it means extracting information from semi-structured web pages *at scale* must be a blend of art *and* science.

Step Four: Data Cleaning

In data mining and data driven research, it is said “, "10 percent is analysis, 90 percent is data cleaning.” This feels about right. It was a non-trivial amount of effort to wrangle my data into a form that was good enough for topic modeling. There is always more cleaning to do, but given the scale they are few and hard to find. As I have discussed above, blogs are not a standardized format and Diffbot’s extraction heuristics are not perfect. Data cleaning, where a bulk of the work of transforming blogs into data occurs, isn’t pretty or sexy, but it is incredibly important to document because the data-cleaning stage involves many decisions about selection and curation.

The data I received from Diffbot, while nicely packaged in JSON, wasn’t necessarily in a format ready to be consumed by MALLET. Beyond the formatting, there was still a significant amount of noise in my dataset. The two foremost issues were fixing the dates and removing duplicates from the dataset.

Cleaning Dates

Diffbot’s article extraction API does a fairly good job of finding the correct date for a given blog post, but it is not perfect. Before cleaning, I had a very curious distribution of dates. For example, there was a very high concentration of posts in 1912, which is impossible. Diffbot identified historical dates within the pages and assigned those as the creation date. This shows the difficulties involved with automated metadata extraction.

[59]This expressive capacity gives HTML its powerful flexibility (and makes implementing web browsers difficult), but results in few standards and many best practices for authoring web documents (if you were to push the implicit designs of semantic HTML to its logical limit you end up with a web page like http://motherfuckingwebsite.com/.

Diffbot appears to use entity extraction to identify and isolate dates from the contents of a page, but sometimes there are multiple dates in the page. The ranking system they are using does not always select the correct date, sometimes I found date metadata assigned to the blog post that were part of the content (particularly short posts), while the correct date was in a different location on the page.

Figure (8) below shows posting rate stabilized around a thousand per month in 2008. The graph also shows a massive spike up to five thousand posts at the very end of 2013. This either implies a massive influx of bloggers or dirty metadata. Figures (9) and (10) show the process of zooming into the data to find the source of the spike.

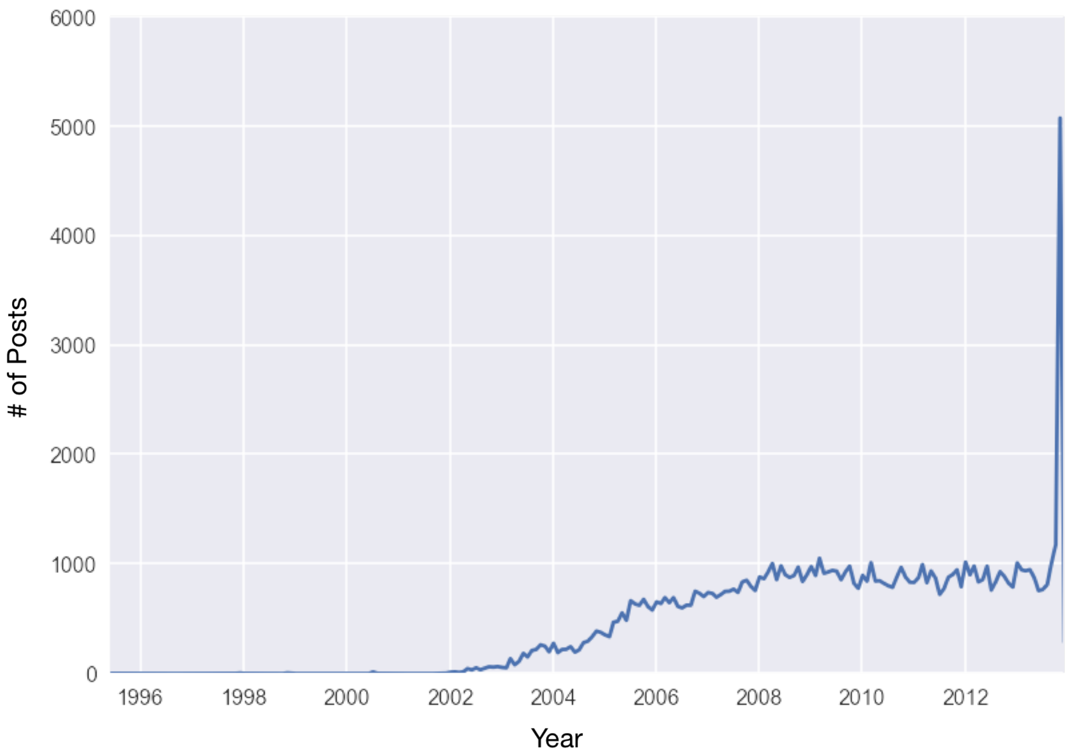


Figure (8). This graph shows total number of blog posts per year for the entire Compendium. The spike is suspicious and indicates dirty data.

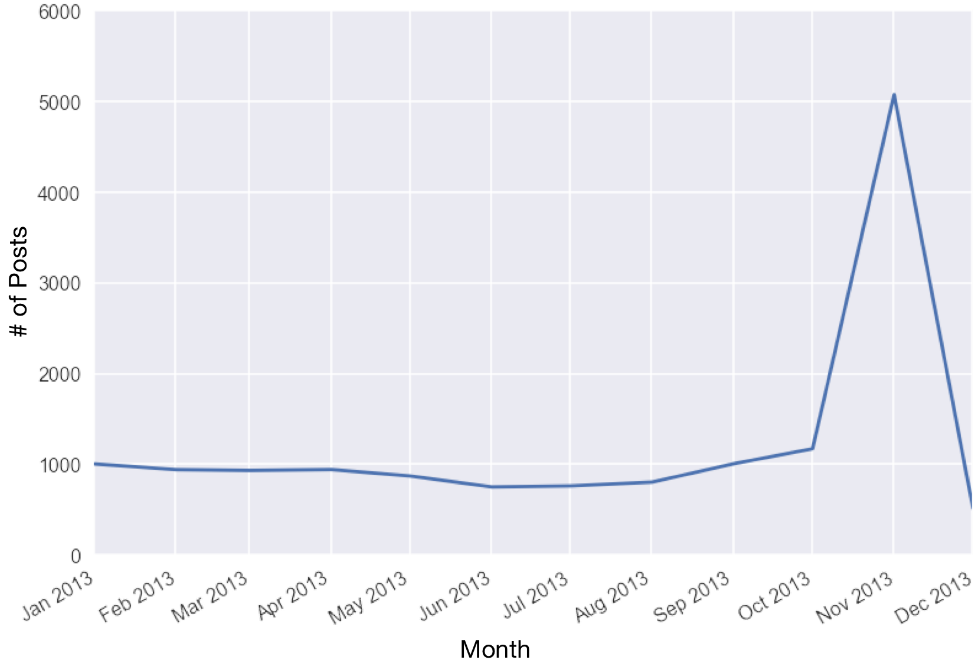


Figure (9). Compendium posts per month for 2013. The spike appears in November 2013.

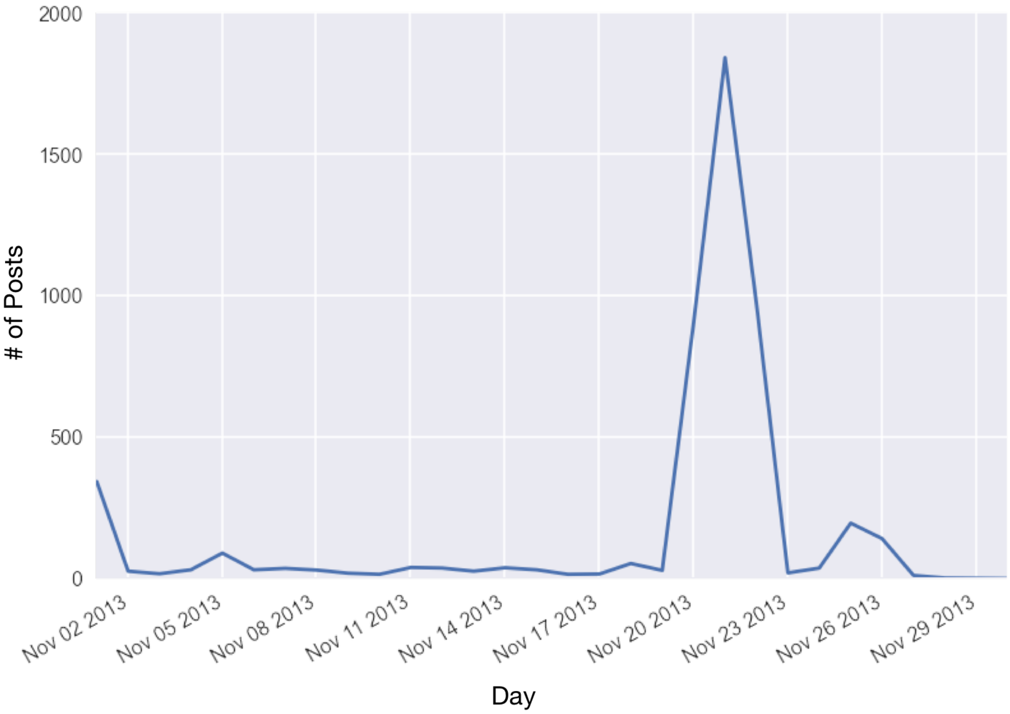


Figure (10). Compendium posts per day for November 2013. The spike occurs around November 21st.

Figure (10) shows a flood in blog posts on November 20th, 21st, and 22nd. A closer look showed several domains (`nataliacecire.blogspot.com`, `grandtextauto.org`, and `ideo.org`) had dirty date metadata. Fixing the dates was relatively easy because each of these sites use the standard WordPress URL design that includes the year, month, and day in the URL body. A bit of regular expression and Python scripting fixed the issue. But, I do not know why those dates got assigned to those posts. The black box strikes again.

Never Trust a Trace

There was one blog, *discontents.com.au* by Tim Sherratt, in which posts with "impossible" dates, but that were not due to dirty Diffbot data. When I visited his site I found posts with "correct" metadata (as expressed by WordPress) extending back as far as November 15th, 1985.[60] Some of these posts pre-dated the web and the open Internet, so how could these blog posts have such a date? This example reveals a particularly interesting aspect about infrastructure and "ground truthiness."

In the case of *discontents.com.au*, Diffbot correctly extracted the date metadata. If you visit the site, the dates listed on the site's blog posts extend back into the mid-'80s. How is this possible? The owner of the domain must be managing his own installation of WordPress instead of using a hosted instance at Wordpress.com. By managing his own instance he has complete control over the database, the data, and how the blog is represented on the web. With this level of control, Sherratt could modify a post's metadata and specify any value in the past or future. These dates may or may not indicate when the content was actually created, but we *do* knowthat WordPress, MySQL, and the Web did not exist in the mid '80s. We have good data from an information extraction perspective—that

[60] I think the "*Australian scientists at the British atomic tests*" post is a narrative excerpt from a book. http://discontents.com.au/australian-scientists-at-the-british-atomic-tests/

is, the Diffbot algorithm worked properly, but it is uncertain how accurate that data can be in a larger sense.

Finding Duplicates

On occasion, Diffbot captured duplicate content. So, I computed the cosine similarity of all documents per blog. Cosine similarity is a technique that compares documents, represented as word vectors projected into an N-dimensional space (N being the number of word features in the corpus). Imagine a high-dimensional Cartesian space with lines projected from an origin to a point representing a document (based on its word features). Document vectors that share word features will be closer together in this Cartesian space because the lines would be projected into similar areas of the high-dimensional space. By computing the cosine between these two vectors we can get a measure of document similarity.

I computed a pairwise comparison of cosine similarity only for posts within a blog—that is, I only compared posts on dancohen.org to other posts on dancohen.org. Computing cosine similarity for the entire corpus would be useful for search and information retrieval tasks, such as finding similar posts across blogs, but I wanted to see duplication errors within each domain. Pairwise measures of similarity can be used to create visual representations of the dirty data.

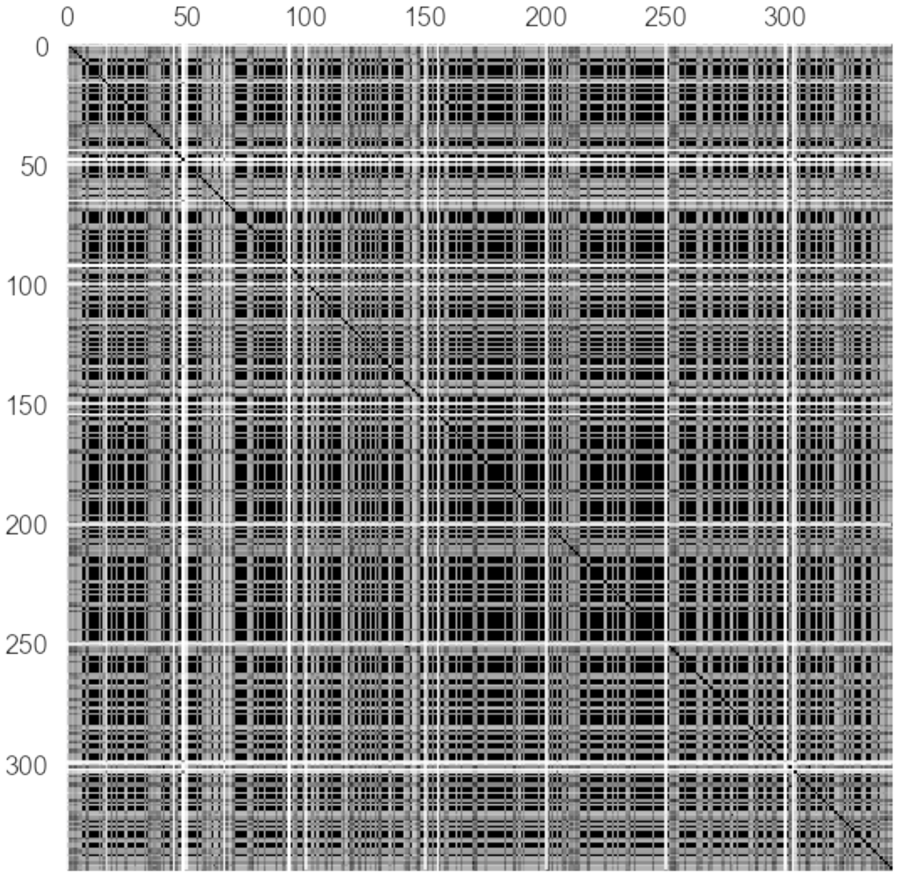


Figure (11). A matrix visualization showing dirty data. Darker points indicate more similarity.

Each point of the matrix in Figure (11) is a pairwise comparison of the cosine similarity of two blogposts from the blog *Ancient World Bloggers*. Darker indicate a higher cosine value, which means the two posts are similar. The black diagonal line shows comparisons of posts to themselves. The matrix shows data with a lot of duplication, this particular blog was difficult for Diffbot to scrape.

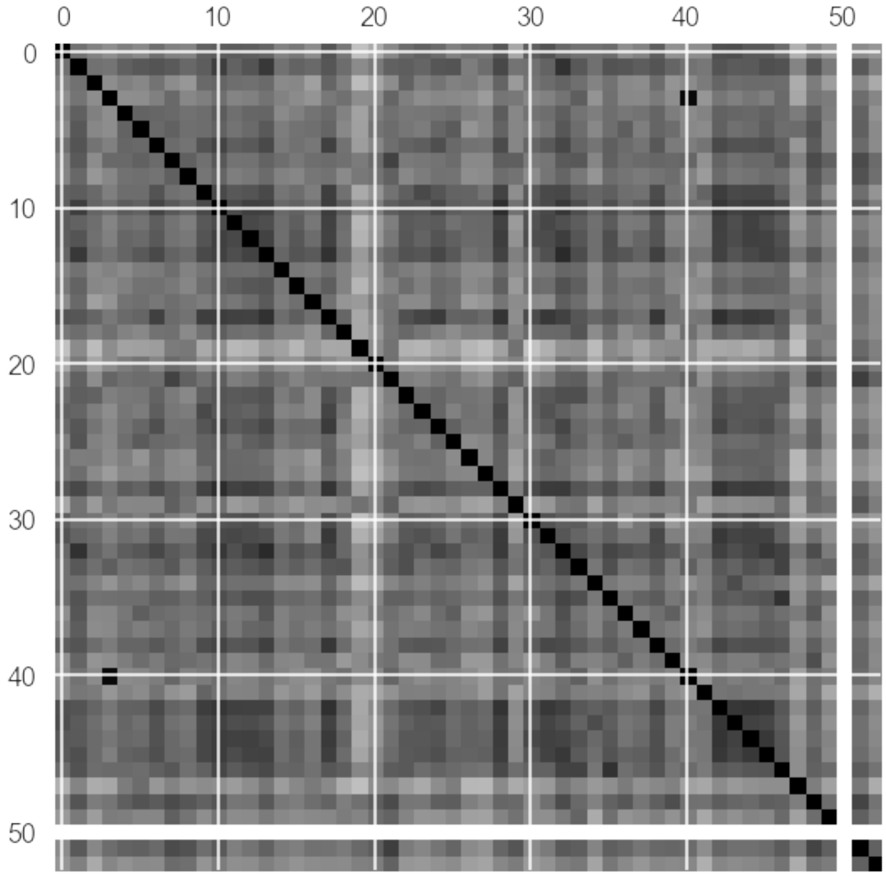


Figure (12). A matrix visualization showing clean data. Notice balance of light and dark gray, this indicates no duplicates.

Figure (12) shows the computed similarity matrix for Quinn Dumbrowski's blog. There are no black points, except the diagonal, which means there are no duplicated posts from the scrape. On a whole, the image appears dark, which would indicate many of the posts are somewhat similar. The density of this matrix is lower because the number of posts on Quinn Dumbrowski's blog is much less than on *Ancient World Bloggers*.

After computing the comparison matrices, I found 13,202 blogs with cosine similarity above .99 or 99 percent. I removed them from the corpus.

Step Five: From Text to Bags of Words

At this point the blog posts were nearly ready for topic modeling, but there was still more data manipulation required to compute a model for the corpus. The data cleaning processes described in the previous steps were at the *document* level, in this phase of the process I needed to clean at the *word* or *token* level. Before I can feed my documents to the MALLET topic model, I must *tokenize* the documents and *filter* stopwords and words with extreme frequencies. This document preparation workflow comes from the Gensim topic modeling library.[61] I used Gensim's wrapper for running MALLET, this allowed me to use the Gensim document preparation functions as well as automate the execution of MALLET from within an IPython Notebook.

Tokenization

While tokenization isn't a very sexy topic, it needs to be mentioned because it generated the fundamental unit of my analysis, the word token. My documents needed to be transformed from sequential text into list of individual words. I used Gensim's tokenizer function, which breaks up text into a set of tokens well suited for text analysis. For example, the Gensim tokenizer takes the following string of text:

```
"This is Matt Burton's awesome tokenizer test; it is good."
```

And return a list of thirteen word tokens:

```
[“this,” “is,” “matt,” “burton,” “s,” “awesome,” “test,” “of,” “the,”
“tokenizer,” “it,” “is,” “good”]
```

There are several things to note in this example, the lowercasing, the removal of punctuation, and the preservation of the possessive *s*. Different tokenizer

[61] Gensim is an open source topic modeling library for the Python programming language maintained by Radim Řehůřek. http://radimrehurek.com/gensim/

functions will break up the string in different ways, such as including punctuation or including or dropping the possessive.

I normalized word tokens to lowercase so that I wouldn't end up with "Digital" and "digital" or "Humanities" and "humanities" counting as separate word tokens. Sometimes capitalization matters, for example if you are modeling a corpus of nineteenth-century German periodicals upper and lowercase words can have different meanings. Whether or not to lowercase should be driven by the specific needs of the analysis and the particularities of the texts. In my case, distinguishing capitalized words from their lowercase counterparts doesn't make a significant difference, as the meaning of both terms is typically the same.[62]

Unlike other common Python language tokenizers, like the standard word tokenizer in the NLTK library, the Gensim tokenizer removes punctuation. Because the goal is to generate a bag of words for topic I didn't want to preserve any of the punctuation in the corpus. Period or comma tokens carry absolutely no meaning when they have been taken out of context.

One artifact of the Gensim tokenizer is the handling contractions. While it might seem strange to allow the possessive "s" to remain in my corpus of tokens, in practice it was removed in the token filtering phase.

Filtering Tokens

The next step in transforming the corpus into a bag of words is to build a dictionary, or an index, of every unique word in the corpus. Building a dictionary calculates the least and most popular words in the corpus and enables filtering based on word frequencies. Technically, a dictionary provides a mapping between an index value (a number) and a word token; basically a giant lookup table. This is the moment in the data processing pipeline when words are quantified and transformed into number for computational analysis. The dictionary is the object

[62] While there could be cases where the only way to distinguish proper names from their common word counterparts is capiltalization, accommodating such unique cases is too costly of a cleaning process for the benefit of preserving them.

that preserves the ability to *go back* to words (and subsequently human comprehensibility) after the topic model has been trained.

Building a dictionary involves iterating over each document in the corpus, tokenizing the document, and iterating over each token adding new tokens to the dictionary or incrementing the count of known tokens. With the dictionary, I could further clean and prepare the corpus for topic modeling by filtering stopwords and filtering extremes.

I used an English language stopword list provided by MALLET, a popular Python library for natural language processing.[63] This list contains 123 high-frequency word tokens whose meaning is lost when they are removed from their sequential context (such as a term-frequency matrix). Stopwords introduce junk into the model producing topics that carry no meaning.

Beyond stopwords, it is often good practice to filter out words with extreme document frequencies. Filtering high-frequency tokens catches corpus-specific stopwords that weren't in the MALLET list. Filtering low-frequency tokens removes unnecessary data that wouldn't inform the topic model and reduces computational overhead. Gensim provides a `filter_extreme()` function that performs three filtering processes. First, it removes words that only occur in *n* documents or fewer. The default value is 5, meaning words that appear in five or fewer documents are removed. Because of the size of my corpus, I opted to remove words occurring in ten or fewer documents. Second, the `filter_extremes()` function can remove frequently words based on what percentage of documents they appear within. The default is to filter any word appearing in more than 50 percent of the documents. Filtering high-frequency words has a similar effect to removing stopwords (as stopwords are typically high frequency words), but is fitted to the specific characteristics of the corpus. I opted to go with the default of 50 percent. The third and final step sets a cap on the

[63] MALLET stopwords can be found in the source code on GitHub: https://github.com/mengjunxie/ae-lda/blob/master/misc/mallet-stopwords-en.txt

total number of unique tokens. The default is to keep the top hundred thousand most frequent tokens.

Before filtering I had a dictionary of 340,827 unique tokens. After filtering I had 53,406 unique tokens.

I used the filtered dictionary to transform each document from a string into a vector; the computable representation of the documents used by Gensim and MALLET. The vectors, which are a list of numbers representing the count of word tokens for a particular index (that maps to the dictionary), can be combined to create a term document frequency matrix representation of the corpus.

Recent research has shown that topic modeling performs poorly with short documents (Tang et al. 2014). While I would have hoped every blog post in my corpus was a long, thoughtful commentary on digital humanities or scholarly communication, the reality is many posts are short one-sentence links to other resources on the web. Additionally, there is still some noise due to the Diffbot text extraction heuristics.

To ensure I trained my model with reasonably good document vectors I decided to filter out any document with a length of fewer than a hundred word tokens. This eliminated 66,828 documents, more than half, from the set of documents to be modeled.[64]

Step Six: Training the Model (Finally)

Only after a significant amount of data collection, cleaning, and preparation was I able to train a topic model. I will be covering the topic model itself in the next chapter, so I briefly summarize the parameters used to train the MALLET model

[64] At first it might seem drastic to remove more than half of the documents from the corpus, but short documents are not very meaningful from both a quantitative and qualitative perspective. This removal is one of the simplifying assumptions of topic modeling as a scalar device, it foregrounds certain features at the expense of others. A different scalar device, like network analysis, might focus on short documents to model the linking behaviors of DH bloggers.

below. MALLET uses the Gibbs sampling as its inference algorithm. Hyper-parameters are continually adjusted based upon the optimization interval.

- Number of documents: 39,976
- Number of topics: 100[65]
- Number of Iterations: 2,000
- Optimization Interval: 20

Summary of the Data

This section answers a series of questions about the collection.

- Total entries in the raw *Compendium* spreadsheet: **615**
- Blogs included in the study: **396**
- Total blog posts scraped: **106,804**
- Total blog posts after filtering duplicates: **91,436**
- Total blog posts after filtering documents with less than a hundred words: **39,976**
- Total blog posts per year:

[65] The determination of the number of topics was a result of iterative testing based around the mapping between topics and interpreted themes. This is discussed in the next chapter.

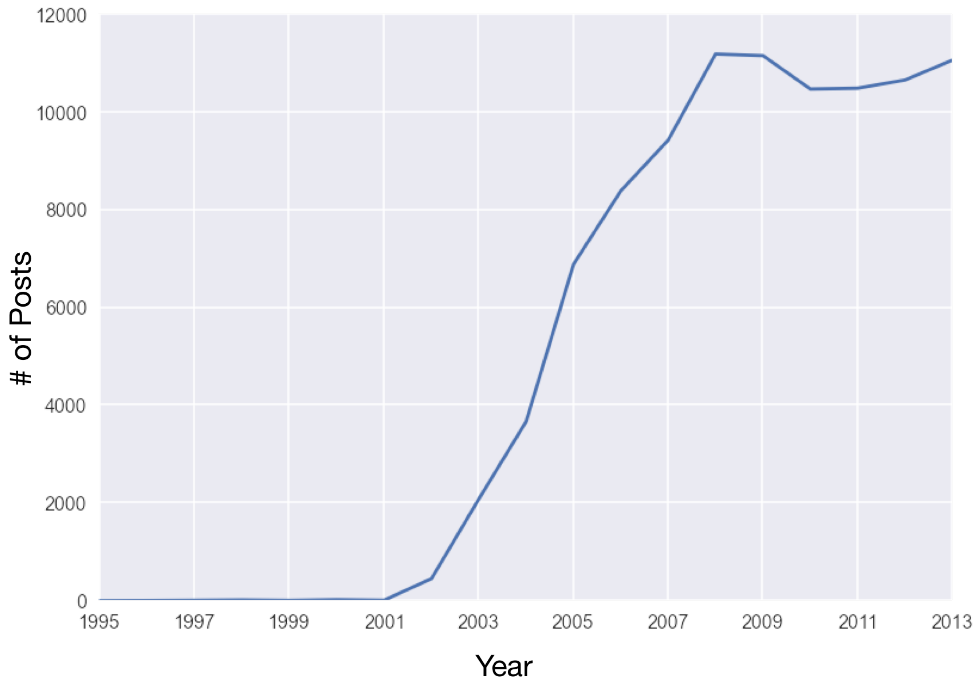


Figure (13). Number of blog posts per year.

- Total number of posts in the top ten blogs:

Table 1: Most prolific domains

Count	Domain
3,581	openculture.com
3,187	askpang.typepad.com
3,158	ancientworldonline.blogspot.com
3,107	lonewolflibrarian.wordpress.com
2,711	infocult.typepad.com
2,614	virtualpolitik.blogspot.com
2,565	digital-scholarship.org
2,460	flowingdata.com
2,148	zephoria.org
1,971	acephalous.typepad.com

- Average length of blog posts:

The average length of posts in my dataset is 388 words. This represents 388 word tokens *before* filtering stop words and extremes. This is the number of word tokens the author of the post might see in their text editor.

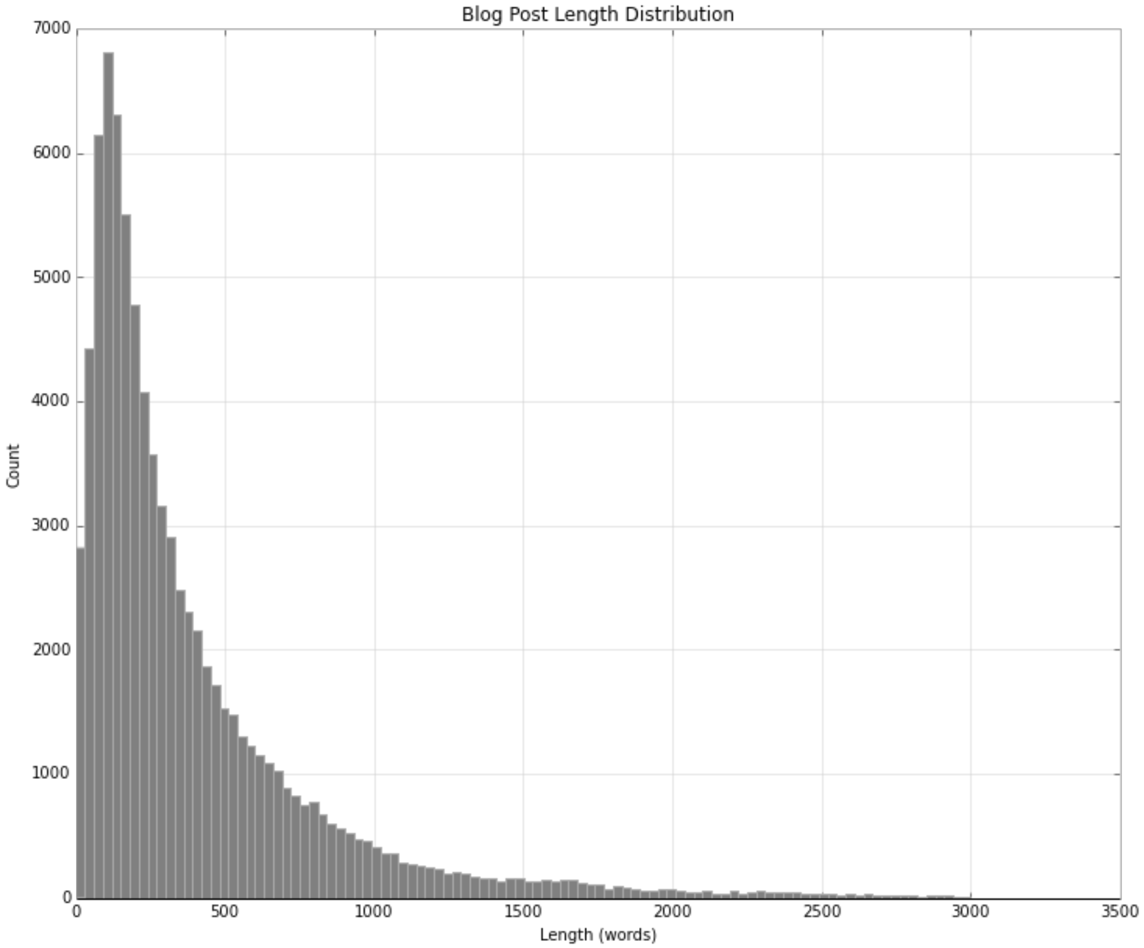


Figure (14). Distribution of post length.

The graph in Figure (17) above shows a distribution of the lengths of blog posts. What is shown in the graph is that many of the posts are less than five hundred words, with a peak around two hundred words. There is however a long tail of longer posts stretching to three thousand words, a cutoff I specified when generating the graph. I purposefully choose to only show the frequency of post length from one to three thousand to preserve the legibility of the graph.

- Who wrote the first blog post?

The very first blog post in my dataset belongs to danah boyd who is one of the earliest bloggers, and one of the first scholarly bloggers (she was an undergraduate when she started her blog in 1997). The post is a very intimate representation of a discussion with her partner at the time and how she distinguishes her online and off-line identities.[66]

- Who blogs the most?

The blog *openculture.com* has 3,581 posts. Open Culture is a group blog dedicated to collecting "the best free cultural and educational media on the web." It appears to be a volunteer effort involving professors, k-12 educators, journalists, technologists, and writers. It is a curated aggregation of free and open content including movies, books, classes, and language resources from across the web.

The Work of Research in the Age of Mechanical Reproduction

The information technologies enabling digital humanities scholars to blog also enable researchers to access and analyze blogs. As I have described above, the *visibility at scale* afforded by blogs on the Open Web have made it possible to transform discursive practice into data to be quantitatively manipulated by scalar devices and qualitatively interpreted through trace ethnography. This process of transforming blogs was non-trivial and included a series of decisions dutifully documented in the six steps above. This extensive level of detail is extremely important because the methods being used are still experimental. As interpretive, yet computational methodologies like distant reading mature and the details are socialized within the community of practice, the level of granularity will probably change.

[66]http://www.zephoria.org/thoughts/archives/1997/06/09/1.html

For the community of scholars who study the Open Web using computation, be they in the digital humanities or interpretive social sciences, there are two methodological and theoretical issues. First, there is the problem of stabilizing the web itself as a research object. The web is dynamic and difficult to isolate. This project used the *Compendium of Digital Humanities* as the set of blogs to investigate; yet that sample is not solid. Some blogs had moved, others had disappeared all together. The digital humanities community did not stop blogging on account of this research, I had to set a stopping point and ignore lots of interesting discussions. Other researchers interested in studying the same phenomena will not encounter digital humanities blogs as I encountered them. The blogosphere has permanently changed and the efforts for archiving that do exist are not adequate for computational researchers.

Second, beyond the data, computational research practice leaves a trace. Workflows and practice can be documented in code. That code becomes a documentary trace of the practice and methodology. New systems like the IPython Notebook allow researchers to capture, organize, share, and possibly preserve data driven and computational workflows. These “scientific notebooks” to facilitate my analysi by constitute a set of documentary traces that can be mixed together with the blogs and the topic model as a collection of data for trace ethnographic analysis.

Scrape Web Archives not the Web

The Open Web is not data. As the last section of this chapter demonstrated, there is a significant amount in transforming the web into data. The five steps, enumerating URLs, classifying blog posts, extracting text content, cleaning, and creating word vectors, are unique to the research questions, methods, and processes of this project. Other scalar devices, such as network analysis, would require a different set of step. This would be going back to the sources on the web and starting a new data preparation process. This is problematic because of the dynamic nature of the web, for some sites there is no source to “go back to” and because I didn’t create an archive of websites or blog posts in the study, the particular configuration of the digital humanities blogosphere I studied no longer exists.

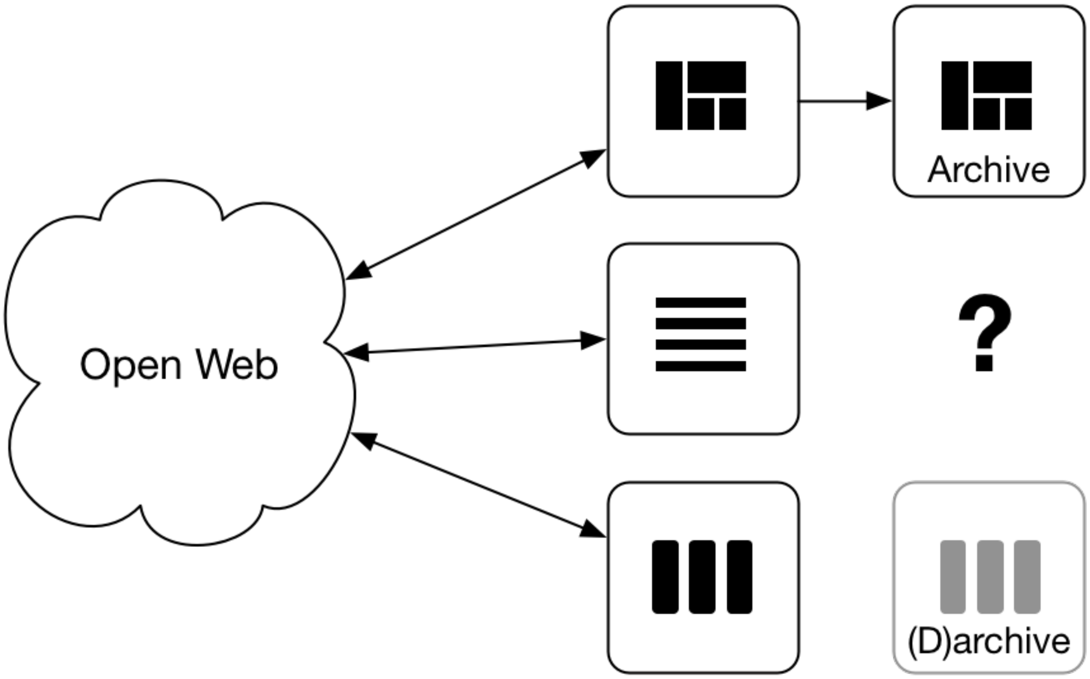


Figure (15). The current flow of data for web research

The current practice in web research scrapes the web directly. From a research perspective, this is the easiest mode of access and there are amply infrastructures supporting automated web scraping. Web data, if it gets preserved at all, is archived in the form of preprocessed data, not as raw web data initially collected. This research workflow doesn't take advantage of existing large-scale web scraping efforts, like ClueWeb or Common Crawl,[67] which provide a more stable research object for scholars interested in studying the web.[68]

[67] Common Crawl is a nonprofit organization like the Internet Archive whose mission is to crawl the web and make the data available for free on the web for researchers. Common Crawl continuously scrapes the entire web every month and Amazon hosts the data on their cloud platform. http://commoncrawl.org/big-picture/

[68] Common Crawl and ClueWeb are datasets *not* archives, they are more interested in serving scientists and quantitative researchers than qualitative social scientists and the humanities. The access model is oriented towards quantitative analysis using big-data tools. It is impossible to access selections of the ClueWeb collection, you have to work with the entire collection as a whole.

Data collection should go directly into a web archive. Then researchers should interact with the archive, performing data extraction, reshaping, or other processes associated with whatever scalar device they employ. Ideally a web archive should sit *in between* the researcher and the Open Web, provide multiple modes of access, and support a variety of qualitative and quantitative research methods.

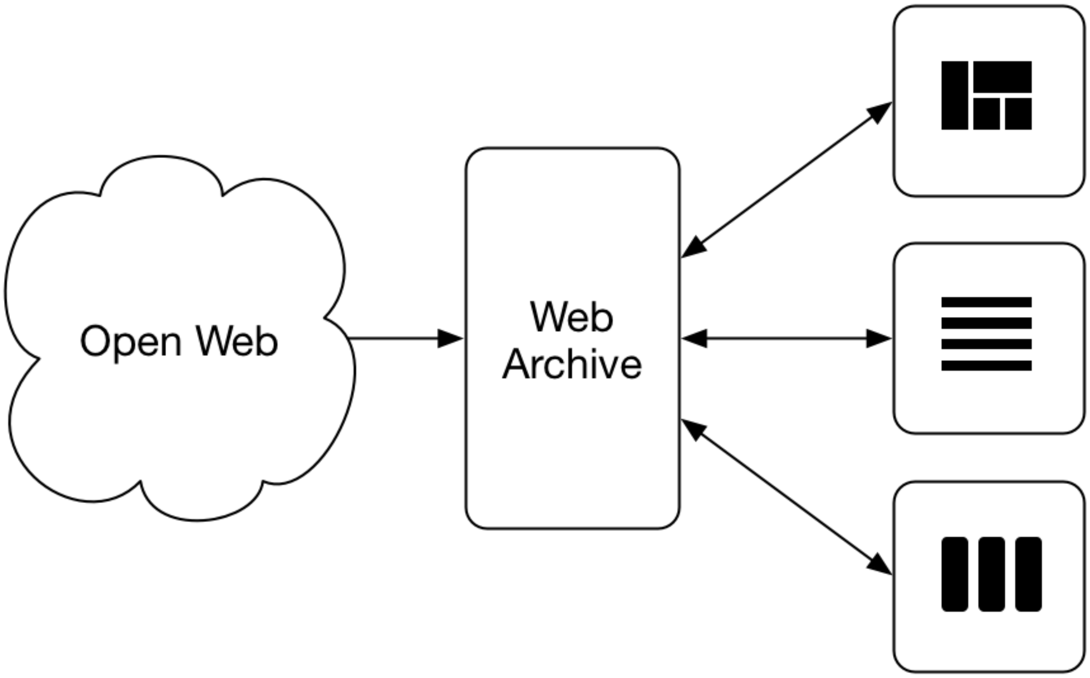


Figure (16). A better model for web research

Web scraping, crawling, and archiving for mixed methods social science have a very particular set of needs that are not well supported in the ecosystem of web preservation. The *selection of* and *access to* the collection need to support quantitative and quantitative research, but not mixed methods research that weaves between the two.[69] The modes of *selection* should be specific, both in term of network topology and temporality. Researchers should be able to specify

While this is good for big-data research, it doesn't work well for researchers like myself who are only interested in studying a tiny subset of the entire web.

[69] Paying $380 for a set of physical hard drives is not a good mode of access to an archive.

exactly which sites *and* time frame they want to preserve. The modes of *access* should permit the computational extraction of features, i.e., word counts or links, as well as browsing the collection in its original representation (or as close as possible).

The current state of web archiving infrastructure is idiosyncratic and doesn't effectively support researchers in the computational sciences. Lin, Kraus, & Punzelan (2014) have argued the current state of the art in web-archiving technology has not kept pace with the current generation of big data technologies such as the Hadoop suite of tools. They have developed a tool, *warcbase* that stitches together web-archives with big-data technologies making it easier for computational researchers to work with the web-archive created by institutions like the Internet Archive or Common Crawl. *Warcbase* and similar technologies will allow for richer modes of access to web-archives beyond basic browsing or per-URL retrieval. Full-text search, network analysis, maybe even topic modeling at *internet scale* are possible when large archives like the Common Crawl are combined with big-data software like Hadoop, and, most importantly, the cyberinfrastructure and people capable of supporting it.

Herbert Van de Sompel's *Memento* project is middleware for accessing "pockets of persistence" by leveraging the technical open standards of the Open Web as a means of accessing older versions of web resources *at the infrastructural level.* Memento web proposes using HTTP Request Headers as a means of accessing archived versions of resources identified by URIs. This means existing web crawling and web scraping systems can still be used while providing loosely coupled and transparent access to existing archives like the Internet Archive, Archive-It, or other web archive partners.[70]

[70] Memento Guide—Introduction to Memento http://mementoweb.org/guide/quick-intro/

There are few researchers using computational methods to understand digital culture "preserved" in web archives.[71] Ian Milligan, a digital historian at the University of Waterloo, is one of the few researchers who is using the techniques and methods of digital humanities to explore the near history of communities, like Canadian Political Parties, on the web. Most notably, these experiments have been published on his blog, not in books or journals.[72] Milligan, along with William Turkel and Mary Beth Start, have been developing a toolkit, HistoryCrawler, to assist historians in computationally analyzing web archives.[73]

IPython Notebooks as Documentary Traces of Research Practice

This chapter, which is an account of my methods and decisions, was derived from documentary traces preserved in an extremely important piece of research infrastructure, the IPython Notebook.[74] The IPython Notebook is a digital document format and computational environment for *writing*, *executing*, and *preserving* code. The Notebook mixes narrative text, computer code, data, and visualizations into a single document that can be executed, shared, and preserved. This has transformative implications for digital research.

Rather than working on the command line or in the Python interpreter, the IPython Notebook saves the command and result of each cycle of interaction with the computational engine, in my case Python. The practice of data cleaning, preparation, and even the execution of the model has been saved in a

[71] Computer scientists use and analyze web archives for in the course of designing computational models and machine learning algorithms. Web archives are like stones being used to sharpen a blade. The computer scientists want a sharper blade, whereas I am more interested in the stones.

[72] Milligan's post, "Using Modularity to Find and Explore Web Archived Communities" uses social network analysis to try and find clusters of relations around political parties. http://ianmilligan.ca/2015/02/03/using-modularity-to-explore-web-archives/

[73] SSHRC Research proposal for History Crawler http://ianmilligan.ca/the-next-project/sshrc-proposal/

[74] For an introduction to the IPython notebook, visit http://iPython.org/notebook

standardized format, the IPython Notebook. These notebooks allow me to review and even reproduce my work down to the most granular detail. My research practice is, with little effort, completely documented. Now, making sense of these notebooks can be difficult and the execution, for technical reasons, cannot always be re-computed because of “side effects.”

IPython Notebooks are a structured document format[75] combined with a browser based interactive computing environment. Notebooks are composed of a series of cells of different types, Python code, computed output, plain text, and markdown/html. The cells can be arranged such that code, graphs, and prose are mixed together as dynamic, interactive scientific notebook. IPython Notebooks are particularly powerful because they are self-contained and portable documents. This means it is easy to document, share, and possibly preserve computational practice.

IPython Notebooks are the documentary traces of my work compressing voluminous blog posts into something more humanly comprehensible. These traces are inscriptions of the deployment of a scalar device and the work surrounding that deployment. These traces capture a level of granularity about my research practice at a much finer grain than the typically methodology section. There is too much detail, which is why through a form of trace ethnography I have tried to synthesize the data, code, description, and visualization fixed in the notebooks into the narrative above.

Topic modeling, as with any computational technique, requires an amount of data cleaning and preparation that is typically not fully articulated in methodology sections. IPython notebooks provide a documentary infrastructure for preserving the implicit or ineffable (in formal publication) threads of work. The act of publishing code, alongside findings and data, is yet another frontier of scholarly publishing. However this will be a challenge because raw code, like raw

[75] The IPython Notebook is stored as a standarized JSON file. While a schema exists for the format, the standard has not yet been submitted to any formal standards bodies. https://github.com/iPython/iPython/ipython/blob/master/IPython/nbformat/v4/nbformat.v4.schema.json

data, is messy and incomprehensible to outsiders. Trace ethnography was a useful reflexive and interpretive frame for *making sense* of the code I produced in the course of this research. This chapter, which synthesizes the code and practice, is the result of that analysis.

Chapter Five

Themes in Digital Humanities Blogs

The purpose of this chapter is to answer the question: *What themes are digital humanities scholars talking about on their blogs?* Answering this question using qualitative methods, such as content or discourse analysis, would be impossible for a single analyst at scale. To solve this “big-ish data” problem, I trained a 100-topic model on a subset of my corpus of 39,976 blog posts using the MALLET toolkit (Mccallum 2002). The topic model is a map of the corpus for a single scholar to study. Using quantitative and qualitative methods I interpret the model’s topic distributions into human meaningful *themes*.

Matthew Jocker’s word *themes* describes the discourses surfaced by a topic model. Themes are the meaningful *subject matter* of the blog posts. The word *themes* instead of *topics* to prevent a conflation between the “topics” produced by Latent Dirichlet Allocation statistical analysis and the “topics” resulting from a human reading of the texts. In this chapter, *topic* means the word distributions produced by the topic model and *themes* means the subject matter of blogs as interpreted by a human reading of the topic model and source documents.

Transforming topics into themes is an act of *interpretation* that begins with topics, a probability density of words, and makes them meaningful. To articulate what Digital humanities scholars are writing about on their blogs is to understand a set of topics as a set of themes. Topic distributions are not meaningful in and of themselves. They are not innately endowed with meaning nor are they the final arbiters of interpretation. The model exists on one side of a *meaning gap* in the ontological state *as data*. To interpret is to hop this gap and *interpret* the *meaning* of the data into narrative.

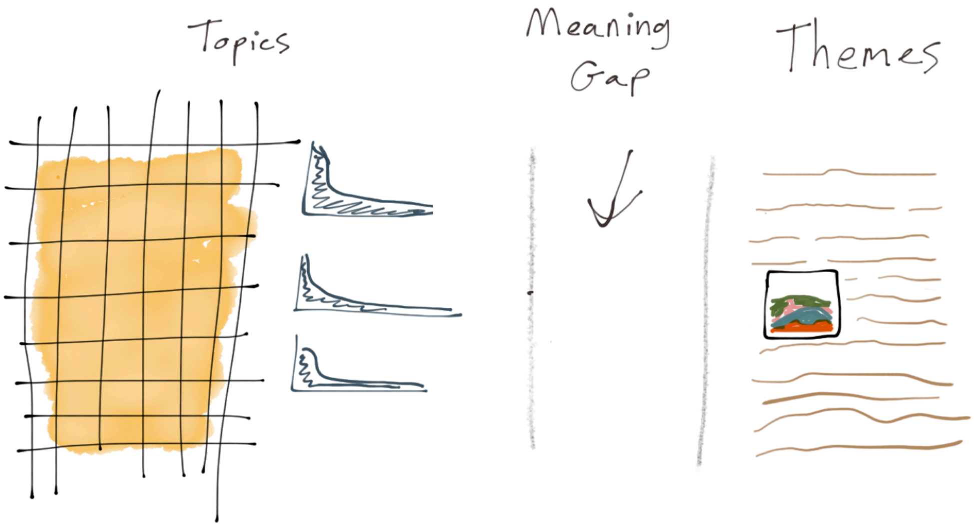
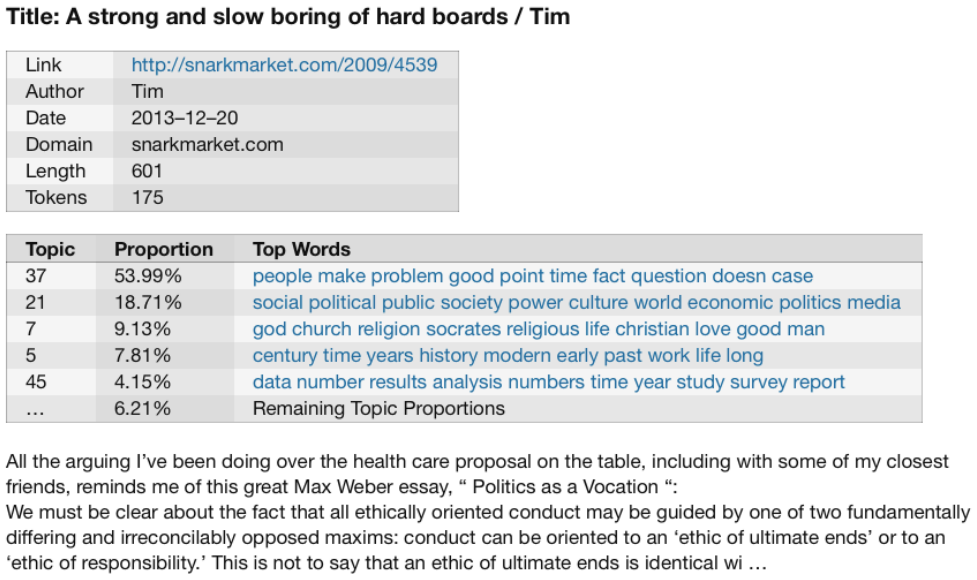


Figure (17). The meaning gap between topics and themes

Looking at figure (17) above, on the left side of the meaning gap we have topics, as data, on the right side we have themes, as narrative. In-between is the process of interpretation. Overcoming this gap means bringing the various pieces of information, the topic model and blog content, together to enable enriched reading of the traces.

Hopping over the gap was accomplished by creating two documents assembling the traces in such a way as to create a synthetic representation of the data. The first document annotates the blog posts with the model's document topic proportions. For each topic, the topic 25 documents were assembled linked together via shared topic connections. This resulted in an inter-linked HTML document used in the trace ethnographic analysis of the data. Figure (18) below shows an example of a single blog entry annotated with document topic proportions. The second document shown in figure (19) shows the top 25 topic proportions for each of the 396 domains in the data.



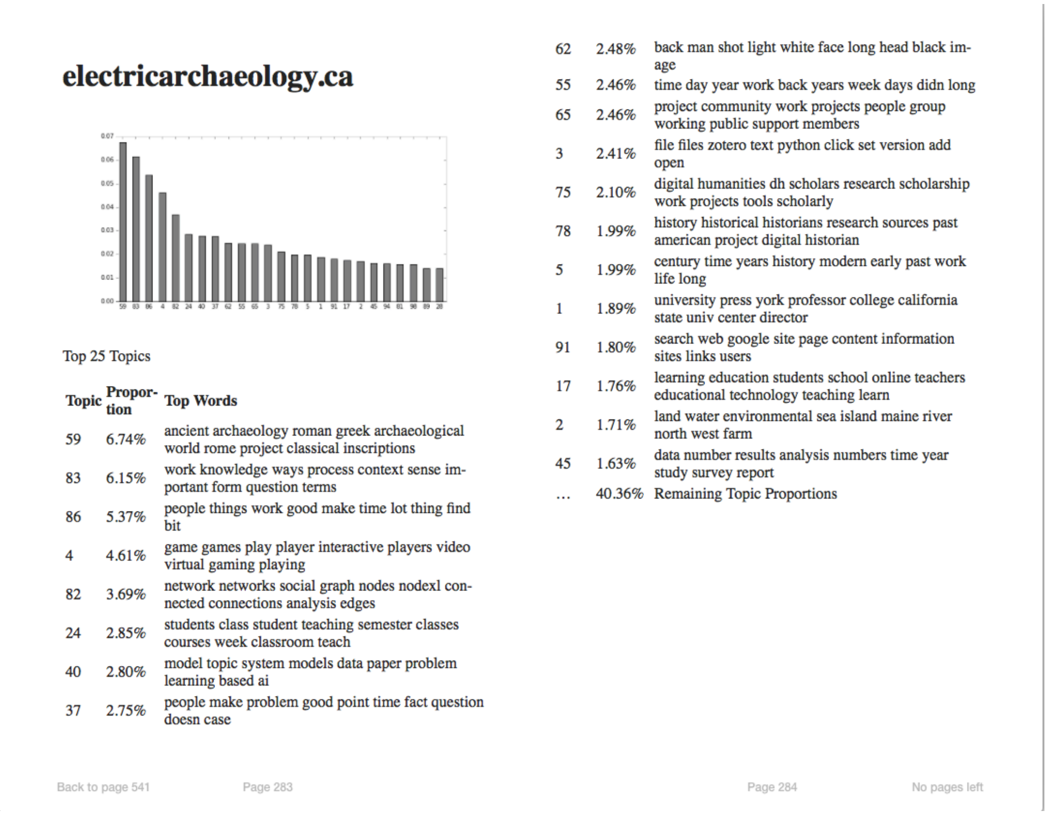
Title: A strong and slow boring of hard boards / Tim

Link	http://snarkmarket.com/2009/4539
Author	Tim
Date	2013–12–20
Domain	snarkmarket.com
Length	601
Tokens	175

Topic	Proportion	Top Words
37	53.99%	people make problem good point time fact question doesn case
21	18.71%	social political public society power culture world economic politics media
7	9.13%	god church religion socrates religious life christian love good man
5	7.81%	century time years history modern early past work life long
45	4.15%	data number results analysis numbers time year study survey report
…	6.21%	Remaining Topic Proportions

All the arguing I've been doing over the health care proposal on the table, including with some of my closest friends, reminds me of this great Max Weber essay, " Politics as a Vocation ":
We must be clear about the fact that all ethically oriented conduct may be guided by one of two fundamentally differing and irreconcilably opposed maxims: conduct can be oriented to an 'ethic of ultimate ends' or to an 'ethic of responsibility.' This is not to say that an ethic of ultimate ends is identical wi …

Figure (18). A blog post annotated with the document topic proportions.



electricarchaeology.ca

Top 25 Topics

Topic	Proportion	Top Words
59	6.74%	ancient archaeology roman greek archaeological world rome project classical inscriptions
83	6.15%	work knowledge ways process context sense important form question terms
86	5.37%	people things work good make time lot thing find bit
4	4.61%	game games play player interactive players video virtual gaming playing
82	3.69%	network networks social graph nodes nodexl connected connections analysis edges
24	2.85%	students class student teaching semester classes courses week classroom teach
40	2.80%	model topic system models data paper problem learning based ai
37	2.75%	people make problem good point time fact question doesn case
62	2.48%	back man shot light white face long head black image
55	2.46%	time day year work back years week days didn long
65	2.46%	project community work projects people group working public support members
3	2.41%	file files zotero text python click set version add open
75	2.10%	digital humanities dh scholars research scholarship work projects tools scholarly
78	1.99%	history historical historians research sources past american project digital historian
5	1.99%	century time years history modern early past work life long
1	1.89%	university press york professor college california state univ center director
91	1.80%	search web google site page content information sites links users
17	1.76%	learning education students school online teachers educational technology teaching learn
2	1.71%	land water environmental sea island maine river north west farm
45	1.63%	data number results analysis numbers time year study survey report
…	40.36%	Remaining Topic Proportions

Back to page 541 Page 283 Page 284 No pages left

Figure (19). The top 25 topic proportions for a domain.

From a methodological perspective, this is a form of content analysis but where the content is composed of *traces* produced by human (bloggers) and non-human (MALLET) processes. Trace ethnography provides the theoretical grounding for the qualitative analysis of the topic model and the original blog content. In practice, this involves an enriched reading that follows the linkages between topic keywords, document proportions, and the original blog posts.

This chapter has three sections. First, a thick description of the topic model and the shape of the data it created in order to situate what we can *know* from the analysis. Second, a quantitative analysis measuring topic diversity and visualizing the model with hierarchical clustering to provide two high-level perspectives of the corpus. Third, a richer interpretation through four categories of discourse, *quasi-academic*, *meta-academic*, *para-academic*, and *extra-academic*, as analytical categories that enrich our understanding of informal scholarly communication.

An Ethnography of the MALLET Model

> All models are wrong, but some are useful. —George E. P. Box

Training a topic model with MALLET produces several types of data. The most useful are the *topics (*distributions over words) and the *document topic mixtures,* (topic distributions over documents.) The topic model has a hundred topic distributions over 154,363 word tokens and a hundred document topic mixtures over 39,976 documents. This, combined with the data from the source documents, means I still a significant amount of data to be analyzed and interpreted. These data are *traces* generated by MALLETs generative process and the raw material of my trace ethnography.

The shape and contours of a topic model are dependent upon the parameters specified when training the model. The three main parameters when training MALLET topic models are hyper-parameter optimization, document quantity and length, and most importantly, the number of topics. While MALLET does have a few other buttons and levers, these have the greatest impact upon the quality of topics identified. Hyper-parameter optimization is a feature of the MALLET toolkit that will periodically readjust hyper-parameters when estimating the model. In terms of output, this means computed topics have different “sizes,” that is, its breadth and depth throughout the document topic distributions. The size of a topic can be informative when analyzing and interpreting the model because it can provide a quick indicator of topic quality. Every topic is assigned an alpha value that measures the topic’s distribution across the corpus. Topics with a small alpha appear in a small number of documents but with a high proportion. Topics with a larger alpha value are spread across a larger number of documents, typically with a smaller proportion. Research using topic models for the interpretation of texts provide little guidance on best way to specify this parameter beyond trial and error. I used 20 because, like Pagano and Maalej, it “leads to the most meaningful results” (2011).

The corpus itself is a collection of parameters to be tuned when training a topic model. The dataset originally contained 106,804 total blog posts, however only 39,976 of these to MALLET. I filtered out 66,828 documents because they were too short (fewer than a hundred words). Tang et al. (2014) found the performance of LDA models decreases with short documents. Also, blog posts with fewer than a hundred words rarely contain meaningful content. I trained the topic model on 39,976 documents with a total of 154,363 unique word tokens.

The most important parameter when training topic models is selecting the number of topics. While topic modeling is an unsupervised machine learning algorithm, it isn’t magic. As with other clustering algorithms, you need to specify the number of topics, or clusters, before training the modeling. Knowing or finding the “correct” number of topics is an area of ongoing research. Researchers have proposed techniques for finding the “right” number of topics (Arun et al. 2010), but these techniques have been found in other research to hinder the human interpretability of the topics (Chang et al. 2009). I trained a hundred-topic model. I tried a smaller, sixty-four-topic model, but found the meaningful themes to be, for lack of a better term, cramped. The larger hundred-topic model produced similar themes, but they were better separated out into individual topics making them easier analyze and interpret. A model with a hundred topics served my purpose, it was small enough allow me to give each topic a closer reading, but big enough where themes were nicely aligned with topics.

Using topic modeling in domains like machine learning or information retrieval is different than in the interpretive social sciences or humanities. MALLET's parameters must be tuned with respect to research goals. In this case, the goal was to identify a set of themes within a large discursive space. If the goal was to build a faceted search engine, I might have tuned the model differently by specifying more topics and not filtering out documents with less than a hundred tokens; for what good is a search engine that doesn't index all documents?

Using Alphas to Evaluate Topics

MALLET's hyper-parameter optimization means topics can be different "sizes" This adds another dimension to consider when interpreting the model. Without hyper-parameters, the topics would all be uniform in size, which may or may not make sense given the kinds of documents in the corpus. Given that blogs have few constraints on content and style, it is reasonable to expect a large diversity of themes. Alphas can then provide additional information about the topics was useful when considering their quality.

A small alpha value implicated topics with extremely narrow themes. At the other end of the spectrum, topics with large alphas had very broad and general themes. Some large topics appeared to contain more than one theme (a sign that perhaps a larger number of topics could have been specified), others had semantics at an extremely general and high level that upon inspecting the top-ranked documents didn't appear to come together in any readily apparent way. Such topics might point to underlying structural features or patterns of the English language the algorithm can see, but don't have meaning to a human reader.

Small Topics / Foreign Language Topics

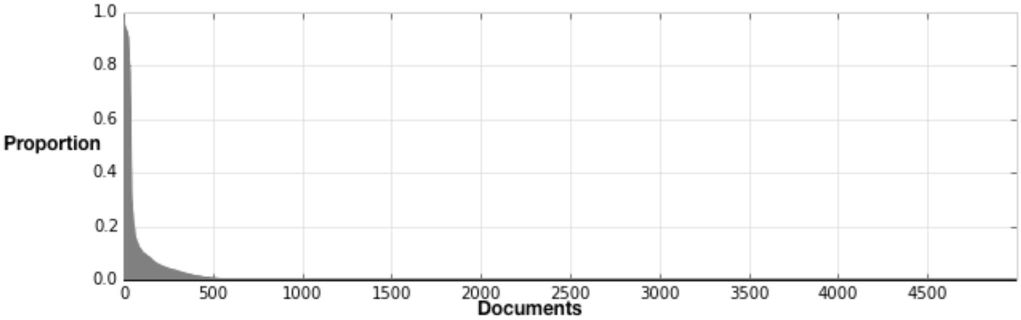


Figure (20). Topic 85 has a high proportion for a small number of documents.

Topic 85, *de da em para software na um os como uma*, is the smallest topic in the model with an alpha value of 0.001193. Figure (20) shows the shape of topic's distribution over the top five thousand documents you can see there is a small number of documents where this topic is *highly* expressed, over 90 percent. There are somewhere between fifty and one hundred documents that are dominated by this topic. This collection is followed by a sharp drop to somewhere around ten percent creating a small tail that extends for the top five hundred documents. After five hundred, the percentage approaches zero. This means that topic expression is limited to a very small collection of documents, and for that collection is is *extremely* highly expressed. Closer examination of the topic shows most of the top documents are in Spanish. This makes sense; documents in other languages are structurally distinct from the rest of the corpus (which is in English).

Topics with an alpha value of less than 0.01 do not surface readily apparent themes. Small topics have a tendency to focus upon shared words of a single blog. The alpha value in this case is a good indicator of how much interpretive weight to give these topics.

Large Topics

Topics with high alpha values have a readily identifiable shape to their top document proportions graph. Figure (21) displays the top document proportions for topic 86, *people things work good make time lot thing find bit*, which has an alpha value of 0.2432. This is the largest topic in the corpus and thematically very general and hard to assign a pithy label.

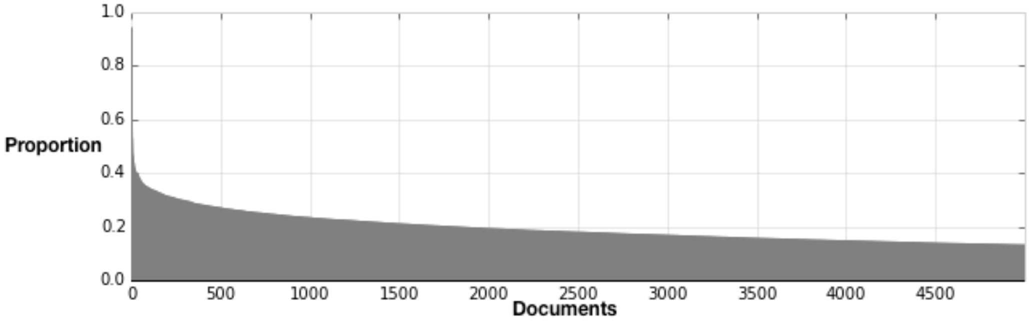


Figure (21). Top document proportions for topic 86, the largest topic.

Figure (21) shows how the weight of this distribution is spread across five thousand documents in the corpus. There are a few documents at the head of this distribution, but for the most part this topic is spread out across a very large number of documents with a modest 10 to 20 percent expression.

Very large topics illustrate the limitations of human interpretation of unsupervised machine learning. Large topics are so big they features that bind them are beyond meaningful comprehension. They reveal a kind of structural coherence, but no obvious semantic coherence to how the individual posts have been pulled together. Topic modeling has a way of detecting patterns that aren't readily interpretable. These topics exist because LDA detected a pattern in the frequent co-occurrence of these words, but when trying to read into the topic via the top words and top documents, there are no obvious themes. The words clustered by these topics might be performing as "stoppish" words; that is, words that *can and do* have meaning independent of their context, but in this particular model they don't come together meaningfully.

The vast majority topics in the model are also the most amendable to interpretation. These topics have midrange alpha values and lend themselves to readily identifiable themes, a mix across domains, and mixed with other topics within documents. These midrange topics qualitatively approximate a one-to-one ratio between machine generated topics and human interpretable themes. Midrange topics are the most interesting because they render visible insights that might otherwise be invisible to a close reading.

The large topics are similar to what Lisa Rhody, in her discussion about using topic models to analyze poetry, called "semantically *opaque* topics." Topics whose top words don't make sense and upon closer inspection are difficult to find apparent coherence. These topics are interesting not because of the insights they impart upon the interpreter, but by the questions they evoke.

> Opaque topics ... in models that have mixed results prompt the kinds of questions we are looking for as humanists. What this ... shows is that topic modeling as a methodology, particularly in the case of highly-figurative language texts like poetry, can help us to get to new questions and discoveries—not because topic modeling works perfectly, but because poetry causes it to fail in ways that are potentially productive for literary scholars (Rhody 2013).

Investigation of the ways in which topic models "break" is a form of *infrastructural inversion,* whereby the normally hidden innards of an infrastructure, or in this case unsupervised learning algorithm, are broken open and exposed to us as users. These unexpected relations are the promise of distant reading, to lead to knowledge and insight previously impossible. That said, how do we distinguish between a productive break and an artifact of the algorithm that leads to a dead end?

Document Topic Distributions

The topic document mixtures are the other collection of useful data produced by MALLET. These data take the form of probability distributions of a topic's expression within each document (derived from the number of words from each topic present in the document). The document/topic matrix contains these probabilities for each of the 100 topics for each of the 39,976 documents.

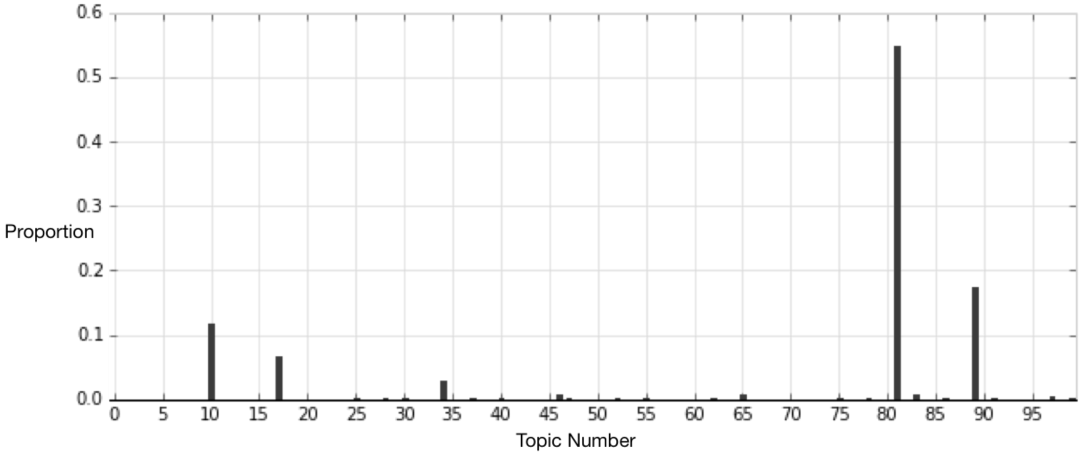


Figure (22). Unranked chart of topic mixture for a single document.

Figure (22) shows shows the topic mixture for a blog post about the Badges for Lifelong Learning competition.[76] There are primarily four topics with any significance and most of that belonging to topic 81, *award competition year prize badges festival awards badge winners winning*. This makes sense given the nature and content of the document. In the generative process, topic 81 would have the highest likelihood of being selected, which in turns means words like *award*, *competition*, and *year* would be more likely to be in the document. However, one of the powerful features of LDA is the ability to model documents with more than one topic. Because the document's topic mixture is a distribution, we also see a probability of 17.53 percent for topic 89, *design web user users content software tools system systems services*. Based on these two topics we can infer a theme in this document about competition and design which isn't far off from the post, which, upon close reading, is on the Digital Media and Learning Competitions for which is about building and designing software tools for education.

[76] The post is titled "Friendly reminder: Important January deadlines for the Badges for Lifelong Learning Competition." To see the post visit: http://dmlcompetition.net/Blog/2012/01/friendly-reminder-important-january-deadlines-for-the-badges-for-lifelong-learning-competition/

Topics, Top Words, & Distributions Over Words

Topics are probability distributions over words and are the main feature used when interpreting topic models. Each topic is a uniquely configured “bag of words” from which the generative process selects a word when composing a document. They can provide an almost spooky level of insight into what a document or corpus is about. In considering the output of a topic model, researcher’s attention gravitates to the top ten or twenty words with the highest rank per topic. A topic’s *keywords* (top-ranked words) are the words with the highest probability in the distribution.

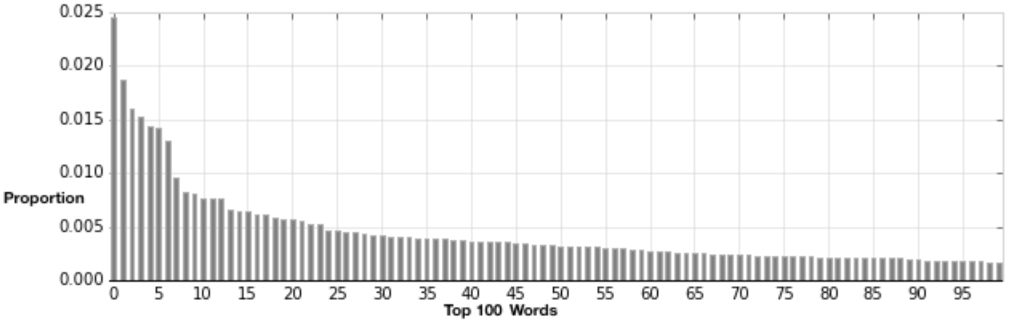


Figure (23). The shape of the word distribution for topic 86. The probabilities of the top-ranked words are on the left side of the chart.

The bar chart in Figure (23) shows the probability of the top one hundred words in topic 86, whose top-ranked words are *people things work good make time lot thing find bit*. The chart shows the probability of the top-ranked words starting from the left. We can see the word “people” has a probability of 2.45 percent. The next-highest-ranked word, “things,” has a probability of 1.86 percent. The distribution extends over the 154,363 words in the model. Each topic has differently shaped distributions; different words are ranked highly with different proportions.

The top words are typically used to represent the topic in analytical narratives, such as this one, describing the model. While a topic distribution contains every word token, only a small number have any probabilistic weight worth noting.

Examining the top N words of a topic provides an interpretable and meaningful way to for a human analyst to understand what a topic is about. However, as Ben Schmidt (2013) pointed out, the long tail of this word distribution contains valuable information; they shouldn't be interpreted divorced from the document context within which they originated.

Thematic Diversity

Given that blogs and the Open Web afford open publishing to what extent do scholars take advantage of the ability to write whatever they want? The question that motivates the section is: *How thematically diverse are digital humanities blogs?* Unlike formal scholarly communication, blogs do not impose structural constraints on the content scholars can publish. They can write about anything and hit the "publish" button on their blogging platform. This means scholars are free to write about teaching one day, philosophical quandaries another, and finally their favorite recipe on a third. However, the affordances of the Open Web only remove *technical* constraints on thematic content, to what extent do digital humanities bloggers exhibit self or social restraint?

I calculated the total cumulative topic proportions for each blog by aggregating the weighted topic proportions of all the documents in each blog. These computations provided a probability distribution of topics per blog, which the breadth of topics a scholar writes in total, not just within individual documents.

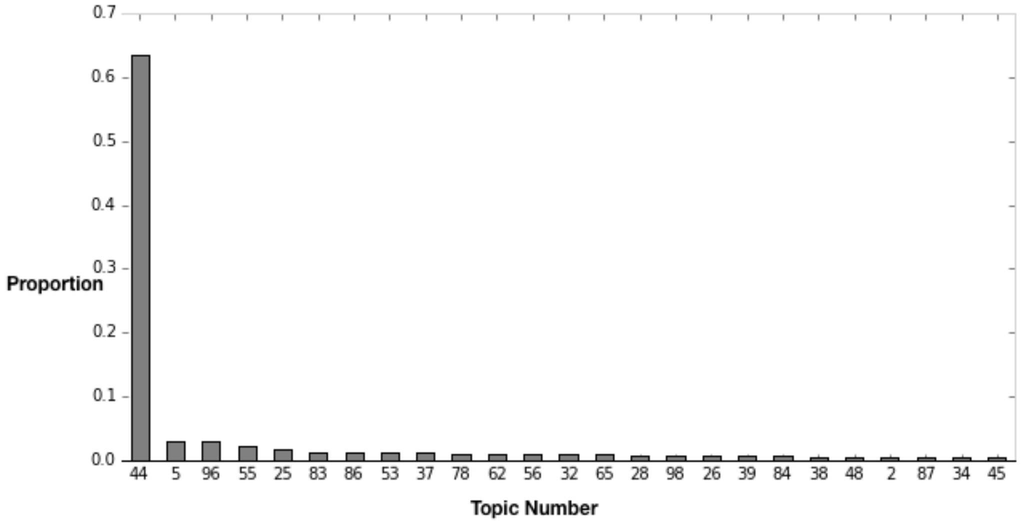


Fig (24). Top twenty-five topics for the blog, vinson.hagleyblogs.org. Most of the probability lives in the first topic indicating this blog has little thematic diversity.

Figure (24) shows the top-ranked topics from `vinson.hagleyblogs.org`, a blog managed by the Z. Taylor Vinson Transportation Collection at the Hagley Museum and Library. It is an institutional blog with a high proportion of topic 44, *car cars taylor collection vinson american sports ford hagley team*, a topic exclusively dedicated to words about cars and the Vinson collection. Cars are not common themes in the digital humanities, so this blog is relatively isolated, content-wise, from others in the corpus.

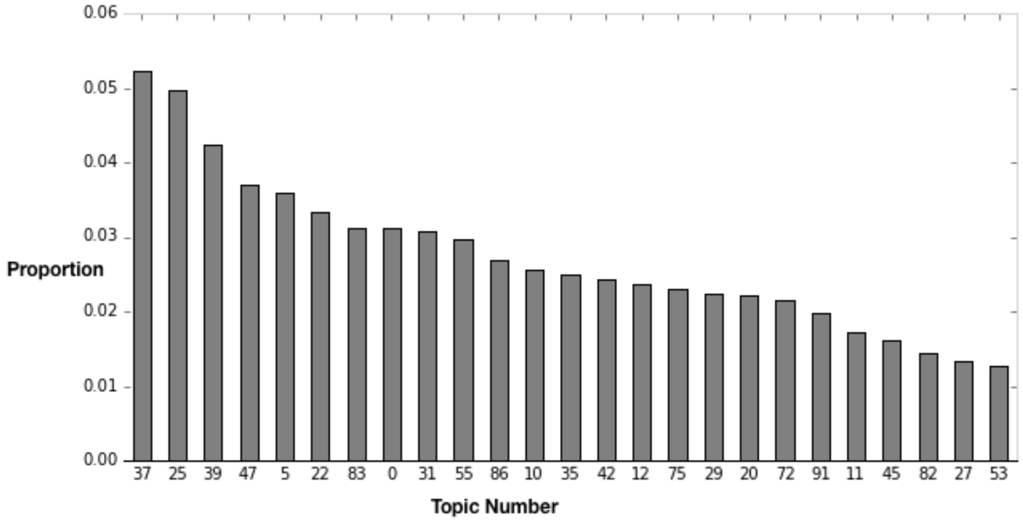


Figure (25). An example of a domain, amandafrench.net , with a very high diversity of topics

Figure (25) shows the top twenty-five topics for a very different blog, `amandafrench.net`, the blog of scholar Amanda French. Notice how the distribution is spread more evenly across the top 25 topics (and beyond) instead of accumulated at the head like the Vinson Collection. The two top topics are topic 37, *people make problem good point time fact question doesn case*, and topic 25, *library digital libraries collections preservation collection access information archives archive*. French, likes to write about a variety of themes and because she manages the blog free from the encumbrance of an institutional apparatus, she writes more freely. When asked about what she writes about on her blog, French responded:

> I keep it academic, although it's a fairly informal version of "academic." Bloggy. A bit journalistic, too. I often used to think that having a blog was a bit like having a newspaper column, certainly tonally (Amanda French interview).

Plotting topic proportions for each blogs provides a visual indicator of thematic diversity. Some blogs stay tightly "on topic," while others vary widely in their subject matter. This information leads to the question, what is the average level of thematic diversity on digital humanities blogs? While visual inspection works a single metric would show the bigger picture.

Shannon’s Entropy as a Measure of Diversity

Shannon’s Entropy, a quantitative measure of information and uncertainty within a message, indicates a blog’s thematic diversity. Each blog has a unique distribution of topic proportions. Some skew heavily towards one or two topics, while others are spread across a large number of topics. Entropy works as a measure of thematic diversity per blog because it characterizes the expected expression of particular topics from the distribution. Blogs whose probabilities skewed towards one or two topics have low entropy because there is more certainty that an “observation,” the topic of a blog post, will be drawn from a small set of topics. Blogs with very high entropy express a larger number of topics so the expectation of specific individual observations is less certain. High entropy blogs draw from a wider variety of topics.

Shannon’s entropy measures the amount of information in a message. Information in this context is the meaningless, mathematical concept developed by Claude Shannon as a means of quantifying the uncertainty or regularity of sequence of observations. For example, in the string “AAAAAA” there is very little entropy because every character is the same and so we expect the letter “A,” whereas the string “v-b49[hq5” has more entropy because we are less certain about each individual character in the string. We can think of the varying expression of topics in blogs like these certain or uncertain strings.

To compute the entropy per blog I used the `scipy.stats.entropy` function from the SciPy Python library.[77] I passed this function two parameters, first a sequence of 100 topic proportion values and second a logarithmic base value of 2. This function condenses the 100 value sequence into a single value from 0 to 6.64 (the value of log2(100)). Zero would indicate a blog with 100 percent of the probability belonging to a single topic, 6.64 would indicate a blog with a uniform

[77] The documentation for the function is available here: http://docs.scipy.org/doc/scipy-dev/reference/generated/scipy.stats.entropy.html and the source code for the function is available on GitHub here: https://github.com/scipy/scipy/blob/master/scipy/stats/_distn_infrastructure.py#L2350

distribution of 0.01 for each topics. After calculating the entropy for each of the blogs, I created a histogram to visualize the distribution of the entropy values and get a sense of where most of the blogs fell on the spectrum between 0, no diversity, and 6.64, maximum diversity.

Visualizing Topic Diversity

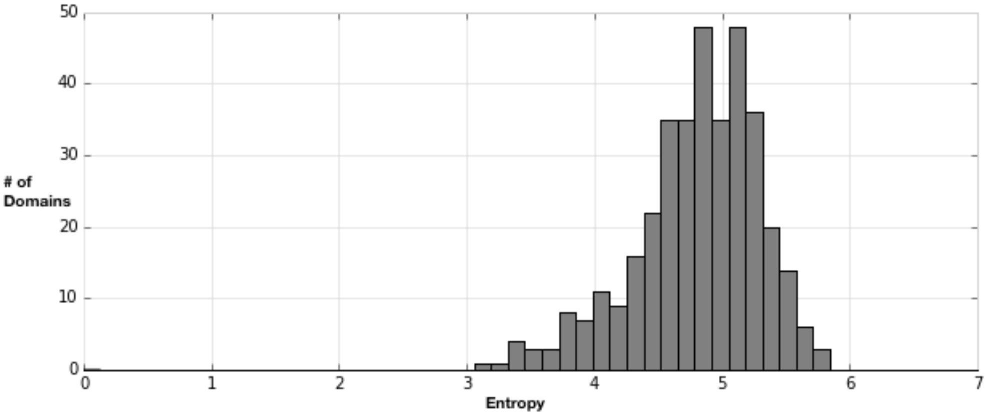


Figure (26). A 50 bin histogram of topic proportion entropy per domain.

The shape of the histogram in Figure (26) resembles a classic bell curve or normal distribution. There is a bit of a tail on the left side of the graph, indicating there are more blogs on the less-diverse side of the median. The interpretation remains the same, if we allow for entropy to be a suitable measure for a blog's topical diversity and zero represents a low diversity of topics and 6.64 represents uniform diversity across all topics, then we can see in Figure (26) that most blogs tend to shift towards thematic diversity in their content.

A quick statistical summary shows another perspective of thematic diversity of digital humanities blogs.

Table 2: Summary statistics of entropy measure

Count	365
Mean	4.799438
Standard Deviation	0.487676

Minimum	3.145530
Maximum	5.790262

These numbers and images give us a “distant” reading of the diversity of topics on digital humanities blogs, but only diving into the data to get a better feel for what a minimum of 3.15, maximum of 5.79, and a mean of 4.8 actually mean when it comes to the diversity of topics on the blogs. The numbers are meaningless and difficult to interpret without context to show what it means to have a high or low entropy score.

Diversity around the Min and Max Entropy

Looking at the shape of blogs at the edges of the spectrum frame the range of thematic diversity for the blogs in the middle. Remember, the blogs with a low entropy score are those that stay *on topic*, that is, most of the topic proportion is drawn from a small number of topics. The blog *dragonfly.hyoptheses.org* has the lowest entropy score of 3.15. The number alone isn’t very informative, so visualizing the shape of the distribution of top topics as well as the keywords of the top 5 topics provides some context.

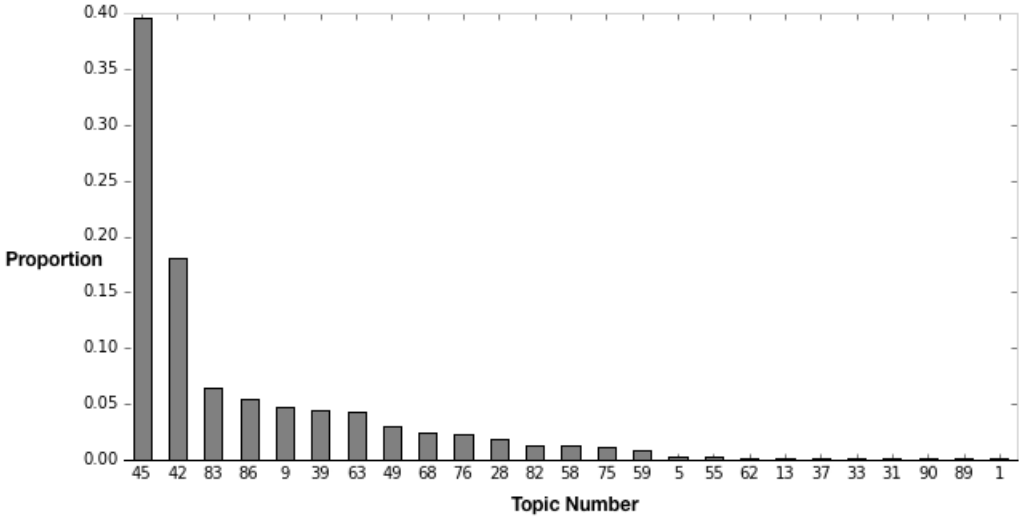


Figure (27). Top twenty-five topics for dragonfly.hyoptheses.org, the blog with the lowest entropy

The plot in Figure (27) above shows a blog almost exclusively focused on a couple topics with a tail that diminishes very quickly. This is characteristic of the blogs with a low entropy score. They are primarily about a single topic with a small expression of other topics periodically in the mix. Only ten topics in dragonfly.hyoptheses.org have a proportion above 2 percent.

To provide a bit more context, the example above, *dragonfly.hypotheses.org*, is a blog all about "computational approaches to literary text analysis." According to the about page:

> This is a personal research and news blog. Its aims are to engage discussion with colleagues interested in computational text analysis, to serve as an observatory of trends, issues and advances in computational text analysis, and to give visibility to the burgeoning and diverse field of computational text analysis.[78]

The keywords from the top five topics fit with the self-reported description of the blog.

Table 3: Top Topics for dragonfly.hypotheses.org

Topic	Proportion	Top Words
45	39.64 %	data number results analysis numbers time year study survey report
42	18.00 %	story stories fiction narrative characters read character writing world author
83	6.51 %	work knowledge ways process context sense important form question terms
86	5.44 %	people things work good make time lot thing find bit
9	4.72 %	data map visualization maps visual information image images color mapping

A bulk of the topic proportion is dedicated to topic 45, one of the topics I labeled as a *humanities computing*. The second highest topic, topic 42, which I labeled *stories*, is all about narrative, fiction, and literature. These themes indicate *Dragonfly.hypotheses.org* is about the computational analysis of fictional texts. It is a technical and academic blog focusing on methodological questions, but

[78] For more information see the blog's "About" pblog's about age. http://dragonfly.hypotheses.org/about

reflecting on methodological challenges as can be see in the third highest-ranked topic 83, a topic about *knowledge work*.

At the other end of the spectrum, with the highest entropy score is *electricarchaeology.ca*, the blog of Shawn Graham an archaeologist and professor of digital humanities and history. His blog, while primarily focused on archaeology, also covers many other topics to a similar degree. The distribution of cumulative topic proportions diminishes much more gradually than *dragonfly*.

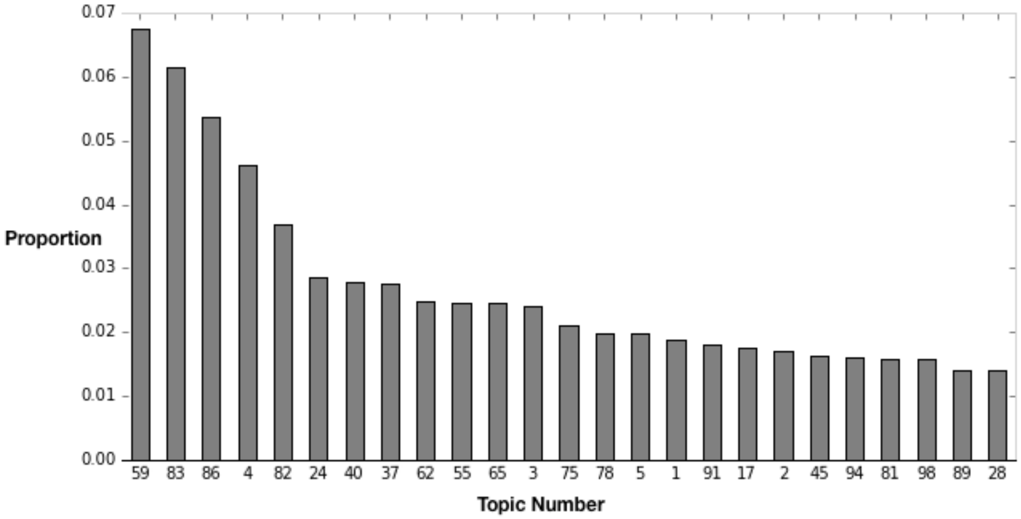


Figure (28). Top 25 topics for electricarchaeology.ca, the blog with the highest entropy score

Figure (28) above shows the top twenty-five topics for *electricarchaeology.ca*. The y-axis, the topic proportion, reveals the top topic 59, which I labeled *Ancient History*, is only expressed at 6.74 percent. Graham writes widely and broadly about a variety of topics including video games, network analysis, teaching, and the digital humanities more generally.

Table 4: Top Topics for electricarcheology.ca

Topic	Proportion	Top Words
59	6.74%	ancient archaeology roman greek archaeological world rome project classical inscriptions
83	6.15%	work knowledge ways process context sense important form question terms
86	5.37%	people things work good make time lot thing find bit
4	4.61%	game games play player interactive players video virtual gaming playing
82	3.69%	network networks social graph nodes nodexl connected connections analysis edges
24	2.85%	students class student teaching semester classes courses week classroom teach
40	2.80%	model topic system models data paper problem learning based ai
37	2.75%	people make problem good point time fact question doesn case
62	2.48%	back man shot light white face long head black image
55	2.46%	time day year work back years week days didn long
65	2.46%	project community work projects people group working public support members
3	2.41%	file files zotero text python click set version add open
75	2.10%	digital humanities dh scholars research scholarship work projects tools scholarly
78	1.99%	history historical historians research sources past american project digital historian
5	1.99%	century time years history modern early past work life long
1	1.89%	university press york professor college california state univ center director
91	1.80%	search web google site page content information sites links users
17	1.76%	learning education students school online teachers educational technology teaching learn
2	1.71%	land water environmental sea island maine river north west farm
45	1.63%	data number results analysis numbers time year study survey report
...	40.36%	Remaining Topic Proportions

The table above shows how probability is distributed in small portions across the topic space. Because the likelihood of any individual topic is so small, there is less certainty about what topics Graham will blog. This results in a high entropy score of 5.79.

Electricarchaeology.ca and *dragonfly.hypotheses.org* represent the upper and lower bounds of topical diversity in my data set. As evidenced by the curve in Figure (26)'s histogram, the diversity of topical discourse on most DH blogs exists at a more moderate level. Looking at blogs whose entropy is closer to the mean (4.79) gives a better idea of the extent and shape of thematic diversity.

Diversity around the Mean Entropy

Looking at the blogs with entropy values nearer the mean shows the extent DH bloggers write about different themes. Figure (29) below shows *nines.org*, a blog with a cumulative topic proportion entropy value of 4.77, which is close to the mean value of 4.80.

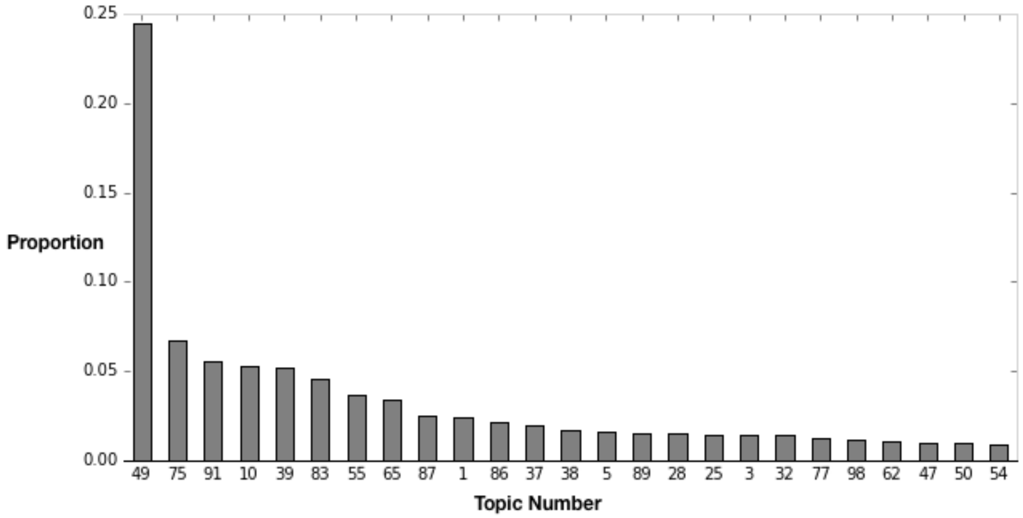


Figure (29). Top 25 topics for nines.org, which shows the average distribution of topic diversity

NINES or *Networked Infrastructure for Nineteenth-Century Electronic Scholarship* is a well-established digital humanities project founded at the

University of Virginia. The main focus of the project is to provide a social and technical infrastructure for research, peer review, and publication of digital scholarship about nineteenth-century British and American Literature.[79]

The NINES blog acts as institutional news feed announcing projects, highlighting NINES research, and engaging with the community. Looking at the top five topics for the NINES new blog yields few surprises:

Table 5: Top Topics for nines.org

Topic	Proportion	Top Words
49	24.42%	text project texts archive edition editions images blake transcription work
75	6.69%	digital humanities dh scholars research scholarship work projects tools scholarly
91	5.52%	search web google site page content information sites links users
10	5.24%	research university information program experience digital library position work faculty
39	5.18%	london century great john life james book english man author

The top topic for NINES, by a large margin, is topic 49, a topic exhibiting a strong humanities computing theme with words like *text*, *transcription*, *TEI*, *database*, *project*, etc. The focus is clearly on the digitization of historical texts, but also images and illustrations. NINES shares this topic with the *blakearchive.wordpress.com* and *textcreationpartnership.org* which are both institutional blogs about digitization and computational analysis of texts.

The NINES blog does include a bit of space for other topics, particularly topic 75, which I labeled as the "What is DH?" It would seem the NINES blog does not shy away from participating in the reflexive conversations about digital humanities, although it is certainly not a primary focus of the blog as it is expressed as a distant second place.

[79] For more information see the "About" page. http://www.nines.org/about/

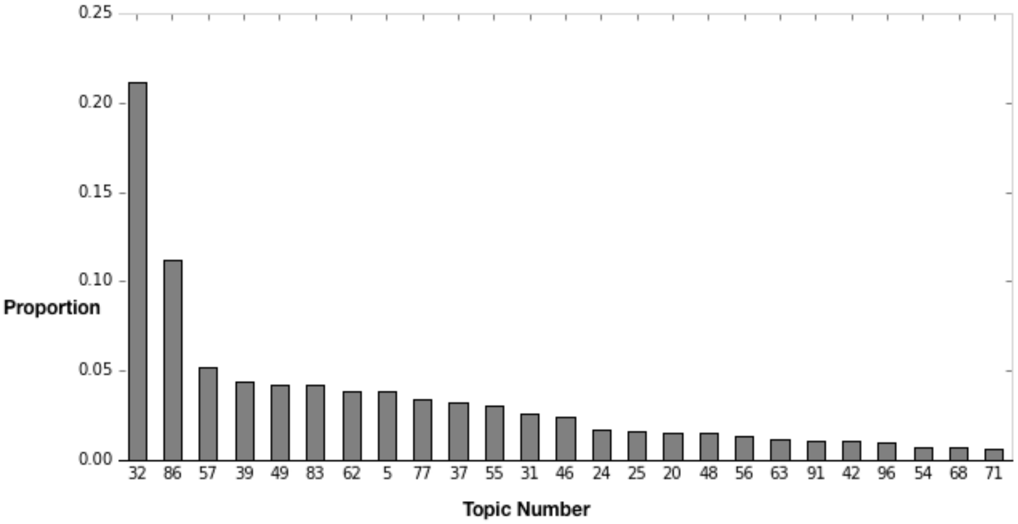


Figure (30). Top twenty-five topics for wynkendeworde.blogspot.com, which also shows the average distribution of topic diversity

Figure (30) shows another blog with entropy near the mean, *wynkendeworde.blogspot.com*, shows a similar pattern of topic proportions. Most of the proportion lies in one topic, with along tail. This is the older blog of Sarah Werner, which she described as:

> This blog shares thoughts specifically about books and early modern culture as well as speculations more generally about the history of books, reading, and printing.[80]

Looking at the top words of the top topics fits with this self-reported description of what the blog is about:

Table 6: Top Topic proportions for wynkeneworde.blogspot.com

Topic	Proportion	Top Words
32	21.12 %	book books reading read text print page readers paper writing
86	11.19 %	people things work good make time lot thing find bit
57	5.16 %	ms manuscripts library manuscript century british royal medieval england st

[80] For more information see the blog's "About" blog's aboutpage. http://wynkendeworde.blogspot.com/p/about-blog.html

39	4.40 %	london century great john life james book english man author
49	4.25 %	text project texts archive edition editions images blake transcription work

Werner's blog focuses mainly on themes about books and reading, but a there is still a fair bit of topic 86, a large general topic that, like topic 83, is about themes around *knowledge work*.

Digital humanities bloggers generally appear write around a single topic with a long tail of other topics. For the individual blogger who manage their own domains, despite the lack of technical or institutional forces determining what they can write about, scholars have a tendency to stay "on topic." This leads to the question, *what are they writing about?*

Measuring and visualizing entropy tells us the breadth and extent of diversity in the discursive space, but it doesn't indicate exactly *what* digital humanists are blogging about. The next sections of this chapter focus upon the subject matter of the topics by visualizing the topic model as a whole and then interpreting the model through four qualitative categories. First, is high-level distant reading of the entire topic space. Using hierarchical clustering a twenty thousand foot view, a truly distant reading of blogs mediated through several layers of quantitative analysis. The dendrogram, labeled with topic keywords, is an image of thin synthetic traces. These thin traces are then thickened through qualitative analysis that labels the clusters of topics based upon their keywords. This representation of is a data-derived *map* of the discursive space.

Visualizing Topic Clusters

To visualize the entire landscape of topics, I used a technique called hierarchical clustering to generate a dendrogram that groups similar topics together on the same branches of a tree-like structure. Appendix B contains the full image.

Hierarchical clustering requires a measure of pairwise distance between each of the topic distributions. Following Mimno (2012), I computed the Jenson-

Shannon distance, which provides a quantitative measure of similarity between each pair of topics. For example, topic 4 (*game games play player interactive players video virtual gaming playing*) and topic 74(*games game history play twitter university digital past http online*) have a distance of 0.99265486. Both topics are about games and game studies, so it makes sense that they would have a very high similarity.

Computing the Jenson-Shannon distance results in a matrix of similarity scores, which contains insight, but is difficult to interpret. This is where hierarchical clustering is useful. This technique creates a visual representation of the distance matrix as a dendrogram or tree-like structure and is much more meaningful to read and interpret. This technique for visualizing a topic model as a whole has been popular with digital humanists trying to analyze and interpret topic model. For example, Lincoln Mullen uses the technique in his in-progress book *Digital History Methods in R*.[81]

The dendrogram that emerged from the hierarchical clustering proved to be an excellent method for visually characterizing the topic model in one picture. The complete dendrogram has been included in Appendix B. In the following section, four specific clusters are highlighted because they give prominence to the dominant themes I discovered in the topics model.

The Technology Cluster

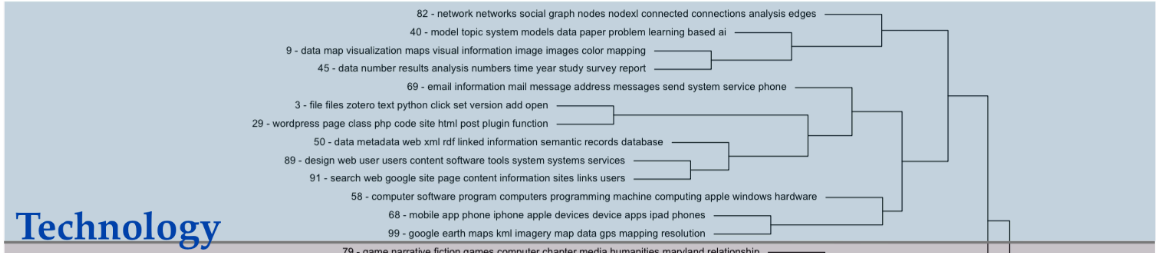

82 - network networks social graph nodes nodexl connected connections analysis edges
40 - model topic system models data paper problem learning based ai
9 - data map visualization maps visual information image images color mapping
45 - data number results analysis numbers time year study survey report
69 - email information mail message address messages send system service phone
3 - file files zotero text python click set version add open
29 - wordpress page class php code site html post plugin function
50 - data metadata web xml rdf linked information semantic records database
89 - design web user users content software tools system systems services
91 - search web google site page content information sites links users
58 - computer software program computers programming machine computing apple windows hardware
68 - mobile app phone iphone apple devices device apps ipad phones
99 - google earth maps kml imagery map data gps mapping resolution
Technology
79 - game narrative fiction games computer chapter media humanities maryland relationship

Figure (31). The methodology and technology topics cluster

[81] I used Mullen's tutorial as a starting point for my own analysis. http://dh-r.lincolnmullen.com/topic-modeling.html

Technology is featured prominently on digital humanities blogs (and blogs more generally). This set of topics has an instrumental relationship to technology. The computer is seen as a tool for analysis, a platform for publishing, or an application for research. Topics engaging technology in other epistemic modes did not feature the same configuration of keywords. These topics featured extremely technical language, like topic 3, which is clearly scholarly with its references to Zotero, but also deeply technical with references to file management on the Unix command line.

As can be seen in Figure (31), there are thirteen topics with technological keywords. This cluster included two major subclusters, one focused on *methodological* themes, and one focused on *technological* themes. The topics with methodological themes featured top words related to quantitative and data-driven analysis. Words like *data*, *numbers*, *analysis*, *model*, indicate a collection of blogs around quantitative and computational methodologies. Some of the topics even singled out specific methods, for example topic 82, which is clearly about social network analysis, or topic 40, which is about topic modeling and machine learning more generally.

Topics with technological themes focused on the very practical and ordinary aspects of technology. These topics evoked themes about installing Linux, configuring web-based content management systems, and how to use the Unix command line. These themes are not typically part of a high-level academic discussion about computation and "the digital" compared to the methodological topics, yet they are an equally important and foundational body of knowledge that needs to circulate. These topics represent the function of scholarly communication that Menzel called the *transmission of the ineffable*. The minutia of installing Apache Cocoon on the Ubuntu Linux distribution is too extreme a detail and not to be uttered in the pages of Literary and Linguistic Computing.

The Digital Cluster

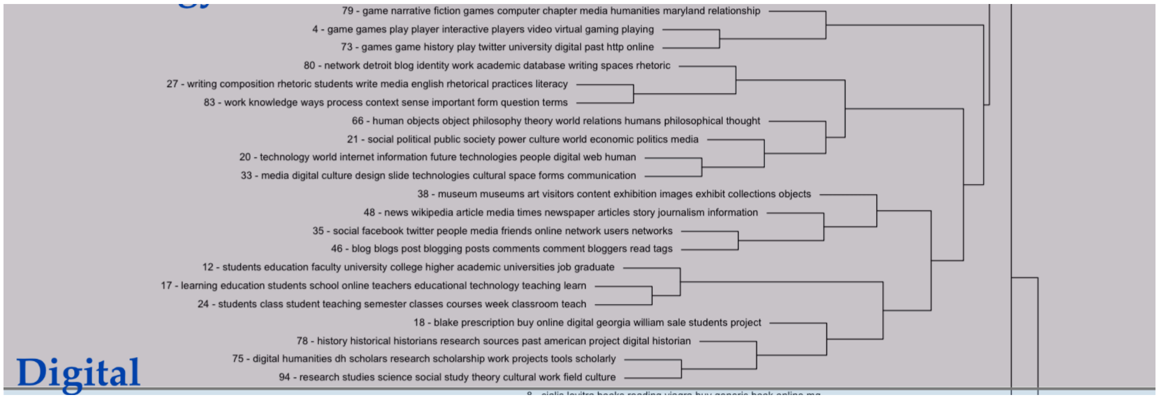

79 - game narrative fiction games computer chapter media humanities maryland relationship
4 - game games play player interactive players video virtual gaming playing
73 - games game history play twitter university digital past http online
80 - network detroit blog identity work academic database writing spaces rhetoric
27 - writing composition rhetoric students write media english rhetorical practices literacy
83 - work knowledge ways process context sense important form question terms
66 - human objects object philosophy theory world relations humans philosophical thought
21 - social political public society power culture world economic politics media
20 - technology world internet information future technologies people digital web human
33 - media digital culture design slide technologies cultural space forms communication
38 - museum museums art visitors content exhibition images exhibit collections objects
48 - news wikipedia article media times newspaper articles story journalism information
35 - social facebook twitter people media friends online network users networks
46 - blog blogs post blogging posts comments comment bloggers read tags
12 - students education faculty university college higher academic universities job graduate
17 - learning education students school online teachers educational technology teaching learn
24 - students class student teaching semester classes courses week classroom teach
18 - blake prescription buy online digital georgia william sale students project
78 - history historical historians research sources past american project digital historian
75 - digital humanities dh scholars research scholarship work projects tools scholarly
94 - research studies science social study theory cultural work field culture
Digital

Figure (32). The "digital" topics cluster

The digital cluster is a large collection of topics with several important sub-structures. Hierarchical clustering identified four significant subclusters within the digital cluster that I have labeled as *games*, *theory*, *public humanities*, and *digital History & Humanities*.

Video games are cultural objects like music, film, and literature. Games studies is emerging as a sub-discipline of and/or adjacent to the digital humanities. Games have also been a topic categorically marginalized by the mainstream humanities disciplines and blogs are one of the places where games studies is constituted. The fact there are three topics about gaming is an indicator that this area of study is significant.

There were also a series of theoretically oriented topics, focused on knowledge work, object oriented philosophy, social and political power, digital culture, and technology. While many of these topics addressed very classic question of knowledge and power, there was a distinctly digital inflection in their keywords. Two of these topics, 27 and 66, actually belong to a single blogger, Alex Reid, who writes quite extensively about the theoretical issues surrounding digital rhetoric and object oriented philosophy.

A small subcluster of topics touched upon news, social media, blogging, and interestingly museums. Threading through each of these topics is a concern with the public and how to engage the public. Blogs and social media have a very obvious public orientation, as do issues of news, journalism, and Wikipedia. The inclusion of the museum oriented topic may seem at first to be oddly placed (and it is in the outer layer of this small subcluster), but it makes sense. Museums are increasing concerned with making the digital collections accessible online to the public through online exhibits. The digital humanities is not simply the concern of English and History, the knowledge infrastructure that supports the humanities, libraries, archives, and museums are also implicated (more on this with respect to libraries below).

Digital history, curiously, does not have a significant presence in the topic model. The few topics with explicit historical keywords didn't have the weight to form a cluster of their own independent from general topics about digital humanities, William Blake, and the social sciences. In some respects this makes sense because history is often categorized as a social science rather than a humanities discipline. The affinity between the history and social science topics makes sense given the boundary spanning nature of the field.

The digital history and humanities cluster is also home to topic 75, a topic that attracted many of the "what is digital humanities" blog posts. Debating the nature of digital humanities is the foundation of DH blogs. The top words for topic 75 are a perfect distillation, into a weighted sequence of keywords, of the essence of digital humanities.

> Digital humanities dh scholars research scholarship work projects tools scholarly humanists project field computing methods history technology studies academic text thatcamp literary scholar collaboration media traditional teaching analysis data questions neh building literature collaborative graduate humanist university tool disciplines pedagogy humanistic lab critical academy technologies students cultural institute center twitter

Debating digital humanities is a genre of digital humanities scholarship into itself. What this topic shows is that blogs are the space where that genre grew up. What makes this topic so powerful is that it emerged as a distinct and coherent topic from a one hundred topic model of 39,976 blog posts. This simultaneously speaks to the spookiness of topic modeling, that anecdotal understandings of DH do hold up in the face of data, and that digital humanists really do blog a lot, perhaps too much, about digital humanities.

English Cluster

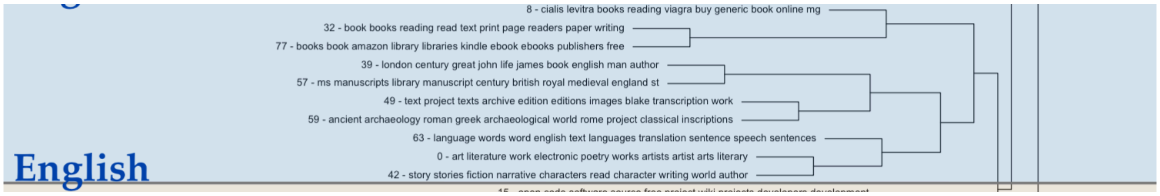

English
8 - cialis levitra books reading viagra buy generic book online mg
32 - book books reading read text print page readers paper writing
77 - books book amazon library libraries kindle ebook ebooks publishers free
39 - london century great john life james book english man author
57 - ms manuscripts library manuscript century british royal medieval england st
49 - text project texts archive edition editions images blake transcription work
59 - ancient archaeology roman greek archaeological world rome project classical inscriptions
63 - language words word english text languages translation sentence speech sentences
0 - art literature work electronic poetry works artists artist arts literary
42 - story stories fiction narrative characters read character writing world author

Figure (33). The English topics cluster

Perhaps it comes at no surprise that a cluster of topics foregrounding the interests of English would pop out of a topic model of digital humanities blogs. As Kirschenbaum (2010) noted, digital humanities is deeply implicated with English, and so English is deeply implicated in digital humanities blogs.

Figure (33) shows a collection of topics featuring top words relevant to English. This cluster has several substructures, like the two topic branch, topic 32 and topic 77, which are about books. Closely related to those two topics, but one branch removed, is topic 8, which appears to be simultaneously about books and drug treatments for erectile dysfunction. This is less a commentary about the virility of English faculty and more about the challenges of maintaining a blog on the Open Web.[82]

Books, literature, and lexicality comprise the three main subclusters of the English topics. The book subcluster is divided with topic 32's keywords featuring the materiality of books, *text print page writing reading*, on one hand and the digitality of books, *ebook amazon kindle*, on the other. The there is a historical literature subcluster that also includes the single archaeology topic

[82] Topic 8 exists because some of the blogs I scraped were the target of automated spam bots that exploit vulnerabilities in out of date Wordpress installations. Spammers can be diabolically clever and embed advertisements into the blog content in such a way that they are invisible to the author, yet visible to readers.

Libraries & Meta Clusters

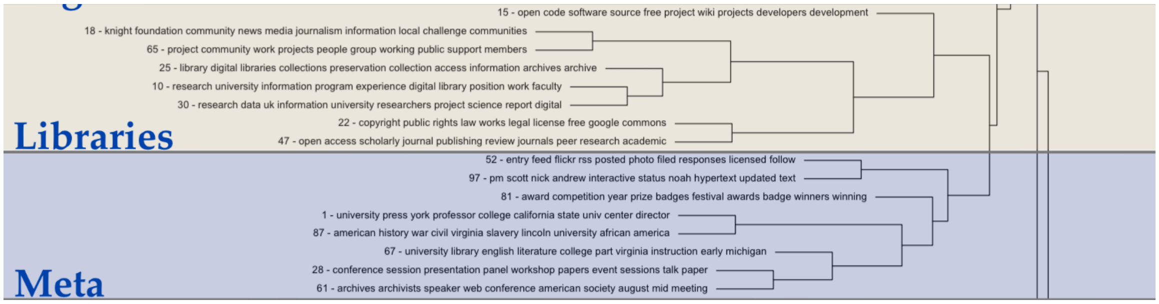

15 - open code software source free project wiki projects developers development
18 - knight foundation community news media journalism information local challenge communities
65 - project community work projects people group working public support members
25 - library digital libraries collections preservation collection access information archives archive
10 - research university information program experience digital library position work faculty
30 - research data uk information university researchers project science report digital
22 - copyright public rights law works legal license free google commons
47 - open access scholarly journal publishing review journals peer research academic
Libraries
52 - entry feed flickr rss posted photo filed responses licensed follow
97 - pm scott nick andrew interactive status noah hypertext updated text
81 - award competition year prize badges festival awards badge winners winning
1 - university press york professor college california state univ center director
87 - american history war civil virginia slavery lincoln university african america
67 - university library english literature college part virginia instruction early michigan
28 - conference session presentation panel workshop papers event sessions talk paper
61 - archives archivists speaker web conference american society august mid meeting
Meta

Figure (34). The library topics cluster

Within the cluster in figure (34) are two distinct, yet interconnected thematic groups, libraries and meta. The two subclusters are distinctly labeled, but the clusters are closely inter-related by both the quantitative measure and qualitative readings of keywords. Semantically, this grouping of topics makes

These two clusters concern topics about "libraries" and topics about various aspects of academic life. The meta-topics are not miscellaneous (those are a separate cluster discussed below) but are about the structural aspects of the digital humanities as a discipline. For example, topics 28 and 61, which are both about academic conferences or topic 10, which is about academic jobs.

Topic 22 and topic 47 are topics on classic library themes. Copyright issues, identified by topic 22, and open access issues, identified by topic 47, are very closely related (the spooky intelligence of topic modeling + hierarchical clustering at work) and are the subject of librarian discourse and expertise. Charles W. Bailey's blog, *digital-scholarship.org*, dominates the top topic proportions for both of these topics,[83] but librarians like Paul Courant and former librarian Molly Kleinman (who worked with Paul on copyright issues) are also present. While copyright and open access issues are important across the digital humanities,

[83] Bailey is an *extremely* prolific blogger who writes about digital scholarship and has self-published, in full open access, multiple books on electronic publishing and digital scholarship. He is a force.

these two topics surface extremely technical discussions about the minutia of creative commons licensing,[84] or the status of court cases about intellectual property[85]. In both of these cases, librarians are bringing their deep technical knowledge forward. The technical details and nuance are not unlike the technology topics discussed above. Instead of the nitty-gritty of Linux distributions, these librarians are talking about Share Alike and NonCommercial copyright licenses.

On a closely related branch, although visually distinct on the dendrogram, was a topic 15, the open source software topic. This was a small topic, focused mainly on Bess Sadler who manages the software engineering team at Stanford University Library. She writes about software development not at a technical level, but at an academic administrative level (hence the topics placement in the library cluster rather than the technology cluster). Again, we see librarians featured prominently in detailed discussions about technology in an open and informative way.

The subject matter of the meta topics are addressed in a subsequent section about higher level themes and discussion on blogs. It is important to note the close relationship between library topics and, for lack of a better term, “administrative” topics, such as topic 1. The overlap reflects the unique organizational position of libraries with respect to the digital humanities. While scholars are often seen as distinct from the administrative and organizational apparatus of the university, the library is intertwined. Perhaps libraries have served as a bridge between the academics and the administrators across which the infectious enthusiasm for the digital humanities has crossed. Administrators are very excited about digital humanities, some argue because of the broader

[84] Molly Kleinman wrote a series of blog posts about how to effectively use Creative Commons licenses. http://mollykleinman.com/2008/10/20/cc-howto-no-derivatives/

[85] Christine Fruit is a lawyer and librarian who bloggs about copyright and open access issues. In one post she wrote a blow-by-blow about a copyright infringement suit brought against UCLA. https://campuscopyright.wordpress.com/2011/02/13/immunity-contracts-and-copyright-an-update-on-aime-vs-ucla/

attention from the public and funders. This enthusiasm has been viewed with caution and skepticism from the academic side of the digital humanities (Grusin 2014; Chun and Rhody 2014) for fears of corporatism and neoliberal economic factors having a negative influence upon the community.

Other Clusters

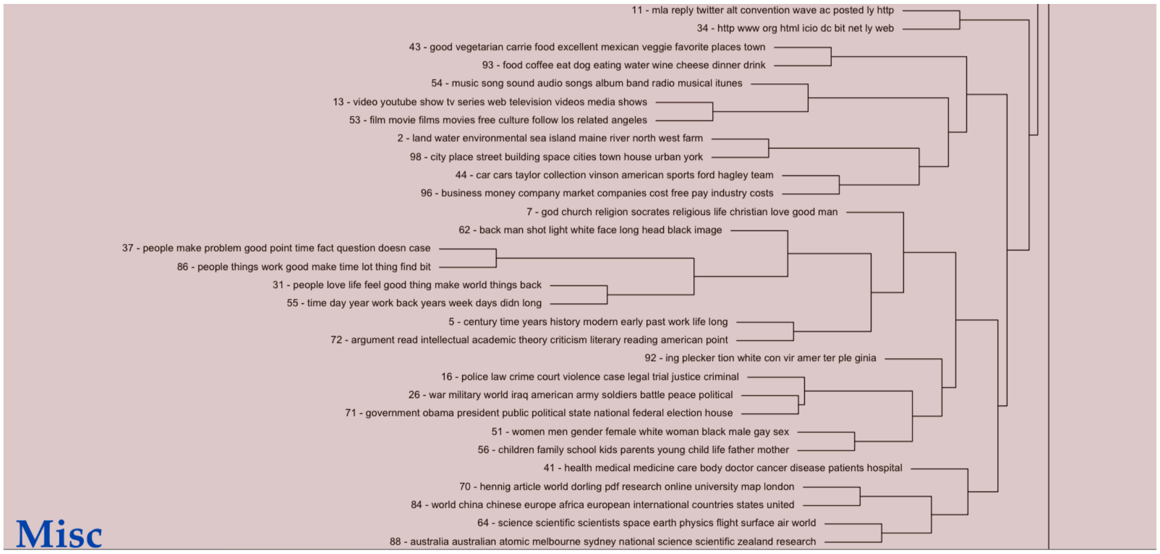

11 - mla reply twitter alt convention wave ac posted ly http
34 - http www org html icio dc bit net ly web
43 - good vegetarian carrie food excellent mexican veggie favorite places town
93 - food coffee eat dog eating water wine cheese dinner drink
54 - music song sound audio songs album band radio musical itunes
13 - video youtube show tv series web television videos media shows
53 - film movie films movies free culture follow los related angeles
2 - land water environmental sea island maine river north west farm
98 - city place street building space cities town house urban york
44 - car cars taylor collection vinson american sports ford hagley team
96 - business money company market companies cost free pay industry costs
7 - god church religion socrates religious life christian love good man
62 - back man shot light white face long head black image
37 - people make problem good point time fact question doesn case
86 - people things work good make time lot thing find bit
31 - people love life feel good thing make world things back
55 - time day year work back years week days didn long
5 - century time years history modern early past work life long
72 - argument read intellectual academic theory criticism literary reading american point
92 - ing plecker tion white con vir amer ter ple ginia
16 - police law crime court violence case legal trial justice criminal
26 - war military world iraq american army soldiers battle peace political
71 - government obama president public political state national federal election house
51 - women men gender female white woman black male gay sex
56 - children family school kids parents young child life father mother
41 - health medical medicine care body doctor cancer disease patients hospital
70 - hennig article world dorling pdf research online university map london
84 - world china chinese europe africa european international countries states united
64 - science scientific scientists space earth physics flight surface air world
88 - australia australian atomic melbourne sydney national science scientific zealand research
Misc

Figure (35). The miscellaneous topics

The miscellaneous cluster in Figure (35) exists because this cluster of topics does not have as strong of thematic ties as some of the other clusters. This large branch of the dendrogram includes topics whose keywords indicate nonacademic themes like personal health or food (i.e., topic 41 and topic 43), as well as topics with keywords literary criticism, theory, and arguments (see topic 72). Based on keywords alone, topic 72 would seem to be the purest form of academic discourse. That topic is on a subbranch featuring a set of very broad topics about time, work, and people. Unlike the Technology or English clusters, the topic model didn't surface a clearly distinguishable theme. This large cluster of topics demonstrates the diversity of topics and themes written about on blogs.

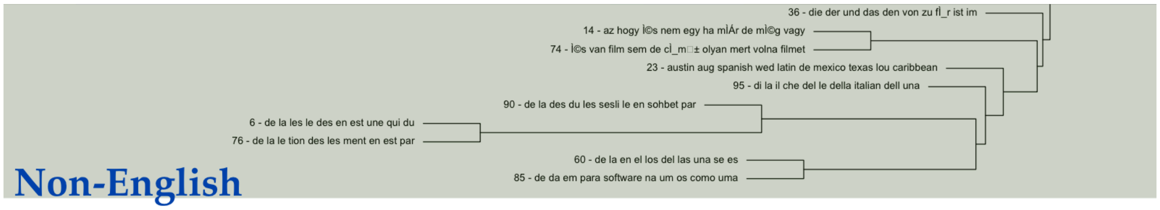

36 - die der und das den von zu fÌ_r ist im
14 - az hogy Ì©s nem egy ha mÌÁr de mÌ©g vagy
74 - Ì©s van film sem de cÌ_m□± olyan mert volna filmet
23 - austin aug spanish wed latin de mexico texas lou caribbean
95 - di la il che del le della italian dell una
90 - de la des du les sesli le en sohbet par
6 - de la les le des en est une qui du
76 - de la le tion des les ment en est par
60 - de la en el los del las una se es
85 - de da em para software na um os como uma
Non-English

Figure (36). The cluster of non-English topics

The finally cluster of topics in Figure (36) to share a branch were the topics with non-English keywords words. Some of the blogs are written in a mixture of languages. The topic model pulled out those posts and the hierarchical clustering put them together. These topics, which are composed mainly of stop words, are mainly French and Spanish. There was a Hungarian topic (topic 14) that shows a hit of the international reach of the digital humanities.

The dendrogram shows a extremely distant reading of the landscape of digital humanities blogs. The hierarchical clustering only takes into account the shape and contours of the topic distributions generated by the model. There is no semantic weight, social, or historical context factored into the tree structure. The broad clusters, Technology, Digital, English, Libraries, Meta, Misc, and Non-English are labels derived from a qualitative coding of the topic clusters and a saturated familiarity with the general themes of each topic. However, the dendrogram representation of the topic space does not give justice to the fact that not all topics are created equal. Alpha values have not been included in the generation of the dendrogram nor has a richer reading of the topics that includes their top-ranked documents.

Hierarchical clustering generated a map helps navigate the thicket of traces collected and generated. This is a map to the territory captured in the HTML document described in Chapter 4. The document connects the topic keywords; it's size and shape, and the top twenty documents for each topic. The next section includes a deeper, closer reading of digital humanities blogs that includes all available traces, the components of the model and the original document content. From an ethnographic reading of the assembled traces emerged four categories of informal scholarly communication.

Categories of Informal Scholarly Communication

This section introduces four categories of informal scholarly communication on digital humanities blogs by digging into a couple themes from each category. These categories emerge from the trace ethnography that includes both the topic model and the content of the original documents. These descriptive categories take into account the document content, not just the topic keywords, as well as a few other quantitative measures, such as proportion over time. The categories themselves are framed by the formal/informal distinction of scholarly communication where each of these categories reveals a distinct relationship to formal scholarly communication:

- **Quasi-academic:** Topics whose subject matter touched upon themes resembling formally published scholarly communication. The content of these posts fits in with the classic cycle of scholarly communication where the pre-publications are circulated informally on a course towards formal publication.
- **Meta-academic:** Topics whose subject matter is focused on maintenance and organization of a social group. These posts do the social work of managing the institutional and organization aspects of the community.
- **Para-academic:** This category is especially significant because it describes posts that might contain serious academic or intellectual content, but they don’t have a place in formal publishing, in form or in content. This includes many of the technical discussions around tools, technique, and methodology that are vital to DH work, but have few outlets in journals or monographs.
- **Extra-academic:** This is the decidedly nonacademic space of discourse that the structural features of blogs allow for. These can be recipes or restaurant recommendations. These posts are the reminders that the people writing these blogs are human beings and are an important signal of a blogger’s humanity.

The topic model revealed that the discursive space of informal scholarly communications on blogs is not uniform in its relation to the formal/informal distinction. There are varying degrees of formality on digital humanities blogs. This spectrum ranges from *quasi-academic* blog posts relating to historical, theoretical, or literary matters that, at least thematically, are not much different from the themes of formally publishing journal articles and books. At the other extreme of this spectrum are *extra-academic* blog posts about favorite recipes, family matters, or local politics. These theme's extreme informality are hard to classify as scholarly communication, but are mixed into this corpus because of the structural features of blogs as a platform. Scholars can mix *quasi-academic* with *extra-academic* themes as much or as little as they would like, the Open Web does not distinguish. I also found themes in the discursive space between the quasi-academic and extra-academic. *Meta-academic* and *para-academic* posts have intellectual value for the digital humanities, but don't have a place in the ecosystem of formal, print-centric scholarly communication.

Meta-academic topics focus on the organizational work of the digital humanities as a community. These are the practical postings about jobs or conferences, but also the existential postings about the nature and definitions of digital humanities as a concept. Para-academic topics show how blogs fill a discursive niche, the "How To," ignored by the traditional print-centric journals, and books.[86] Para-academic topics cover deeply technical themes, like how to install, configure, and use Linux, or topic modeling, or discussions about project management. Given the nature of the work in the digital humanities, this is extremely important subject matter, but (at least historically) these topics have no place in the lexicon of formally published humanities scholarly communication.

[86] Digital humanities, and humanities more generally, does not have a conference paper culture. The DH conference might be one of the few that asks for something more substantial at conferences, but they are still extended abstracts, not formal publications like we might see at a more technical ACM style conference with ten-page paper submissions.

Quasi-Academic

It comes as no surprise that academic bloggers write about academic themes. Of the hundred topics in the model, thirty-two are “quasi-academic.” This category represents blog posts focused on academic themes. Academic themes are subject matter and language that could be reasonably expected to appear in formal scholarly publication.[87] The top ten blogs with the highest cumulative proportion of quasi-academic topics belonged to faculty, students, or scholarly projects. The top words, top documents, and top domains of these topics featured academic themes. However, a closer reading of these topics showed writing that was *close* to formal academic discourse, but didn’t quite fit, due to content or style, within the normal boundaries of traditional academic publishing.

For example, topic 51 has the top words *women men gender female white woman black male gay sex*. This topic was labeled “Identity Politics” because the top-ranked documents discuss a variety of issues relating to race, gender, sexism, power, and privilege. The politics of power and identity are growing themes within the digital humanities, which have been rightly criticized for historically ignoring these areas, especially in the days of “humanities computing.” Blogs have become one of the places where thinking, writing, and discussing these themes is embraced. The challenge then shifts away from publishing, but finding readers and attention for new blogs and initiatives.

The top blog post for topic 51 is from the blog *muckleado.com*, by Candace Nast, a scholar, instructor, and technology consultant who specializes in “the intersections of technology, history, and gender, both in and out of higher education.”[88] The blog post, titled “Constructing an Identity” compares and contrasts two book chapters on “how the multiple dimensions of identity impact

[87] I cannot know or make claims about the extent to which the ideas and themes are actually published, that would require a different study performing a comparative analysis of scholar’s formal and informally published works. Such a comparative study would also provide a means of testing the robustness of my four categories.

[88] See muckleado.com’s “About”com's about page http://muckleado.com/about/

an individual."[89] The post was written in August of 2005, when the author was an undergraduate in women's studies. The post seems to be a response to weekly readings in a seminar class. The posts before and after "Constructing an Identity" share a similar structure, i.e., comparisons of two or three academic articles or book chapters.

> I didn't blog regularly until I did an independent study on digital edition design with Matt Kirschenbaum the spring of my first year in the UMD literature Ph.D. program, when I began doing a reading journal blog for my weekly readings. Visconti

From Class to Community

Several of the blogs in the collection began as coursework but evolved into spaces for public writing. Trevor Owens, a historian and archivist who blogs about digital history, told me about how his blog emerged out of a course requirement from a digital history class he took while getting his masters degree.[90] The social media research danah boyd started her blog while as an undergraduate at Brown and now she is a highly regarded public intellectual (I'll discuss boyd more below). The blog *Wonders and Marvels*, which is now "A community for curious minds who love history, its odd stories, and good reads" was created by historian Holly Tucker at Vanderbuilt while teaching an undergraduate history course. The blog has serendipitously evolved since its humble beginnings,

> And then suddenly and with little warning, it all morphed into something much bigger and wonderfully unexpected. … First other professors started coming by and offering up their insights on the past. They were followed by equally talented writers of historical fiction. … Wonders & Marvels is now a place for specialists and non-specialists to revel in the stories of the past. It also offers learning opportunities for interested Vanderbilt students to work with Professor Tucker on building the site. Student interns have a chance to interact regularly with scholars and other experienced authors, and even write a few posts of their own.[91]

[89] To see the blog post visit http://muckleado.com/2005/08/constructing-an-identity/

[90] Trevor posted the answers to my interview questions on his blog: http://www.trevorowens.org/2014/11/wherein-i-answer-13-questions-about-digital-humanities-blogging/

[91] The "About" page at wondersandmarvels.com has an account of the blogs origins: http://www.wondersandmarvels.com/about

The Open Web made it possible for other professors and professional writers to discover, comment, and contribute to *Wonders and Marvels*. The contributors to *Wonders and Marvels* are not only academic faculty; the list includes student interns *and* professional fiction writers.[92] Today, Wonders and Marvels is a highly respected and award winning blog in the History blogging community.[93] While not every class blog evolves into a community, open access and open publishing are prerequisite *conditions of possibility*. Had the blog been locked behind courseware tools and access control, it wouldn't have been discoverable. Its publicness also created a bridge spanning the ivory tower and the historical fiction community, a relationship Hunter has explicitly cultivated by including professional writers as regular contributors. This is scholarly communication and public history at its best and it wouldn't be possible without open infrastructure paired with compelling content and a willing base of contributors (who were able to discover it because it was open).

Academic Themes in the Public Context

Another highly ranked post in a topic labeled *Identity Politics* is by dana boy titled "Considering racism…" The post, from March of 2001, reflects on a talk about the "language of racism" she had attended earlier that day. The post is casual in tone, serious in subject matter, and raw in its conceptual formation; like an entry in an intellectual diary. The post is deeply personal, and probably not appropriate in formal academic venues, but it never the less engages themes about the structures and hierarchies of power and racism in our society; these are deeply academic themes. This post engages academic subject matter, but situates the discussion in a personal context.

The top blog post is by Mike O'Malley, a history professor at George Mason University and "pioneer in digital media and history." It is titled "A Brief History

[92] See the list of regular contributors here: http://www.wondersandmarvels.com/regular-contributor

[93] See the award announcement here: http://www.wondersandmarvels.com/2012/01/wonders-marvels-wins-cliopatria-best-group-blog-award.html

of American Money" and is a long post responding to a *Wall Street Journal* news article on the GOP's rhetoric about returning to the gold standard. The post is simultaneously a response to current events and pop history.

> The GOP associates [the gold standard] with a return to tradition, something the history of American money doesn't really bear out. The U.S. was on a gold standard for less than thirty years, all told, and the history of our money reflects not stability and tradition but innovation, compromise and experiment.[94]

The post traces the history of various experiments with money starting in the colonial period and through the Revolutionary War then the Civil War, and discusses the shifts between metal and paper currency and the *several instances* of adoption and rejection of the gold standard. The history of money in America is complicated, and in a nice rhetorical flourish, O'Malley shows how even the *Wall Street Journal* flipped its opinion of the gold standard (against the standard in the post WWI era, and support in 2012).

The post enriches the reader's understanding of a contemporary political and economic issue through historical contextualization. O'Malley specializes in American history and even wrote a book on the history of race and money in America.[95] Here is a case where the directionality of ideas is working in the *opposite* direction of the current models of scholarly communication. O'Malley is writing informally about themes he has *already published formally*. His blog is not a forum for working out ideas; rather it is a place to make them available (in both the material and intellectual sense) to the "public." The post doesn't necessarily represent any new historical thought or arguments; rather it mixes O'Malley's historical work and a contemporary issue to provide public commentary. Like danah boyd's post above, we can see deeply academic themes, racial politics and history, contextualized with more informal themes, personal reflection and current events.

[94] “"A Brief History of American Money,”" http://theaporetic.com/?p=4101

[95] *Face Value: The Entwined History of Race and Money in America* http://www.amazon.com/Face-Value-Entwined-Histories-America/dp/0226629384/ref=sr_1_1?ie=UTF8&qid=1331924266&sr=8-1

By Academics For Academics

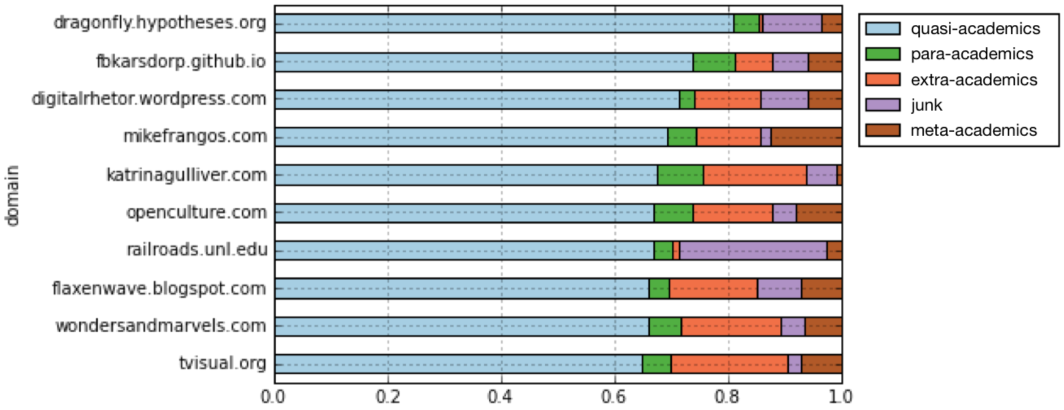


Figure (37). The top 10 blogs with highest proportion of academic topics. All of these domains belong to scholars or primarily scholarly activities.

All of the blogs with the highest cumulative proportion of academic topics belong to scholars, groups of scholars, or are scholarly projects. Not all of the posts they write are public oriented. Another common mode of the *quasi-academic* genre is as a dumping ground for the piles of nearly formal academic writing which have no place to go. This could include classwork by students, but also transcripts of academic conference talks.

The last example comes from an *academic* topic labeled *Latin Studies*. The topic brings together posts touching upon themes about Latin American studies, Central and South American history, and the Spanish language. One of the top posts is a scholar's prepared statement for a roundtable discussion. The talk highlights the "the multivalence of early modern Spanish legal identities, identities that often hinged upon the perpetuation of legal fictions easily absorbed by the labyrinthine, contingent space of the judiciary."[96] Clearly, an academic wrote this.

[96] For the full post see: http://parezcoydigo.wordpress.com/2008/12/29/clah-andean-section-roundtable/

At conferences, humanities scholars typically read from prepared documents rather than speak extemporaneously as in technical or scientific scholarly conferences. Scholars often post this material on their blogs to attract a broader audience than just the attendees of the conference.[97] A humanities conference talk, especially one with some peer review, is obviously a kind of academic discourse, but also one that doesn't always have a home in the scholarly record.[98] Blogs have then emerged as a place for this form of writing.

The posts described above are examples of the kinds of posts found in the *quasi-academic* category. The topic model successfully brought together blog posts whose content engaged classical academic themes, but upon close reading shows how they are distinct from formal scholarship in ways beyond the publication venue. The *quasi-academic* genre raises (at least) one crucial issue. The existence of valuable academic content in this discursive space raises the challenge of *discoverability*. Blogs, in their most raw form, do not have an editorial process. Open publishing means anyone can start a blog, and the size of the DH community shows how members take advantage of these dynamics.

Meta-Academic

Meta-Academic contribute composition to the digital humanities as a community. These themes function at a social and organizational level. These topics can be structural like topics10 and topic 28, which are job listings and CFPs. They can be functional like topic 17 and topic 25, which are about students, learning, and managing a classroom. Finally, these they can be reflexive like topic

[97] I also suspect there is a bit of two-bird-one-stone labor saving, they can write a conference talk and use it in multiple ways. Finding the time to blog was one of the issues my respondents said was hard in my interviews about their blogging practice. Anything they can use as content for the blog is fair game, including my interview questions!

[98] The digital humanities conference posts extended abstracts, but classic humanities conferences like MLA or AHA don't provide have archived conference proceedings. For another example, Alex Reid posted "Composing objects: prospects for a digital rhetoric #cwcon," which was the text of his keynote at the 2012 Computers and Writing conference. http://alex-reid.net/2012/05/composing-objects-prospects-for-a-digital-rhetoric-cwcon.html

12, which waxes about the state of the humanities, or topic 75, the "What is DH?" topic.

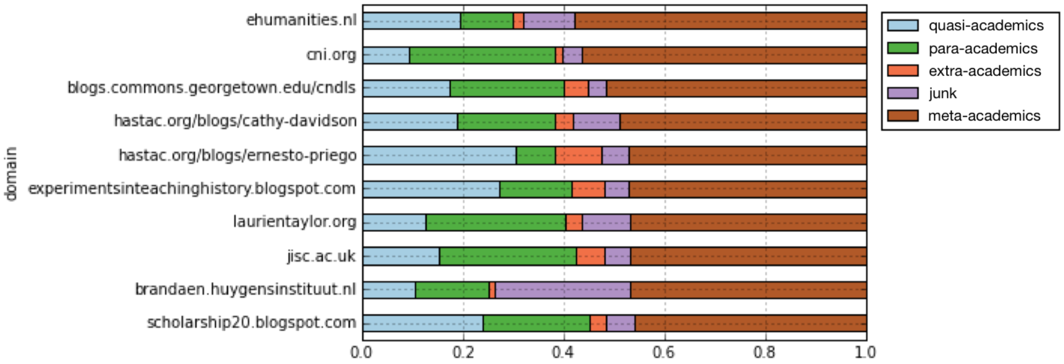


Figure (38). Blogs with highest cumulative proportion of meta-academic topics. Most are institutional blogs.

Of the blogs with the highest proportion of meta-academic topics, four belong to institutions, three are scholars, two are librarians, and one is a technologist. The four institutions are clearing houses for information about digital technology and the academy. CNI, the Coalition for Networked Information, is a membership organization dedicated to "supporting the transformative promise of digital information technology for the advancement of scholarly communication and the enrichment of intellectual productivity."[99] Similar to CNI, JISC, the institution formerly known as the Joint Information Systems Committee, is another membership organization advocating digital technology, in the UK. The eHumanities group does the same in the Netherlands, but is sponsored by the government. CNDLS, or the Center for New Designs in Learning and Scholarship at Georgetown, focuses on digital technology for teaching. Each of these are institutions whose purpose is to generate and circulate information about technology and the academy.

Much of the content these organizations produce comes in the form of *reports*, an underappreciated genre of grey literature and informal scholarly communication.

[99] CNI "About" page http://www.cni.org/about-cni/membership/key-benefits/

Based upon the model's analysis of their blogs, most of the attention of these institutions oriented towards science and technology. The top documents for topic 30, *research data uk information university researchers project science report digital*, are mainly blog posts on reports produced by JISC and CNI, many of which are focused on Open Science, data-driven research, and digital curation. However, some of this work, namely the data-driven research and digital curation, has relevance in the digital humanities. Blogs have been the engine by which they circulate their reports.

DH Jobs for Everyone

While it is a banal observation, blogs, RSS feeds, and now Twitter have been crucial community infrastructure for the circulation of job postings. Topic 10, *research university information program experience digital library position work faculty*, consolidated blog entries that were job postings. Notice the word "library" in that list; this is significant because it shows how digital humanities is connected to broader structural transformations within the humanities (and beyond). *All* of the top twenty documents in topic are job postings, but they weren't all tenure track humanities positions.

The myth that the digital humanities revolution will lead to more jobs in and for humanities graduates[100] has seemed to vaporize.[101] That said the proportion of job posting related writing has increased over the past decade.

[100] In 2011 former Google executive Marissa Mayer exclaimed Google would be hiring four to five thousand people "from the humanities or liberal arts." http://www.timeshighereducation.co.uk/416190.article Unfortunately, the myth did not appear to become reality: http://carefullydisordered.blogspot.com/2012/04/google-humanities-and-what-do-you-get.html

[101] https://www.insidehighered.com/news/2014/05/08/digital-humanities-wont-save-humanities-digital-humanists-say

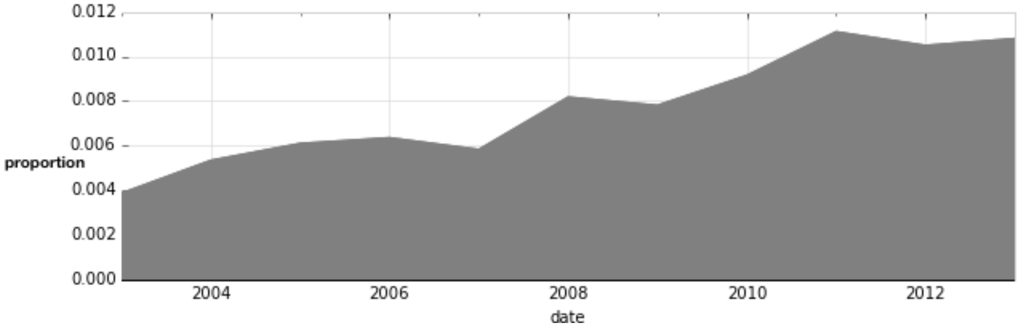


Figure (39). The increasing yearly proportion of topic 10, job postings, from 2003 to 2013.

Figure (39) shows the topic proportion per year for topic 10, the job postings topic. The figure shows the yearly proportion gradually increasing in relative proportion to other topics in the corpus year after year. What this figure, and topic 10, are measuring are clusters of words used specifically in the context of job postings. The top twenty-five keywords for topic 10 reveal a bit more insight into the nature of these conversations.

> research university information program experience digital library position work faculty applications graduate department management development job staff skills application center technology including professional science support knowledge projects services school grant librarian programs apply project candidate students degree academic media communication campus working committee http ability assistant director technologies required training

The fact that the terms *research* and *university* are the top two keywords might hint at the kind of positions being discussed, but they might also describe the type of university. Looking at the remaining top words, we see terms like *library*, *librarian*, *staff*, and *director*. These terms indicate the discussions of DH jobs trend towards the more administrative side of the academic job market, though, the top words only show so much. The top twenty documents, which are all job postings, paint an even clearer picture.

Of those twenty positions, only seven of them are the coveted, tenure-track, traditional academic jobs. The remaining thirteen are non-tenure track, staff, or administrative positions. Not all of the tenure track positions were specifically digital humanities, although all of them had a "digital" flavor such as "digital storytelling," or "digital culture." Two of the academic positions were in schools of information/library science, the rest were faculty positions in traditional humanities departments. Of the nonacademic positions, most of them were in academic libraries.

The rise of non-tenure track academic jobs has been marshaled by the "revolutionary" rhetoric of digital humanities. In an article in the *Chronicle of Higher Education*, William Pannapacker reflected upon the 2011 Digital Humanities conference at Stanford University, a conference he claims "marked the launch of the #Alt-Academy." Alt-ac, or alternative academic careers, is a term of solidarity for PhDs who have chosen work as technologists, librarians, administrators, or otherwise opted out of the tenure-track yet remain within higher education. Alternative academic positions, as defined by the alt-ac media commons website, those that are

> Off the tenure track, but within the academic orbit ... in universities and colleges, or allied knowledge and cultural heritage institutions such as museums, libraries, academic presses, historical societies, and governmental humanities organizations.[102]

Alt-ac gives scholars a unifying narrative; this is particularly true of humanities PhDs who have taken academic positions as librarians, curators, administrators, or other academically collateral positions. Increasingly, traditional disciplinary conferences like Modern Language Association (MLA) include workshops dedicated to scholars with, or interested in, alternative academic careers, driven in part by increasing anxieties about humanities graduate's uncertain futures.

What began as a hash tag on twitter (#alt-ac) has now become a rallying point for discussions about alternative careers for PhDs in the humanities. The discussion quickly outgrew twitter and evolved into an online edited volume at MLA's MediaCommons community platform, whose structure and form resemble many

[102] This definition of alt-ac comes from *#alt-academy* a MLA MediaCommons project aimed at supporting the alternative academic community. The site includes a series of essays and discussions about the structural shift away from tenure track towards alt-ac in academic institutions.

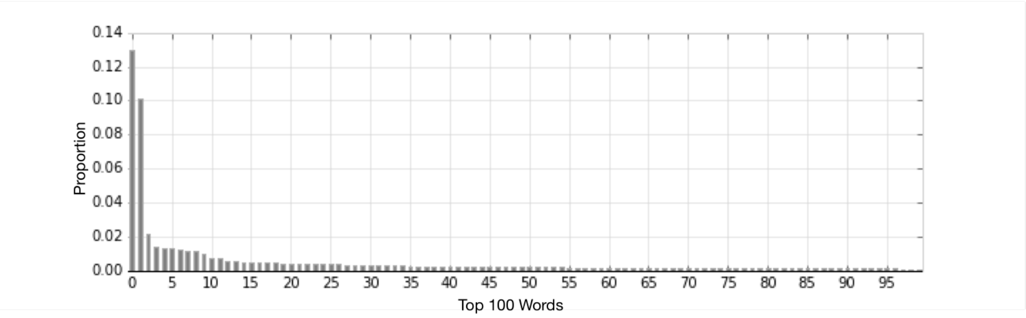
of the micro-community structures I encountered in and through digital humanities blogs:

> *#alt-academy* is both an edited collection and the embodiment of a grassroots, publish-then-filter approach to networked scholarly communication. All community members can comment on existing essays and freely publish relevant work as part of this site, thereby making their content available at a stable URL and discoverable through search. However, only selected contributions will be featured on the #alt-ac home page as part of our edited “clusters.”[103]

The alt-academy site at MediaCommons espouses the lofty ideals of the Open Web as a platform for scholarly communication. It is open-access, post-publish peer reviewed, and contributions are measured solely by their merit (not the status of their authors). The cast of contributors includes senior and junior scholars, post-docs, librarians, administrators, developers, and even at least one graduate student (although he identified himself as a designer).

Topic 75: The “What is Digital Humanities” Topic

Topic 75 is the “digital humanities” topic. The top ten words are *digital humanities dh scholars research scholarship work projects tools scholarly* and of those, the top two words, *digital* and *humanities* have significantly more probability than the remaining eight. These top words are the quintessential list of digital humanities jargon, which is interesting considering this collection comes from an unsupervised algorithm.



[103] #alt-academic“ “How It Works” http://mediacommons.futureofthebook.org/alt-ac/how-it-works

Figure (40). The probabilities of the top-ranked words for topic 75. The top two, digital and humanities, are much higher than the remainder.

Tracking the proportion of the digital humanities topic over time tracks the dramatic rise of digital humanities, especially in the past five years.

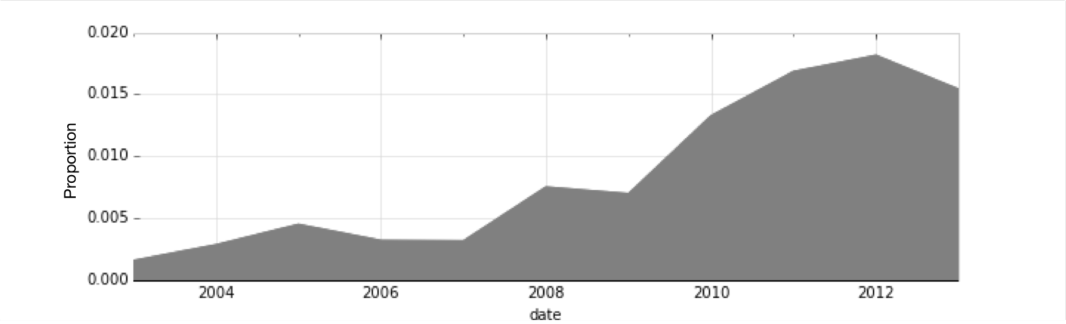


Figure (41). The rise of "digital humanities" as a theme across DH blogs.

The appearance of the *Compendium of Digital Humanities* is generally considered to be the moment when the term "digital humanities" went public, but Figure 41 shows the term really didn't start becoming thematically popular on blogs until 2008, the year of the "Digital Humanities Manifesto," and before the first big wave of digital humanities monographs. *Digital Humanities Quarterly*, the community's premiere journal, had just launched in 2007.

The topic peaks in 2012 and begins to drop in 2013; unfortunately there is no data for 2014 (or beyond) yet to see if this downward trend continues. Scholars have grown tired of writing and talking about digital humanities as a subject and maybe have shifted into discussions about *doing* DH rather than *about* DH. What this high, topic/thematic-level perspective shows is a rough approximation of the meta-conversation and how bulk of these conversations occurred between 2008 and 2013.

Visibility of Meta-Academic discourse

The job postings and #alt-ac discussions on MediaCommons represent the meta-academic themes that perform the *visible* maintenance and plumbing work of the digital humanities as a community of practice. Meta-academic topics are the *visible college of scholarly communication*. Remember, in 1986 Price defined the *invisible college* as a

> [Commuting circuit of institutions, research centers, and summer schools giving them an opportunity to meet piecemeal, so that over an interval of a few years everybody who is anybody has worked with everybody else in the same category (1985).

Price was writing in an era without the same open infrastructure to enable heavily mediated forms of informal interaction. Today, blogs perform some of the same functions of connecting members of the community *and* even helping organize some of those older functions. Blogs are one way of circulating information about the many digital humanities summer schools like the Digital Humanities Summer Institute (DHSI) and Humanities Intensive Learning + Teaching (HILT).[104]

Meta-academic themes enable us to uncover what Leah Lievrouw called the *process* of scholarly communication. While the *structure* of scholarly community can be revealed in citations networks, the *process* has historically been difficult to study because it leaves no trace. Today, those informal interactions are mediated through digital technologies like blogs, which have the side effect of leaving a discoverable trace.

Community building efforts like the #alt-ac MediaCommons site or the distributed, but discoverable, conversations of the "what is dh" topic, are documentary residue of the *processes* of becoming and maintaining a scholarly community. Even if the public residue of a community's organization is not the entire picture (there is still important work that happens off the Open Web) analyzing these conversations traces of meta-academic discourse are more than we've ever had in the past and provide an important record of a scholarly communities *emergence*.

[104] The Digital Humanities Summer Institute has been offering summer workshops since 2001. http://www.dhsi.org/ HILT is a newer program launched in 2014. http://www.dhtraining.org/

Para-Academic

Para-academic themes have two constituencies. First, scholars who need a space to discuss the technical and methodological details of their research. Second, a space for librarians and technologists to discuss the details of *infrastructuring the digital humanities*. While technical discussions bring these groups together, they have different implications for scholarly communication and publishing.

The Para-academic category encapsulates technical and methodological themes. The themes of the para-academic genre ranged from deeply technical posts about installing Linux to using topic models to commentary on intellectual property. The para-academic category is also a place where the two largest constituencies of the digital humanities community, tenure track scholars and librarians, come together. One of the noticeable features of the para-academic genre is extremely technical language. Such technical writing emerges from one of the most important forms of writing on blog, the "how to" post, which might even warrant its own place amongst the genres of informal scholarly communication.

The "How To" Post

Topic 3, "How To & Tools," aggregated blog posts where authors documented their experiences installing, setting up, and using software. The "how to" post is not only its own subgenre of digital humanities blogging it is a genre of web in general.

The top twenty-five words for topic 3 are a collection of extremely technical terms related about managing software and tools.

> file files zotero text python click set version add open line script download code create run server command directory install format folder save select make document pdf txt step list xml work copy import program database image application options start time process mac note running key system find browser type

These terms about file management, the manipulation of data and information, and the maintenance of systems. What is important to note about these terms is that they are *not* related to programming. For example, even the title in a post by librarian and technologist Quinn Dombrowski, "Installing cocoon on Ubuntu,"[105] is mired with technical jargon. The post is a detailed description about how to install Cocoon, a web application server, on a Linux distribution called Ubuntu. But, as the post immediately points out, this "guide was written for Intrepid." which means it is for a very specific version of the Ubuntu Linux distribution, version 8.10, also known as Intrepid Ibex.

While Dumbrowski is a librarian and technologist, the "how to" theme also has a small share of tenure-track faculty who write about similar subject matter. The para-academic category shows how blogs are a place where academic faculty can *publicly* write about the technological and methodological dimensions of computational humanities research practice. It is also a place where they can experiment with computational methods without the same recourse as formal publications (i.e., rejection, dejection, or even hostility). This is not to say there is no space for writing about technical or methodological topics in the (digital) humanities. The point is blogs are a space for hashing out the details and best practices of techniques separate from the requirement to work out rhetorical, theoretical, and epistemological implications.

For example, the post "Using SEASR's Workbench to Explore the Past ...,"[106] by historian Ian Milligan introduces readers to SEASR, describes how to use it, and most importantly, explains how to install it. SEASR, or the Software Environment for the Advancement of Scholarly Research, is an extremely powerful, but also complex, collection of software tools for text and data analysis; getting it up and running is not for the feint of heart. While SEASR has documentation[107], reading the personal narrative of someone actually *using* the system is not only as informative, it is much more entertaining.

[105] Dumbrowski, Quinn. 2010. "Installing Cocoon on Ubuntu" http://quinndombrowski.com/blog/2010/04/14/installing-cocoon-on-ubuntu

[106] Mulligan, Ian. "Using SEASR's Meandre Workbench To Explore the Past, Part One: Overview" http://ianmilligan.ca/2012/06/27/using-seasr-to-explore-the-past-part-one-overview/

[107] See the SEASR documentation page here: http://www.seasr.org/documentation/

Blog posts in the para-academic genre, such as the “how to” themes, serve an extremely important function within the digital humanities community because they are a *public record* of technological dabbling, fooling around, goofing off, and other “ineffable” (Menzel 1968) experiments. Scholars document their successes and failures using tools in their research and this in turn helps others in the course of their own research. The “how to” post testifies to a culture of honesty, uncertainty, and openness about knowledge and information, but also satisfies, at least partially, the demand for technical training in the humanities.

Humanities Computing and the Skills Gap

Graduate training in the humanities does not typically teach the command line. Programming, scripting, and other information literacies have an uncertain place in the humanities curriculum. The skills gap in the humanities has been a challenge for the adoption of humanities computing style digital humanities into the mainstream academic discourse. While graduate training slowly accommodates this need,[108] digital humanities blogs have rushed in to fill the gap.

Until recently, blogs were the most effective source of information about topic modeling. There was little formally published literature on *using* topic modeling in 2012, despite a plethora of papers in computer science *about* topic modeling. However, starting in 2010 topic modeling became “a hot topic” on blogs in part due to often-cited posts by Mathew Jockers and Cameron Blevins.[109]

[108] For example, in a long post about teaching Shawn Graham describes a digital history course he teaches that introduces students to tools and project-centric, as opposed to paper centric, research. He also discusses the independent studies he advises for students who are able to conduct self-directed learning. What is important to reflect upon is the idiosyncratic the distribution of technical training in the humanities, which is reliant upon technically savvy faculty. http://electricarchaeology.ca/2012/03/29/a-teaching-philosophy-in-practice/

[109] Matthew Jockers, as you know, wrote the book on topic modeling [-jockers_macroanalysis_2013] and his blog has been one of the focal points for spreading awareness of topic modeling in the digital humanities. His first post in March 2010, topic modeling blog posts from the Day of DH, used MALLET to try and identify latent thematic relationships between bloggers who are not explicitly connected. http://www.matthewjockers.net/2010/03/19/whos-your-dh-blog-mate-match-making-the-day-

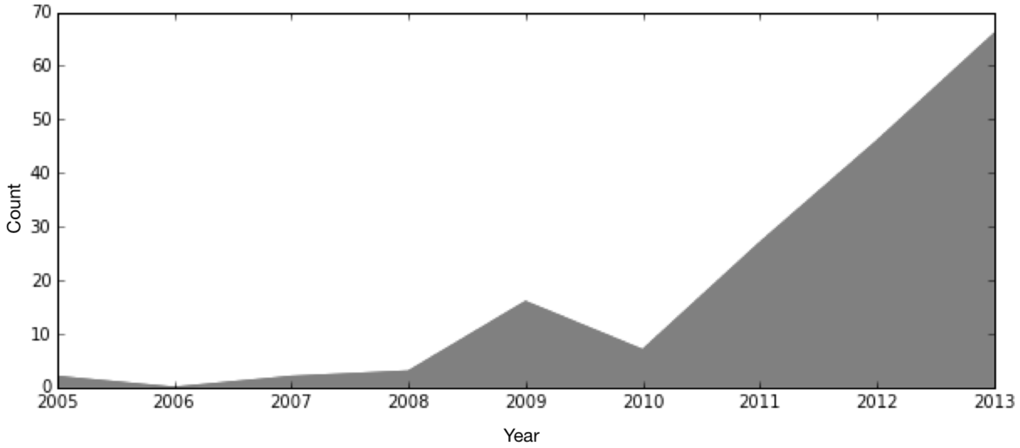


Figure (42). A crude count-per-year of the term "topic model" in the corpus

The graph in Figure (42) is a very crude approximation of the interest in topic modeling by counting occurrences of the string "topic model." The graph shows only a few instances of the term until a bump in 2009. From 2005 to 2008, the only blog using the term was the Natural Language Processing blog *http://nlpers.blogspot.com/*, a computational linguistics blog that was included in the *Compendium*, but isn't written for a digital humanities audience.[110] but the term didn't really take off until after 2010. From 2010 to 2013 occurrences of the bigram "topic model" linearly increase from seven in 2010, twenty-seven in 2011, forty-six in 2012, and sixty-seven in 2013.[111]

of-dh-bloggers-with-topic-modeling/ However, I'd argue topic modeling didn't *really* become popular until September of 2011, when Jockers wrote the extremely popular post, "The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors," which included an "imperfect attempt at making the mathematical magic of LDA palatable to the average humanist" by describing the generative process modeled by LDA using a metaphorical story rather than mathematical notation. This post had a significant impact on the popularity of topic modeling in the digital humanities. Never underestimate the power of a good explanation.

[110] The term "digital humanities" hasn't occurred on the blog a single time, however, the blog was included in the Compendium.

[111] I readily admit this is a very crude measure. Firstly, because I'm only searching for the string "topic model" and doing little normalization. The data, as I discussed in Chapter 4, is incomplete do I cannot report the exact number of occurrences. However, the sample of blog posts and frequency counts I present here is a good approximation of the increasing trend in writing about topic modeling.

Who are the Para-academics?

Para-academic discourse involves a larger constituency than just tenure-track academics writing about topic modeling. Many of the blogs represented in the para-academic genre are written not by credentialed academics, but librarians and technologists who support and *do* digital humanities research alongside faculty, students, and alt-academics.

Looking at the blogs with the most *para-academic* content shows how themes about certain types of academic work correlate with certain kinds of workers.

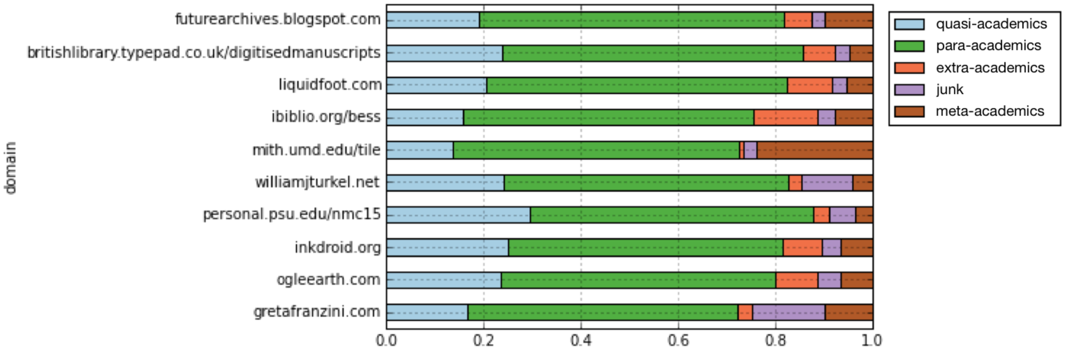


Figure (43). The top 10 blogs with highest proportion of para-academic topics.

Of the blogs with the most *para-academic* content, six belong to librarians or technologists, two belong to scholars, and two belong to library or DH center projects. Para-academic discourse, at least in the top 10, has far fewer tenure track scholars. Of the two scholars in the list, one was an archeologist at Penn State whose blog and web presence has vanished, and the other was William Turkel, an associate professor of history at the University of Western Ontario.[112]

[112] Turkel is not the typical historian. While he participates in history's monograph culture he better known in the community for his digital project *The Programming Historian*. The Programming Historian has been an extremely important text for introducing technical skills in the history discipline.History. The project is currently in its second edition. http://programminghistorian.org/

Extra-Academic

Extra academic themes highlight the most-informal genre of informal scholarly communication. These are the posts whose content would be *clearly* out of place in a journal article or most academic books.[113] These topics show scholars taking advantage of open publishing to write about nonacademic themes. Posts about cooking, restaurants, and food appear within the corpus alongside posts about literary theory, installing Linux, and network analysis. While some of this writing engaged these nonacademic themes with the style and tone of detached academic prose and analysis, other themes, especially religious, were deeply personal in nature. Five of the topics in the model honed in on writing about personal health, relationships, family, and other themes from deeply personal blog posts. Scholars can and do write about serious and intimate aspects of their personal lives. The presence of these themes is a glimmer of humanity within the emotionless and aloof tone of scholarly writing.

For example, topic 31 has the top words *people love life feel good thing make world things back*. Most of the top-ranked blog posts belong to danah boyd, whose blog was not only one of the earliest, but one of the most personal. Most of the personal posts are from the late 90s or the very early 2000s, before she established her identity as a public scholar. Her blog shifted towards a more formal and scholarly voice in tone over time as she increasingly used her blog as a venue for her scholarship.

[113] It is my perception, as a bookish scholar, that academics have a little bit more leeway with respect to formality in monographs. The subject matter, the publisher, and who is writing a monograph factor into formality's horizon of expectation. For example, anthropologist Heather Paxson wrote a book (University of CaliforniaUC Press),, about *The Life of Cheese,* but her treatment of cheese is quite scholarly.

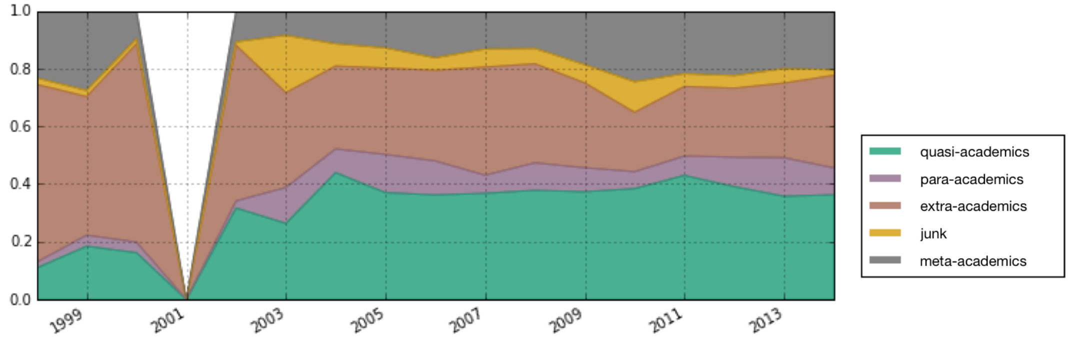


Figure (44). Proportions of different categories on danah boyd's blog per year. There is no data for 2001

Figure (44) shows boyd began writing increasingly more in the quasi-academic genre and less in the extra-academic starting in 2003. Although, it is important to note that boyd still continues to write about personal themes, which is a characteristic of her public identity. Figure (44) also shows an important aspect about the public records scholars leave behind. Their ideas, beliefs, and opinions are (or should be) always changing. Blogs are really a record of ideas, beliefs, and opinions they may no longer hold. Scholars are vectors through space and time; constantly learning, developing, and shifting academic directions.

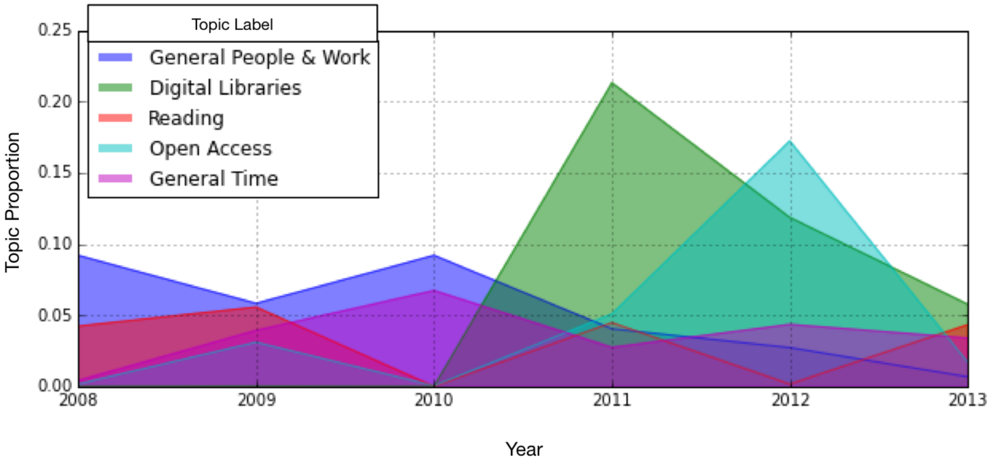


Figure (45). Proportions of topics from amandafrench.net from 2008 to 2013. The proportions show shifting interest and attention over time.

Figure (45) shows the top five topics for *amandafrench.net*. French is an alternative-academic scholar who has held several different positions from 2008 till 2013. The rise and fall of particular topics reflects different stages of her academic career and interests. Notice the spike in the topics on “Digital Libraries” and “Open Access” in 2011; this coincides with her position as a Research Assistant Professor and THATCamp coordinator at George Mason from 2010 to 2014.[114]

The *extra-academic* genre of scholarly communication highlights the humanity of scholars, not only though personal writing or informal tone, but also in how the themes and subject matter *change*. Blogs, especially old ones, are an accumulation of public writing over time showing change, growth, progress, regress, and maybe even regret. This raises questions about accountability when extra-academic writing collides with the systems of hiring, firing, promotion, and tenure.

Summary of Findings

The findings articulated in this chapter came from three main forms of analysis, two quantitative and one qualitative, but all mixed to various degrees. With respect to the theory of methods and ways of knowing, these findings emerge from a trace ethnography that used several scalar devices to “compress“ the enormous discursive space into a more manageable collection of traces. This process includes the data preparation, the topic model, the measure of entropy and hierarchical clustering. Each provides a unique perspective of digital humanities blogs, foregrounding certain features (i.e., clusters of English topics) at the expensive of others (i.e., the diffusions of digital history themes within other clusters).

[114] French started her blog and website as a space for posting information about her professional life. There is a wealth of information, beyond the blog, on or linked on her website. It would be interesting to dig into web archives and see how scholar’s representation of their (professional) self changes over time. http://amandafrench.net/resume/

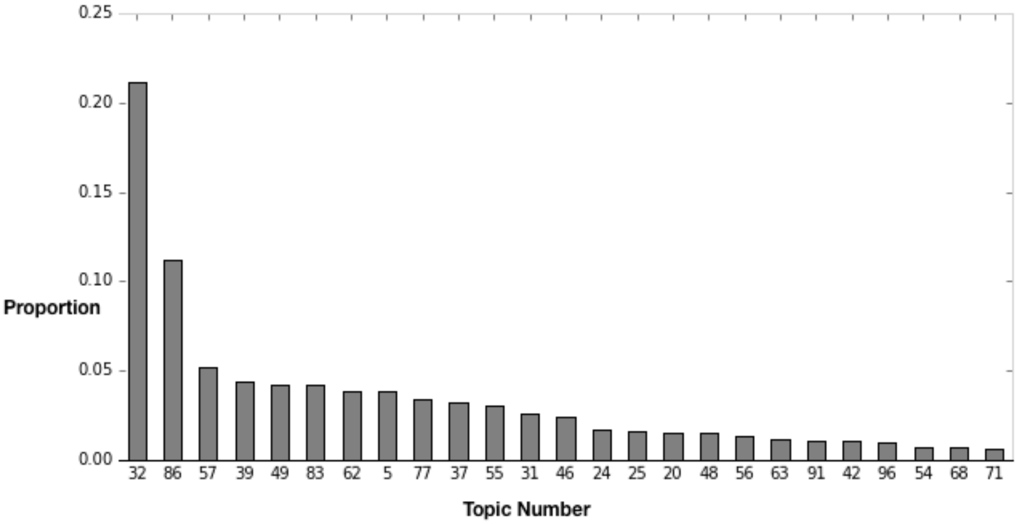


Figure (46). An example of a blog with average topic diversity.

The analysis of entropy provided a measure of the diversity of topics per blog. While digital humanities bloggers as a whole write about a variety of topics, as Figure (46) shows, at an individual level they typically write mostly about a single topic, but touch upon a long tail of other topics. What is not determined in this analysis is the degree to which topic expression is correlated. If a digital humanities blogger writes about topic X, does topic Y have a higher likelihood of expression? To what extent to blogs themselves cluster around a set of themes? If anything this chapter shows there is a wealth of questions and answers in these data.

The analysis of topic clusters begins to explore the computationally generated topics using a distance measure and hierarchical clustering to produce a dendrogram of topic's relationships.[115] The whole tree (Appendix B) shows the entire landscape of a hundred topics generated by MALLET. Using qualitative coding, labels were applied to the clusters based upon the top keywords: Technology, Digital, English, Libraries & Meta, Miscellaneous, and Non-English.

[115] There are extensions to MALLET to support hierarchical topic models (Blei 2003), but those haven't been as extensively used as vanilla LDA. How to document and interpret the model, the results, and the significance using trace ethnography are opportunities for future work.

The technology cluster has two subgroups. The first gathers deep discussions of technological subject matter like configuring Linux or installing and using text analysis tools. The second cluster brings together methodological discussions around data, quantitative analysis, and machine learning. The digital cluster is less semantically compact than the technology cluster. Digital topics focus on technology as an object of study, teaching with digital technology, digital writing, and the important "What is DH" topic. The English cluster features topics relevant to the subject matter of English as a discipline, such as British literature, storytelling, and narrative. The prominence of the English cluster is an indicator of English's dominant presence in the community (despite many arguments that DH is just as much about history as it is about English). However, the significance of libraries in the digital humanities should not be ignored; those discourses were strong enough to form their own cluster (which is interestingly adjacent to the meta/administrative cluster). The library cluster features deeply technical discussions of digital libraries, copyright, and open access alongside more administrative themes like job postings and general talk about the academy. The miscellaneous cluster captures a variety of topics ranging from personal politics and food to high theory and knowledge work. This cluster collects many of the smaller themes distributed through the digital humanities blogosphere.[116] Non-English topics are accurately relegated to their own outlying branch of the tree. These topics indicate that while the community is mostly Anglo-centric, there are spaces in the community for non-English writers.

The four categories of informal scholarly communication, *quasi-academic, meta-academic, para-academic, and extra-academic* come out of an ethnographic analysis of traces. Quasi-academic discussions evoke classic humanities themes and subject matter. Meta-academic discussions focus on the maintenance and administration of the digital humanities as a community. Para-academic discussions address themes that are vital to the digital humanities, but don't (yet)

[116] Digging into these topics is an area for other researchers interested in exploring how other themes and subject matter are discussed and represented on digital humanities blogs.

have a place in disciplinary and formally published venues. Finally, extra-academic themes show how blogs enable scholars to write about anything.

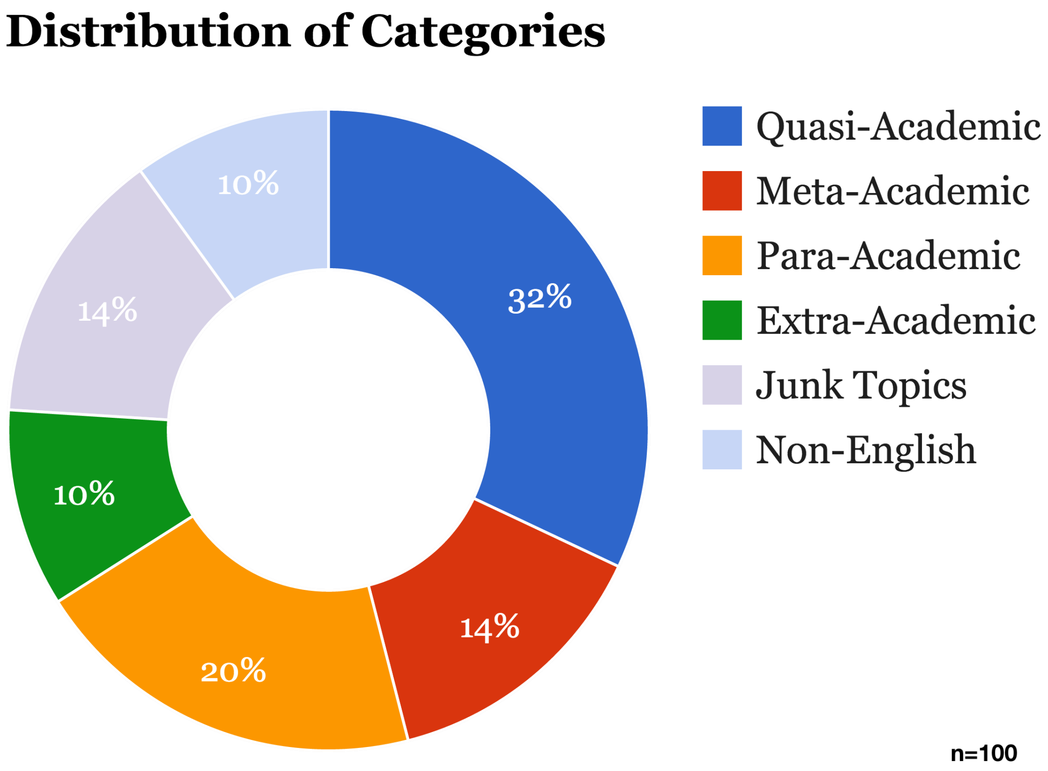


Figure (47): The distribution of categories over 100 topics. Each topic was assigned to one of the four categories of informal scholarly communication or labeled as junk or non-English.

The chart in figure (47) shows how the four categories of informal scholarly communication are distributed across the 100 topic model. Each topic was assigned to one of the four categories of informal scholarly communication based upon a coding of the keywords and top ranked documents. This is a recursive mapping of the analytical themes and categories, derived from a content analysis back onto the topics, i.e. computational representation of the corpus. That is to say, it is an approximation of how the four categories are distributed across the corpus by seeing how they are distributed across the 100 topics of the model.

Quasi-academic discourses are the largest portion, which is an indicator that digital humanities scholars focus most of their writing on academic themes. While the analysis above shows this category is not entirely composed of new ideas, the size of this category shows that scholars are using blogs as a space to do academic work in some capacity. The second largest category is para-academic, which is especially important for the digital humanities because this means blogs are supporting a significantly large discourse around themes that don't necessarily have a natural outlet in formal modes of scholarly communication. Writing about technical subject matter, how to use tools, or supplementary details of computational research is not a small or isolated interest of a small group, it is a significant chunk of how digital humanities scholars use blogs. The number of meta-academic topics reveals the management and maintenance of the community is a tertiary, but still important, function of blogs. The extra-academic topics constitute a small, yet present, portion of the whole. What is missing from this synchronic representation is how the four categories are distributed across the corpus over time. As was shown by looking at the temporal allocation of the four categories in danah boyd's blog was an increase in quasi-academic and decrease in extra-academic writing, might the same be true across the entire corpus?

Clearly the data has many more stories to tell and only a few have been addressed in this chapter. The data-driven and empirical representations of the digital humanities put forth here may be somewhat alien to members of the community, which is why it is important to emphasize that these are *representations* to be interpreted and not declarations of authority or truth. Scalar devices, as a theoretical construct, are useful for understanding this nuance. Topic modeling is a form of *lossy* compression, like an mp3, whereby certain features of a phenomenon are emphasized, while others are deemphasized.[117] However, like an mp3, just because some information is lost doesn't mean the resulting object is not useful or insightful.

[117] For an amazing infrastructural inversion of the mp3 compression algorithm watch Ryan Maguire's videos of what gets removed: http://www.theverge.com/2015/2/19/8068923/mp3-compression-ghost-suzanne-vega-toms-diner

Chapter Six

Discussion & Conclusion

This chapter brings together a wide variety of open threads and moving parts that characterize the research project. The chapter situates the findings from the previous chapters into three analytical clusters: *methodological*, *empirical*, and *conceptual*. The methodological findings concern the dynamics of transforming the web into data and ethnographic ways of interpretation that are augmented with computation. The empirical findings concern what has been learned about the digital humanities blogging. The most recent manifestation of the community grew with the maturation of blogging. Prominent exemplars, such as Dan Cohen's influential blog, showed the digital humanities community new ways to use the medium. The most significant subject of this chapter is the conceptual contribution of the *(in)visible college*. An (in)visible college is a descriptive model for *informal scholarly communication* that is *publically visible* by virtue of its existence on the Open Web.

The methodological, empirical, conceptual findings have implications for the study of infrastructures, the digital humanities, and scholarly communication. Trace ethnography at scale provides a way of knowing large-scale, infrastructural phenomenon that is in theoretical harmony with the traditional qualitative and interpretive methods typically used. This is the first systematic and data-driven study of the digital humanities and as such empirically reveals the community's diversity, especially in libraries, and the importance of informal communication in the creation and ongoing maintenance of the community. Informal scholarly communication is understudied at scale because of practical issues with data and access. Blogs and the Open Web are a fundamentally new space for scholarly communication, an (in)visible college, that is now possible to explore, to study, and participate in.

The collection, analysis, and presentation of findings have been guided by the following research question:

What roles are played by sociotechnical infrastructure, as represented by scholarly blogs, in facilitating scholarly communication in the digital humanities?

This high level question was broken into three sub-questions:

- *What are the methodological dynamics and tensions in generating data from the Open Web?*
- *What themes are digital humanities scholars writing about on their blogs?*
- *How does scholarly blogging expand current understandings of informal scholarly communication?*

Chapter 2 and Chapter 3 reviewed the literature on scholarly communication, digital humanities, and infrastructure studies and set up the theoretical and methodological space for trace ethnography of scaling. Chapter 4 ethnographically documented the data collection, cleaning, and preparation for topic modeling to be used as a *scalar device*. Appendix A further documents the topic modeling algorithm. These *thick descriptions* of data preparation and of the unsupervised learning algorithm LDA are a necessary requirement for trace ethnographic analysis at scale because they reveal what aspects of the large-scale phenomena are lost in the scaling processes. Chapter 5 analyzes the topic model both quantitatively and qualitatively, calculating the diversity of topics per blog, clustering topics, and categorizing the kinds of informal scholarly communication.

Blogs, Infrastructure, & Scholarly Communication

How do blogs as infrastructure inform our understanding of scholarly communication?

This question betrays an assumption about the nature of blogs. In the initial formulation of the project, blogs were characterized as infrastructure and it was their infrastructural dynamics that were subject to investigation. While blogs may be an infrastructure for scholarly communication, the infrastructural dynamics are difficult to distinguish from the Open Web.

I argue blogs *are* infrastructure as evidenced by considering Star's (1999b) nine infrastructural properties:

- *Embeddedness:* Blogs are intimately tied up with the Open Web, which is in turn sunk into the lower level communications networks of the Internet.
- *Transparency:* Platforms like WordPress and Blogger make the process of hosting blogs and authoring posts easy enough for the non-technical members of the digital humanities community.
- *Reach or scope:* Digital humanities blogs extend both spatially and temporally across the globe. They are distributed and managed across individuals and institutions.
- *Learned as a part of membership:* Blogging has been normalized within the community. The fear of being identified as a blogger has passed and now it could be argued you *have* to have a blog to be a digital humanist (it
- *Links with conventions of practice:* One of the defining characteristics of the digital humanities is an ethos of *openness*. Blogs are by definition open access and embody those ideals of openness, both in the practical meaning of access, but also in a larger spirit of openness related to reaching non-academic audiences like the "public."
- *Embodiment of standards:* Humanities computing has a tradition of informal communication, as evidenced by the long history of the

HUMANIST mailing list. Informal channels have always been important because formal channels of communication can't accommodate multi-modal forms of scholarly expression or the alien epistemologies of computational research. Blogs, as evidenced by the para-academic category, enable these forms of expression by plugging into the expressive capacities afforded by open standards on the Open Web.

- *Built on an installed base:* Digital humanities blogging is possible because blogging is a generalized practice on the Open Web. It leverages existing platforms, like WordPress and Blogger, but it also initially inherited the expectations of blogging as an informal and unruly discursive space with little to no epistemic value. The extra-academic category shows that installed base does exist, but it is a small percentage in comparison to the other three categories of informal scholarly communication.
- *Becomes visible upon breakdown:* The work of maintaining a blog is invisible to readers, until something breaks. Topics 8 and 18 of the model included keywords like *cialis* and l*evitra*, mixed in with words about books and literary history. Several of the blogs in the sample had been hacked filled the posts with links to virility drugs. Managing a blog with comments on the Open Web means constantly filtering out spam; one person's invisible infrastructure is another person's daily work.
- *Is fixed in modular increments, not all at once or globally:* Being hacked is often the result of an out-of-date WordPress instance. Blog owners must update their self-hosted blogging platforms, digital humanities blogs are not centrally managed by the community. Control is highly distributed across individual scholars, institutions, and commercial companies. This is the flip side of Open Publishing, anyone is free to create their own discursive space, but it they are responsible for management and repair.

In the context of this project, I argue for blogs as an infrastructure for informal communication in the digital humanities, but this is a soft-argument. Scholarly blogs or blogs more generally may or may not be an infrastructure that is readily distinguishable from the Open Web. The nine infrastructural properties of blogs enumerated above are mixed together with the Open Web. This raises a question, if blogs are not their own unique infrastructure, then what are they?

Are blogs a *platform?* The emerging discipline of *platform studies* (Bogost & Montfort 2009) might provide a useful lens by which to unpack this quandary. According to the official website, "platform Studies investigates the relationships between the hardware and software design of computing systems and the creative works produced on those systems."[118] Platform studies is related to what Sandvig (2013) calls the "new materialist" approach to studying infrastructures. It has direct connections to new media studies and the digital humanities. Such an approach might focus more on WordPress and Blogger.com, colloquially called "platforms," which are a material manifestation of blogs in software.

Are blogs a *format* or a *convention?* Blogs are composed from open standards, HTML and RSS, but there is no W3C standard for blogs. They are a *convention of practice,* which is an infrastructural property, which illuminates the sociotechnical dynamics. Yet, considering blogs as a convention doesn't readily provide a theoretical scaffolding or lens with which to examine blogs as a conceptual object.

Are blogs a *genre?* I hesitate to use the term genre because it invokes the wrath of literary scholars. I am not equipped to fully situate blogs within the breadth and depth of genre studies, but as a category of creative expression, I would argue blogs are not a single genre. Rettberg (2013) identified three styles of blogs, personal blogs, link blogs, and topic-centric blogs. My analysis shows that scholarly blogs exhibit four different categories of communication. Genre is not specific enough as an analytical category and it doesn't highlight the technical and material facets of blogs.

So what are blogs? Platform studies, genre studies, and other theoretical frameworks from new media studies can and should be brought together with infrastructure studies to understand blogs and reconcile the similarities and

[118] http://platformstudies.com/

differences between these various frameworks. This project leveraged the theoretical and conceptual frame of infrastructure to motivate an inquiry into digital humanities blogs and specifically address the issues of scale.

Informal Scholarly Communication at Scale

Informal scholarly communication is a discursive space that previously has not been mapped at any significant scale. The modes of access and lack of public documentary residue have made it challenging for methods other than the labor intensive, yet rich and contextual, ethnographic or other micro-sociological qualitative modes of inquiry. The world of informal scholarly communication has been popularly conceptualized for half-century as “invisible colleges” and for that time has remained methodologically out of reach for researchers (Lievrouw 1988; Lievrouw 1989). The invisible college is not an adequate descriptive metaphor for the kinds of scholarly communication I studied because, to put it simply, I was able to *see* it. Even more importantly, automated web scrapers and information retrieval systems *see it at scale*.

The community that created the 39,976 blog posts I assembled do not constitute an “invisible” college; their work is posted on the Open Web and retrievable because of the technological affordances of open access. Using Diffbot’s infrastructure and APIs I was able to index every page on 396 web sites, identify individual blog posts, and extract their text content. This is a kind of human and computation visibility that makes possible both quantitative and qualitative analysis. So, blogs are an infrastructure that affords a *technical visibility*.

However, blogs as a mode of scholarly communication are not the same as books or journals. Most models of scholarly communication are focused on the formal production cycles of scholarly literature. Borgman’s model of scholarly communication posits three functions, *legitimization*, *dissemination*, and *preservation*. While echoes of these three functions exist in the digital humanities blogosphere, her model, based on the dynamics of book and periodical publishing, does not fully capture the sociotechnical dynamics at play on blogs. Books and journals *structurally* delineate scholarship that has quality. Scholarship in the form of a journal article has been legitimized and peer reviewed to signal its value. When a scholar publishes a book or journal article, the format implies a level of formality and distinction.

The format of blogs may seem to signal the opposite condition: that content is not scholarly or valuable. Articles like the *Chronicle of Higher Education's* "Bloggers need not apply" instill a sense of danger with respect to scholarly blogging. This pushed the perception of blogs deep into territory of informal communication. Yet, the topic model developed for this dissertation revealed deeply scholarly content posted on digital humanities blogs. The clusters of English topics or the topics unpacking computational methodologies show, from a content perspective, scholarly and potentially valuable discourses that may or may not make their way into the formal sphere via books or journal articles.

However, not all content on digital humanities blogs is scholarly. The digital humanities blogosphere is not simply a preprint circulation network; it is more complicated. Bloggers write about a variety of subjects, most of which are "scholarly" to some degree, but are not all directed towards formal publication cycles. Some of this content is posted for the purposes of sharing experiences or information, other forms of content are oriented towards engaging a nonacademic audience. All of these forms of writing circulate with no structural distinction in the digital humanities blogosphere.

Studying Blogs and the Open Web

- *What are the methodological dynamics and tensions in generating data from the Open Web?*

When studying phenomena as infrastructure it is important to not privilege the social over the technical, or conversely, the technical over the social. Latour's Actor Network Theory and his *principle of generalized symmetry* posit that human and non-human actors should be considered as equal agents within a conceptual frame. This project considered an array of documentary traces, some generated by human actors, others generated by computational actors, as a mixture through which neither set of traces privileges the other. All trace data are sociotechnical.

The principle of generalized symmetry renders the discussion in this dissertation to be distinct from social computing or computational social science, where findings inform general principles of social behavior online. My research dives into technical minutia and social context, surfacing unique aspects of the digital humanities, instead of generalized scholarly bloggers. I am not an objective observer; I am web developer and an active participant in the digital humanities community. By understanding both the social and technical contexts of the community of study, I invoke a *sociotechnical* mode of interpretation that calls attention to certain aspects of the complex traces being studied. This is the ethnographic sentiment that resists the "scientizing" and totalizing tendencies of quantitative researchers and the way they interpret results and make knowledge claims about online phenomena. This research is operating under an *interpretivist* social science tradition, but one that uses quantitative and computational methods. While some critics might disagree, such methods do not inherently "scientize" phenomena. Numbers do not inherently yield truth; it is people wielding numbers who make truth claims.

Ground Truthiness

The field of machine learning invokes the notion of "ground truth," in which the accuracy of a statistical model or supervised learning algorithm is judged in relationship to objectively known data. For example, the performance of a supervised classifier can be evaluated by testing to see how well it categorizes information whose assignments are previously known. The already known categorical designations are called the "ground truth." This research problematizes the notion that there can be an "objective" ground truth on the web.

An example: The process of date cleaning the dataset revealed a blog, *discontents.com.au,* with posts from November 15, 1985. The dates are the ground truth as far as the infrastructures were concerned. Yet, the way in which this writing is represented, as blog posts written in the mid-1980s, cannot possibly be true. To what extent can web servers lie? What other pieces of information or metadata in my corpus are not as they seem? This was not simply a matter of incorrect date extraction by the algorithms; this is a "ground truth" that should not be taken at face value. Hence, I would argue the data I encounter on the web has no objective truth, but rather a "ground truthiness"[119] rooted in a performance of technical representations, but requiring further investigation into the infrastructures supporting that technical performance; that is the hosted instance of WordPress, the technical history and provenance of data in the database, and the social history of the person, institutions, and companies responsible for setting up and configuring the infrastructure.

This perspective on the nature of blog data has methodological implications for researchers who treat the web as a repository of "found data" and attempt to make claims about individuals, groups, or society. All significance and meaning is local and purely quantitative; computational techniques for trying to understand the meaning of the web will always fall short because they fail to enter into the localized spaces of membership and sociality, accomplished through close reading, ethnography, and interpersonal qualitative techniques. Yet, qualitative and interpretive methods cannot possibly scale to the level of the web, so what is a researcher supposed to do?

Ground Truthiness highlights one of the properties of scholarly communication in an (in)visible college. Information in an (in)visible college has not been validated or legitimized, it can be "published" without review *and* as the case of *discontents.com.au* shows, the infrastructure and the structural forms themselves can be manipulated at the database layer. Trust in the legitimacy of

[119] Credit to comedian Steven Colbert for coining the term "truthiness" to characterize the kinds of intuitive or illogical "truths" often wielded by politicians.

the content cannot be delegated to the structural forms (e.g., peer reviewed journals) because those structural forms are plastic. While this study focused on the text content of posts, the design, the structural, and the interconnections to other resources constitute a networked mode of expression that can, depending on the platform and its owners, be considered a kind of discourse.

Trace Ethnography of Scaling

This project assembled various theoretical, methodological, technological, and descriptive elements to conduct *trace ethnography of scaling*. Trace ethnography at scale rests upon three sociotechnical assumptions. First, the achievement of co-presence, rather than co-location, is a sufficient basis for ethnographic inquiry (Beaulieu 2010). This move opens up a theoretical space for conducting ethnographic inquiry online. Second, the residual documentary traces of online activity are first-order mechanisms by which users interact and establish social order (Geiger and Ribes 2011). This means that traces, even when accessed after-the-fact as historical documents, can be subjected to ethnographic inquiry. Third, the problem of scale can be overcome through the careful deployment of *scalar devices*, that is, techniques and tools that reconfigure the object of investigation by generating a "smaller" representation of the phenomena. However, when deploying a scalar device it is absolutely *crucial* to recognize which dimensions of the phenomena are foregrounded or backgrounded by the scaling process. This recognition is achieved by thoroughly implementing and documenting topic modeling as a scalar device.

This project treats digital humanities blogs as documentary traces of the community's informal scholarly communications. The architecture of the Open Web and of blogging platforms means the history of a scholar's writing is *observable* as an investigator. Treating blogs as traces makes them available to trace ethnography inquiry, but also introduces problems of scale. To examine the full history of all the blogs listed in the *Compendium of Digital Humanities*, which at the time listed nearly six hundred blogs, means embarking on a long and complicated process of identifying and indexing blog posts from semi-structured websites. Chapter 4 is a thick description of the five major steps in transforming blog posts into data. Only after the five stages of data enumeration, classification, extraction, cleaning, and transformation are the data “raw” enough for computational analysis. Each of these five steps involves saving some pieces of information and throwing out others. Raw data is an oxymoron (Gitelman 2013).

Understanding at scale is at the heart of this project’s methodological contribution. How can we *know* a large-scale phenomenon? How is that knowledge is achieved? What, in the particular achievement of that knowledge, becomes un-knowable? With digital humanities blogs, the text content of the blog posts at the expense of comments or links. Including comments as structured metadata *at scale* involves a considerable amount extra of work beyond that needed to prepare the text posts proper. In the interviews within the blogging community, respondents indicated they either turn off or ignore blog comments because most are spam. Comments on digital humanities blogs are not large-scale phenomena because extensive comment threads are rare. Comments never the less are extremely important and informative to bloggers. The knowledge to be gleaned from comments is best achieved via close reading and content analysis, rather than through text mining.

An analysis of embedded links in blog posts may provide valuable knowledge about digital humanities blogging but text extraction strips out the HTML markup, including links. This is a moment where the choice of scalar device plays an important role in what we can or can’t know about a large-scale phenomenon.

The Blogging Humanities

The Digital humanities love to blog and blogging has now become part of the community's discursive practice. The 396 blogs included in this study are only a sample of the community. Digital Humanities Now reports the community is posting over five thousand posts per month.[120] Studying an actively growing phenomenon means any analysis is immediately out-of-date.

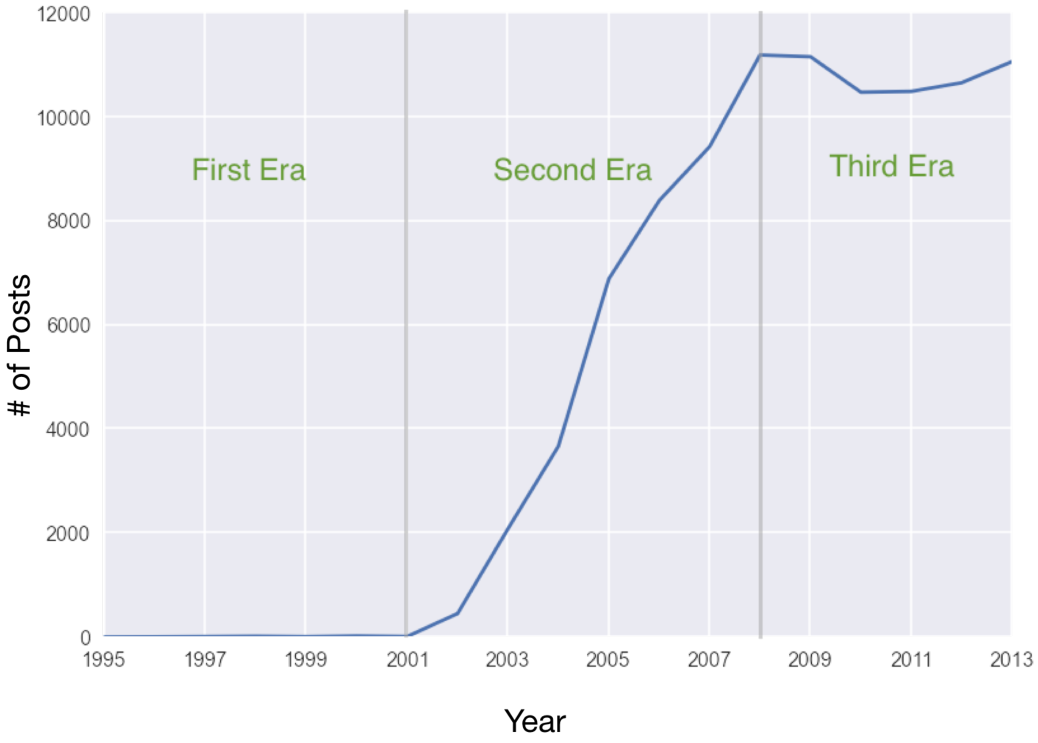


Figure (47). Annotated from Chapter 4, total blog posts per year divided into the three eras of DH Blogging

The graph in Figure (47) shows the *birth of an (in)visible college*, reaffirming the fundamental tenant that these networks of scholarly communication are *technically visible*. The graph shows the number of blog posts per year, as reported in Chapter 4, incorporating the dates of all 106,804 posts in the original dataset, not just the 39,976 posts included in the topic model. The growth of the

[120] https://twitter.com/westcottchnm/status/542757755019603969

digital humanities blogging community from hovering around zero to a sustained level of eleven thousand per month shows the scale and emergence of this (in)visible college is not a trivial matter. The emergence has three distinct periods of growth or "eras." Each era has a different dynamic and relationship to blogging and the Open Web in general.

The first era from the late '90s until 2001 had very little activity; very few scholars, were writing during this period. The analysis in Chapter 5 demonstrated the fluid intermingling of para- and extra-academic writing in this vanguard population. In this early period of proto-blogging on the broader Open Web, the format was dominated by personal narrative and documenting lived experience (Rodzvilla 2002; Serfaty 2004; Bruns and Jacobs 2006; Dean 2010; Rosenberg 2010). The early scholarly bloggers in this dataset, like danah boyd, were blogging about personal themes, *extra-academic* scholarly communication, mixing reflections from inside and outside academia. In this era blogging hadn't actually become a *thing*, people just posted their writing on the Open Web.

This era overlaps with the finals days of what Cathy Davidson calls "Humanities 1.0," that is, the period when computation was merely a tool for research and analysis. However, as the ethos of technology-as-tool faded, the first era of blogging ushered in a period of innovation and experimentation at the intersection of the humanities and computation for communication. One example of this experimentation is Alan Liu's *Voice of Shuttle,* a massive collection of links to humanities resources from across the web started in 1995.[121] *Voice of Shuttle* was a link blog and a social bookmarking site before there were link blogs and social bookmarking sites.

The second era is the "eternal September" of DH blogging.[122] During this period, the practice skyrocketed from almost nothing in 2001 to around eleven thousand

[121] The website, and now database, is still live although many of the links are broken due to link-rot. I wonder how much of VoS we can recover from web archives. http://vos.ucsb.edu/about.asp

[122] Credit to Bethany Nowviskie who explores the pairing of DH with the eternal September in a blog post. I am using the idea mainly to highlight growth and new membership in the community. http://nowviskie.org/2010/eternal-september-of-the-digital-humanities/

posts per year in 2008. During this period, blogging was not only popular in the digital humanities, but it also gained attention of the broader Internet community. Web 2.0 emerged in this era and the number of blog posts by scholars in the *Compendium* increased year by year. WordPress, a favored platform for digital humanities blogging, launched in 2005; Dan Cohen's often cited post, "Professors, start your blogs," was published in 2006. Cohen started blogging in 2005, when he was an assistant professor at George Mason University. When asked about why he started blogging, he stressed the advantages of early adoption:

> I really started to see some good academic blogging and felt I was in a field (DH) where this could be a great way to get ideas out and have an influence. (Dan Cohen interview)

As blogging entered the mainstream, scholars and other academics began using blogs as a platform for *scholarly communication*. In turn, digital humanists began exploring the possibilities of platform and the format as a mode of *scholarly* and *professional* communication. Lisa Spiro, another early and influential digital humanities blogger writing about themes of digital scholarship, cites Cohen's influence in getting her to start blogging:

> I was inspired to blog by people like Dan Cohen, who both made a good case for the importance of academics engaging in conversations online and whose own work contributed so much to my own thinking and learning. Initially I planned to blog the process of transforming my dissertation (completed in 2002) into a work of digital scholarship, but I got diverted by other interests. (Lisa Spiro interview)

Jason Heppler, a historian and academic technology specialist at Stanford, also cites Cohen's blog as an early influence, but also notes the transformation of Caleb McDaniel's blog from personal to a more professional digital humanities-oriented blog:

> I'd have to say the bloggers I've been reading the longest are Dan Cohen and Caleb McDaniel. … I learned about Dan years ago when I started graduate school as I was introduced to the great work going on at RRCHNM. Reading Dan's posts gave me a great window into what DH could be. I started reading Caleb's blog for a different reason. He was still writing on his previous blog *Mode for Caleb* and, if I recall correctly, wasn't doing much with DH at the time. The blog was his space away from his dissertation to write about whatever struck him. The pieces I recall the most are those about jazz music, an interest him and I share. (Jason Heppler interview)

The work of these early digital humanities bloggers set a new tone for blogs, reshaping the community's expectations of the format from personal to professional narratives. This was part of the additive shift that contributed to the *growth* of digital humanities blogging. Cohen and Spiro are exemplars of scholars who showed others how blogging could be used in a more *scholarly* fashion. In this era, scholars in the digital humanities (and beyond) were doing more than writing on the web—blogging had become a *thing*.

The third era, beginning in 2008, is when the growth of digital humanities blogging plateaus at approximately eleven thousand posts per month. The data shows a very rapid decline in rate of growth of blog posts per year. There are two possible explanations for this leveling off. First, 2009 is the year when the *Compendium of Digital Humanities*, the list used as a sample of DH blogs in the research, was compiled. The period leading up to the launch of Digital Humanities Now 2.0 (discussed below) was the only time when a substantial amount of work went into adding blogs and RSS feeds to the list. After 2009, blogs were added to the list, but no exhaustive search of new blogs was conducted. In June of 2012, the *Compendium* contained 509 entries; in December of 2014 the *Compendium* contained 705 entries. The *Compendium* list sampled contained 615 entries, although not all of those were usable blogs. So while no effort was made to actively find and add new digital humanities blogs to the *Compendium*, the list was still passively maintained. Given the *Compendium* continued to grow (by one hundred blogs per year) the shift from active to passive additions cannot completely explain the sharp downturn in the number of blogs per year in 2008. It would seem that by 2008 digital humanities blogging reached a saturation point. The community is not very big and by 2008 a bulk of scholars interested in blogging had started their blogs.

A second possible explanation for the leveling off of blog posts after 2008 is the rapid mainstreaming of the social web, including Facebook, Twitter, and Tumblr. The plateau seen in the third era from 2008 to 2013 could be explained by scholars moving to new and better platforms for publishing certain kinds of content online, sharing links, or offering short commentary on current events.

Why have digital humanists taken to blogs? This analysis and trace of the historical trajectory provide modest and decidedly incomplete insight into the answer. A potentially significant factor in the community's adoption of blogs has to do with timing. The digital humanities, as a distinct brand from humanities computing, started in 2004. This was also the time when "Web 2.0," the technological and cultural shift towards dynamic and user-generated content on the web, was popularized. The emergence of these two sociotechnical assemblages, digital humanities and Web 2.0, are not isolated from each other (although, I should point out the direction of influence is unilateral). We can see the influence of Web 2.0 ideas in Cathy Davidson's characterization of digital humanities as "humanities 2.0."

The interview data confirms the speculations of Kirschenbaum, Davidson, and McPherson. Digital humanities, if nothing else, is about using technology to communicate online and blogs are one of the main digital spaces were that communication occurs. The digital humanities as we know it today would not exist without blogs. The particular infrastructural affordances of blogs, as a native format of the Open Web, coupled with a sociotechnical enthusiasm for writing on the web, gave rise to a vibrant community of informal scholarly communication.

(in)visible colleges

- *How does scholarly blogging expand current understandings of informal scholarly communication?*

Given the dynamic and ever-changing nature of the Open Web, the models and conceptualizations of scholarly communications on the Open Web (Cronin et al. 1998; Kling, McKim, and King 2003; Ginsparg 2004) are out of date or out of touch with the broader ecosystems of interaction and communication. The impact of web 2.0, and now the social web, upon scholarly communication has not been fully accounted for in existing models. Furthermore, the models of informal scholarly communication are focused exclusively upon the relation to formal publication, that is, preprint circulation, peer review, dissemination, and preservation. The infrastructures and processes associated with formal scholarly communication have not, despite best efforts, changed in any significant way with the advent of the web. While some of the platforms and modes of access are new, the underlying process is still the same (and that is a good thing!).

What has changed with the Open Web is the emergence of public, yet *informal* channels of scholarly communication. Open standards, open access, and open publishing have impacted the ways in which scholars informally share data and information. The dominant model of these informal interaction networks, the invisible college, is a purely theoretical construct which, for reasons of data availability and research methodology, could only hypothesize about *latent* networks as a counterpoint to the highly observable and measurable formal communication strategies. The kinds of informal scholarly communication characterized by the invisible college were, by definition, invisible and only available via close ethnographic other micro-qualitative research (Lievrouw 1989). Blogs and the Open Web constitute a fundamentally new discursive space; one that is *observable* in ways the invisible college was not.

What is an (in)visible college?

An (in)visible college is an original conceptual model developed for this dissertation to describe and explain the kinds of scholarly communication that occur on blogs and the Open Web. It is a formulation of what Halavais, drawing on Oldenberg, called a “third place” (2006), “a space for developing the social networks that help drive the more visible institutions of research (117).” Blogs, as a native format of the Open Web are technical *visible*, but as Halavais points out the “blogosphere provides its own intrinsic reputational rewards, but these may not extend to the wider academic (and other) contexts in which scholars work and are valued (123).” This tension around credit, reward, and value is where blogs are simultaneously visible and invisible, hence (in)visible. Blogs, because they are not part of the legitimization processes of peer review, live just below the surface of the visible and formal networks of scholarly publishing.

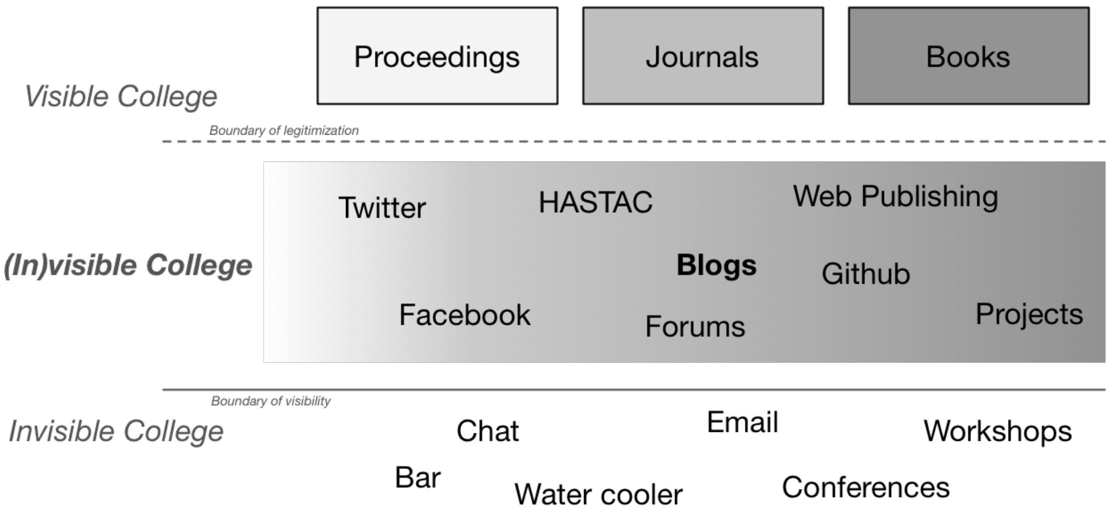


Figure (48). The (in)visible college situated within the various modalities of scholarly communication

Figure (48) represents the (in)visible college as the discursive space between the visible and invisible college. An (in)visible college has two important aspects:

- They are *informal* channels of scholarly communication.
- That is *visible at scale.*

The notion of the *(in)visible college* is specifically a concept for understanding informal scholarly communication on the Open Web. An (in)visible college is *visible* in the technical, functional, and pragmatic sense, but it is *invisible* and informal in the legitimated and peer-reviewed sense. Because of the technical affordances of the Open Web, specifically *open publishing*, blogs are considered an *informal* channel of communication. Scholar's ability to write about anything, both personal and professional, without editorial review or control, means there is no structural affordance with respect to the format itself to distinguish personal vs. professional, scholarly vs. colloquial, good vs. bad.

Informal Scholarly Communication

Discourse in an (in)visible college is informal and that means the expectations of content and form are loose. In theory, digital humanists can write about anything on their blogs. And they do! The topic model revealed, as a collective, digital humanists write about subject matter ranging from deep discussions about methodology (as seen in the technology cluster in Chapter 5) to engaging “the digital“ (as seen the cluster of topics about digital research), to matters of a personal health and well being (discourses in the extra-academic category of scholarly communication).

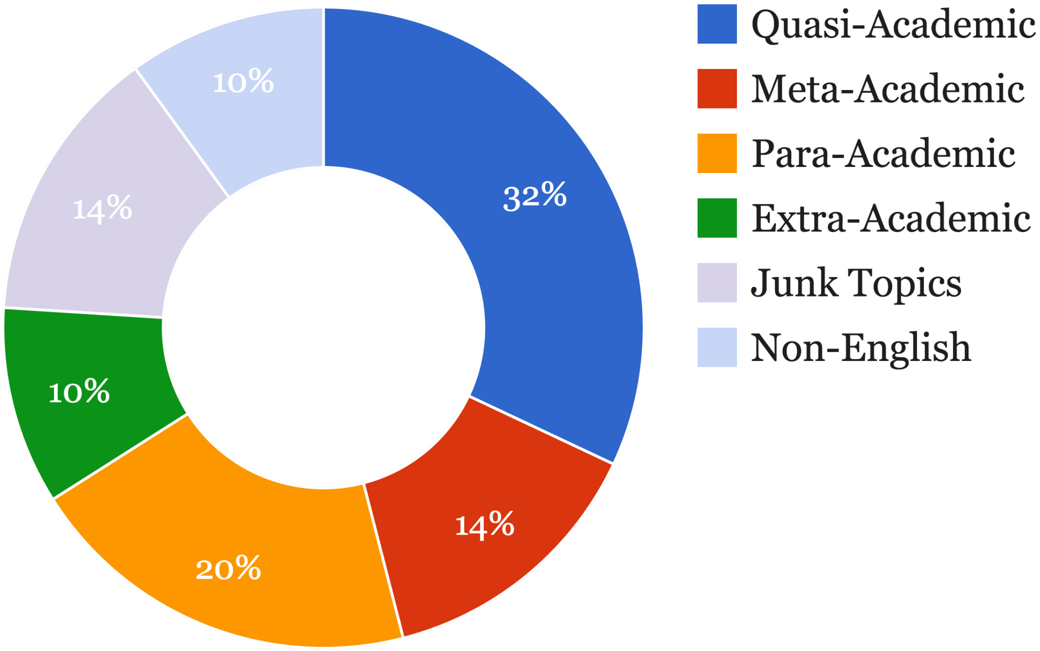


Figure (49): The distribution of categories over 100 topics. Each topic was assigned to one of the four categories of informal scholarly communication or labeled as junk or non-english.

Revisiting the chart from chapter 5 that shows the distribution of categories of informal scholarly communication over the 100 topics of the model

This chart in Figure (49) shows how most of the discussions on digital humanities blogs are focused on quasi-academic scholarship. Blogs are often criticized for being filled with writing with little to no scholarly value, but this chart shows a digital humanities blogs are actually the opposite. Extra-academic topics, discussing food, personal health, or local politics, make up on 10 percent, or ten, of the 100 topics in the model. The chart also shows the relative significance of each of the four categories of informal scholarly communication. The large portion of the quasi-academic category shows how digital humanities scholars are clearly using their blogs as a space mainly for discussing academic scholarship.

While the collective discourses are diverse, the entropy analysis in Chapter 5 showed digital humanities blogger's write mainly about only one or two topics with a long tail of other topics. This indicates there is a dynamic of self-discipline that manages an individual's breadth of subject matter. Part of this topic policing might be a result that while blogs are informal, they are also a public representation of their professional selves. Scholarly blogs are different from gadget blogs, mommy blogs, teenage angst blogs, or the many other genres that appear every day. The interviews, coupled with the focus on quasi- and para-academic topics, indicate blogs function as a form of self-promotion.

> [I] built up a professional reputation that I think has been important not only for carving out my niche in the field of history but also allowed me to, in a way, advertise myself to potential employers. Networking is an important skill in the academy, and I think having an online presence can go a long way in helping you cultivate a network of people across institutions. (Jason Heppler interview)
>
> At this point, Trevorowens.org is a professional/personal blog. I offer running commentary on issues in the field, but for the most part, it is not a place where I present original research as much as a place where I offer and develop my perspective on issues in this area of professional practice and scholarship. (Trevor Owens interview)
>
> I try to keep my blog to my research (no personal stuff) but my work is wide-ranging, so it covers many different fields. (Whitney Trettien interview)
>
> I think I [started my blog] partly as a career move: at that time, 2007–2008, I was working as a "Visiting Assistant Professor" in the English department at NCSU on a one-year contract and was looking for more permanent employment. But I wouldn't deny that there was a genuine desire to join a scholarly and quasi-scholarly conversation, as well. (Amanda French interview)

Blogs are a discursive platform or infrastructure for publishing; they also perform as *sociotechnical*, *public*, and *cross-institutional* performance of self. The institutional affiliations of scholars change; basic sociotechnical identity systems like email are constantly shifting. With the rise of adjunct and alt-ac labor, physical and institutional permanence is less and less stable. Blogs and scholar's personal websites are *semi-stable* across professional transitions. Though she's managed a professional website since the late ’90s, Amanda French registered the independent domain `amandafrench.net` in 2007 (Amanda French interview). As noted in Chapter 5, French has held multiple positions since 2007, working in multiple distinct capacities in and around academic institutions. Tracking her through these professional transitions, and the work product she produced at these various institutions such as any formally published materials or reports or projects are harder to track. Through all of this, her domain and her blog have remained constant. French’s changing interests are reflected in the rise and fall of topic proportions shown in Figure (45) in Chapter 5. The domain and the blog represent a stable sociotechnical identifier, *amandafrench.net*, upon which *she* decides how she is represented online.[123]

Herbert Menzel’s (1968) six functions scaffold this section that explores how blogs performs as a channel for informal scholarly communication. Key to their informality is the technical affordance of open publishing and how blogs provide no structural indicators of quality or value. Unlike formal publications like journal articles, which are delineated *as scholarship*, there is no structural delineation between the quasi-academic, para-academic, meta-academic, or extra-academic categories of informal communication. Blogs are a multiplex of scholarly communication and this constitutes their informality.

Promptness

Menzel argued that one of the beneficial functions of informal channels of scholarly communication is the promptness, in contrast to the slowness of formal channels, by which information can circulate in and through these networks. Even with electronic publishing, formal channels of scholarship such as journal articles and books still take a long time to materialize and, some argue, should take even longer (Levy 2007)!

[123]Unlike institutional websites, Academia.edu, or the dehumanizing ORCID.

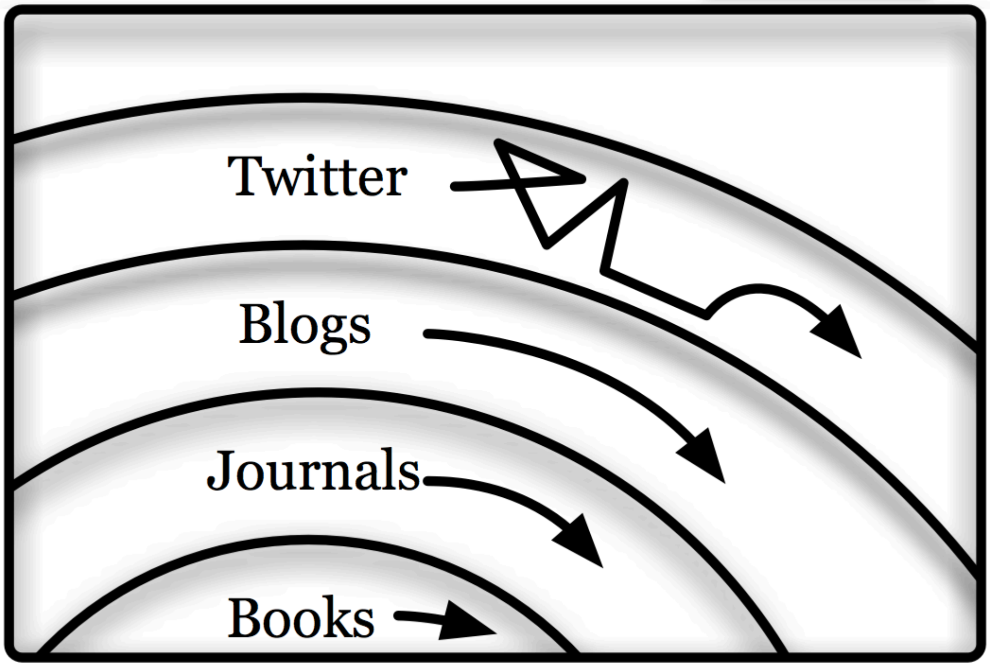


Figure (50). Pace layering, riffing on Brand (1994), of various forms of scholarly communication.

The diagram in Figure (50) shows the pace of different forms of scholarly communication. Books and journals at the bottom are a very slow form of publishing. Blogs are faster. Twitter's so fast that it has its own crazy temporal dynamics that are out of scope for this project. The Open Web affords open publishing, which means the structural barriers to posting online are basically nonexistent. This gives humanities scholars the ability to comment on current events *as they are happening* rather than long after the fact.

> I think blogging keeps things fresh. We're working on a book; the blogged draft has already had a bit of an impact. I'm worried the paper version will already be dated by the time it comes out (though this is one of the fastest book projects I've ever been involved with), precisely because the most interesting conversations are happening across the blogs, faster than the formal apparatus can keep up. But that's ok. (Shawn Graham interview)

> In some instances, blogs can fulfill the role of formal scholarly communication outlets (i.e., journals, book chapters), allowing scholars to quickly distribute their work. They also function as less formal means of communication. I believe Dan Cohen has described Twitter as allowing for sidewalk conversations like you would have in a physical neighborhood. I think blogs serve a similar function, allowing people to engage in asynchronous, geographically distributed conversations about the "state of DH," rant about labor practices, or comment on an idea presented a colleague's article. (Zach Coble interview)

Chapter 5 showed how blogs are used to discuss *academic themes in a public context* with Mike O'Mally's blog, *theaporetic.com*. O'Mally commented on the discussions about the gold standard in the mainstream media by providing important historical context about how non-standard and non-traditional the gold standard actually was historically, despite GOP rhetoric. The pace of blogging platforms enables an expert to comment about current events quickly and easily. If his commentary took three years to publish, it would no longer be relevant or interesting to the public. Timelessness, or at least the illusion of it, is a feature of formal scholarly communication, especially in the humanities where "findings" (interpretations) aren't made obsolete as much as fall in and out of fashion.[124]

Selective Switching

This is the function whereby scholars are made aware of information through social networks rather than formal indexing schemes.

> Faster turn around [with blogs, they are] less formal, [and] better for fostering discussion (in this case, discussion more often comes in reply blog posts by others, or fruitful twitter conversations). I feel like it reduces a lot of the systemic barriers that just get in the way of circulating research. (Scott Weingart interview)

Open publishing, open access, and open standards each contribute to the selective switching function of blogs. Links and RSS feeds provide easy ways to circulate information, open access and open publishing reduces the barriers to finding and pointing to information across the web. The open standard of RSS feeds lets scholars subscribe to a set of bloggers and get notified of new content whenever it is posted. The social web has transformed this stream or river of news; Twitter is replacing RSS feeds.

> Thanks to Twitter I'm exposed to many, many more posts written by people doing DH. (Jason Heppler interivew)

[124] One could argue this is also the purpose of academics' opinion pieces in mainstream newspapers, but I would respond that most op-eds in the media are now hosted on blogs.

> I think Twitter has somewhat supplanted blogs as a medium for ongoing debates, although Twitter also helps to bring blog posts into public visibility. (Lisa Spiro interview)

While the underlying platforms and infrastructures may change, the function is still the same. Digital humanities scholars are constantly sharing links to what they read, either in an anemic tweet or as a responsive blog post. This circulation performs the function of selective switching by informing blog readers, RSS subscribers, and Twitter followers of potential relevant or interesting information.

The difference between Menzel's selective switching of the late 1960s and selective switching today is an indicator of how the network of relations has reconfigured to accommodate scale. Rather than a one to one relationship selective switching operates on a one to many, or many to many, peer-to-peer relationship model. Additionally, technologies such as RSS enable a more passive form of selective switching. RSS automatically notifies subscribers of new content, Twitter feeds are a continually updated stream of links, snark, and commentary of the community.

> The way I interface with the DH community online is mostly via reading Twitter and following tweeted links to blog posts that expand on Twitter conversations or contain DH research thinking. I see blogging as a place where a lot of medium-length DH thinking is made public, and Twitter as how we hear about it and recommend it to others (or give negative peer review by not sharing it with others, or creating blog posts in response that disagree). (Amanda Visconti interview)

Menzel emphasized the *social* dynamic of selective switching, that is, your peers know what you are interested in better than a formal classification scheme.[125] Selective switching is about peer-networks helping each other finding relevant content. In the wide diversity of themes in digital humanities blogs, finding interesting posts to read would seem challenging (especially when considered

[125] Some of this function has been automated with recommender systems and collaborative filtering. I know I have made several purchases of academic books based upon Amazon's "Customers who bought this also bought" recommendations. Journal publishers and now even citation management services are using recommender systems. Scholarly publishers are increasingly aggregating data to extract value and insight, in part to further knowledge, but also to find new premium, for-pay services. http://blog.mendeley.com/design-research-tools/whats-relevant-to-me-right-now/#more-200

from the traditional, formal, print, and librarian centric information retrieval perspective). Yet, in practice Twitter and RSS guide the community's attention.

While the digital humanities community as a whole writes about diverse topics, the analysis of topic entropy in Chapter 4 showed that individual bloggers focus on one or two topics with a long tail of other content.

> I mostly talk about interpreting history as represented in new media, discussion of methods of research and scholarship in digital history and the digital humanities, and issues around the design, development, and process for the use of digital technologies in collecting, preserving, and providing access to cultural heritage materials. I upon occasion will delve into other issues [such as] changes in scholarly communication. Another way to say this is that the thematic unity of the blog is that it covers the things I have an academic/professional interest in. (Trevor Owens interview)

The RSS subscription model works because digital humanities bloggers tend to write about only one or two topics. Readers interested in those topics can subscribe to an RSS feed which will generally yield a steady flow of relevant content. While individual bloggers write about a small number of topics, the community as a whole is diverse and using RSS readers assemble a feed that is tailored to their interests. It is almost like a recommender system, but through curation instead of algorithmic ranking.

Screening, Evaluation, & Synthesis

Selective switching via Twitter and RSS feeds is not a perfect system. Twitter and blogs suffer from link-bait and contribute to the "Buzzfeedification" of scholarly communication. Blog posts of value are not always the ones that get the most attention; those that are the most inflammatory win the eyeball games. In a system with a lot of diverse content, popularity is not an effective measure of value and quality. How then, beyond the management and curation of individual RSS feeds and Twitter followers, can the community perform Menzel's functions of screening, evaluation, and synthesis?

Digital Humanities Now is a proven, and replicated, model for finding good content on the Open Web. Digital Humanities Now began in 2009 as a collection of Twitter feeds curated by Dan Cohen (yes, he crops up a lot). A commercial service, *Twittertim.es* (now known as *The Tweeted Times*[126]), monitored the social media feed and *algorithmically* selected the most discussed articles from twitter conversations. As DHNow evolved the service became less and less dependent upon algorithmic selection. The 2.0 release completely overhauled the process and structure moving to blogs and RSS feeds as the source for content and away from the commercial black box algorithms to a human powered editorial process. [127] A small group of editors from CHNM (and now a group of volunteers from the community) nominate and reviews blog posts each week and select one or two as "editor's choice." The immediacy of the reviewing process is important.

The subject matter of digital humanities blogs is very diverse by virtue of open publishing. The DHNow editorial model is an attempt at sorting through this diversity and finding good or valuable content *at scale*. It isn't traditional peer review, it is a kind of *post-publication peer review* (Fitzpatrick 2011), but with an important caveat, *not everything is evaluated*. Given the rapid pace, selecting only relevant content week-by-week, DHNow editors miss out on a wealth of quality blog posts. Some weeks might have ten great posts, while others only have one. The model, unlike traditional submit-then-review models, is set up to move fast and scale at the price of coverage. The DHNow model is not a replacement for traditional peer review, but it is an effective model for screening, evaluating, and synthesizing from the vast diversity of blog posts some of the ones worthy of attention.

Transmitting the Ineffable

"Know how" or "how to" posts are a very popular genre of blog. In the interviews, multiple bloggers mentioned their "how to" posts were their most trafficked and

[126] http://tweetedtimes.com

[127] The Digital Humanities Now editorial process is now concretized in an open source WordPress plugin. Plugin. The plugin may serve as a process of *stabilization* (Siles 2010) and *normalization* (Bowker and Star 1999) in the sociotechnical shaping of the screening, evaluation, and synthesis. https://github.com/PressForward/pressforward

linked to entries. The clustering showed a well-defined collection of topics that, upon closer reading, were about technology and methodology. The prevalence of these topics in the model and their popularity as discussed in the interviews speaks to a vitally important function blogs have had in the digital humanities.

As reported in Chapter 5, the “how to” post is a kind of para-academic communication in keeping with the insights from Menzel over half a century ago. He argued informal communication was a channel for the circulation of

> unpublished minor details of already published findings; information about the use of techniques, the adaption of apparatus, or the availability of materials; generally the fruits of experience and know-how (Menzel 1968, 160).

For the digital humanities this genre of communication is absolutely essential because it fills a gap in both the published literature and in graduate student training. The hope that others will read and learn from technical blogs posts is a motivator for writing them up.

> I like to futz about with new (digital) toys, to make them do unexpected things, to think through how they might be of use to others, to figure out how to tell others how they might want to use them. I do bits of analyses, munge data together to share with others. (Shawn Graham interview)

Where else can a historian learn about how to set up a text editor and transform articles written in LaTeX into submission ready manuscripts? How-to blogs in this case have served a couple of roles. First, they are the place where the technical details of digital humanities *in practice* are being documented and shared. Second, they are serving a pivotal translational role in contextualizing valuable information from other disciplines.[128]

Transmitting this “ineffable” (in the face of formal scholarly publishing) information is one of the primary functions of blogs. Technical communication is an important component in maintaining digital humanities as a community of

[128] I bootstrapped my knowledge about topic modeling by reading blogs, not from the formally published articles. The blogs contextualized and explained topic modeling in more humane terms providing me with a baseline level of knowledge so I could read and better understand the original articles. Furthermore, most of the early discussions about how to use and interpret topic models occurred on blogs, the books and journal articles came later.

practice, but what is especially important is how these transmissions have occurred out on the Open Web were they are accessible and searchable by anyone, not just members of the community.

Instantaneous feedback

Comments are one of the defining features of blogs. They provide instantaneous feedback for scholars who blog about their latest ideas or experiments in digital methodologies.[129] Comments or what Joseph Reagle (2015) calls "the bottom half of the internet," are an important component to factor into the conceptualizations of the (in)visible college.

> I have often gotten a lot of comments, and they've been one of the best aspects of the whole blogging experience. But increasingly I am being selective about which posts I open for comment. (Ted Underwood interview)

Underwood disables comments when he writes about already-published and peer reviewed work on his blog, "comments would really be reviews, and I think it's a conflict of interest for me to be moderating a comment thread reviewing my own research" (Interview). Underwood uses his blogs not only as a means to circulate already-published research, he also workshops ideas.

> But where I'm presenting work in progress, or discussing matters of general interest, I'm still going to try to keep the blog open for comments. They've often been really valuable contributions to my thinking. (Ted Underwood interview)

As an example of a set of interaction with a positive outcome there is an interesting thread that began in the comments in a blog post by Ted Underwood about topic modeling.[130] In the comments, Lisa Rhody asked a question about methods and Underwood's specific interpretation of a topic. Underwood

[129] Comments as instantaneous feedback assumes an audience. External analysis can infer the presence of a readership when comments exist, but the absence of comments does not imply the opposite. A blog post might get no comments because of a lack of commentary from the readership, or it might imply no readership at all. Comments and subsequently the dynamics of readership were not part of this research, both by design and for pragmatic reasons.

[130] Underwood, Ted. "A Touching Detail Produce by LDA."" http://tedunderwood.com/2012/03/25/a-touching-detail-produced-by-lda/

responded in the comments, but Rhody had more to say about the subject and wrote a blog post of her own in response.[131] Rhody and Underwood continued to converse in the comments to her blog post, and then Underwood wrote *another* response on his blog. Rhody cites this exchange in her article about topic modeling in the *Journal of Digital Humanities*.[132] Notice how the informal influences the formal.

This example is the ideal outcome for scholarly interaction on blogs. A new idea is worked out in and through comments and follow-up posts. The interaction is cordial, productive, and the ideas worked out lead to a formally published article *that credited the original conversation instead of hiding it*.[133] Interactions like this are possible because of the technical affordances of blogs and the Open Web. This does not mean however that such interactions are always so "nice" or productive.

The gap in disciplinary epistemologies and orientations to the production of knowledge are wide and deep. Formal scholarly publishing reifies disciplinary distinctions, placing the burden upon individual scholars to "read widely" (if they can find the time). The structural dynamics of blogs and the Open Web reduce the transaction cost of interdisciplinary discourse, but when the gap has been bridged, ideological and epistemological tensions can emerge. Comments are one of the places where interdisciplinary conflict or misunderstanding manifest. For example, in the fall of 2013 there was a discussion about the intersection of natural language processing and comparative literary studies on a blog called the Language Log. Scholars from literary studies wrote a critique of an article

[131] Rhody, Lisa. "Chunk Topics and Themes in LDA" http://www.lisarhody.com/chunks-topics-and-themes-in-lda/

[132] Rhody cites this conversation in footnote in her article. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/#topic-modeling-and-figurative-language-by-lisa-m-rhody-n-9

[133] I can follow the relationship between the formal journal and the informal blog thread as a trace ethnographer. The rendered visible and public in and through infrastructure. How many of these connections are lost and how much knowledge work is invisible because print-centric publishing occludes these relations.

published in a computational linguistics conference by computer scientists David Bamman, Brendan O'Conner, and Noah Smith titled "Learning Latent Personas of Film Characters" (2013). The critique by Alpert-Abrams and Garrette criticized the article, from the perspective of literary scholars who study film, arguing archetypal theory is no longer widely used in literary studies and the source of data, Wikipedia articles, reveals "not how films work, but how a specific subcategory of the population talks about film."[134]

In a follow-up post, Bamman argued the criticisms of their computational linguistics article centered on a distinction between the *model* vs. the *output* as the article's main contribution. Bamman and his co-authors considered the model to be the contribution, and the data/findings to simply be a test of the model. The critics, from the humanities, did not understand this notion of a methodological contribution in the form of a computational model.[135]

The dynamics of open access means anyone can visit a blog, even unwanted visitors. Such interactions reveal a tension between *writing for the pubic versus writing in the public*.

> I have had a few anti-Stratfordians (people who deny that Shakespeare wrote his plays) email me because of my blog, one quite aggressively, demanding I read and comment on his work. I told him it was ridiculous and he continues to email me randomly, whenever he finds someone else to agree with him and thinks I'll be impressed. (Whitney Trettien interview)

Fortunately, these interactions have usually been at most a nuisance, but they are still a form of invisible labor levied upon scholars who maintain comments on their blogs. Managing a blog is work, not just technical work, but social and emotional work as well. The capacity for easy and rapid feedback means people can respond before really thinking through the implications of their response and that has led to several extremely negative incidents on digital humanities blogs (and especially Twitter).

[134] The critique was published in full on the blog. http://languagelog.ldc.upenn.edu/nll/?p=6968

[135] http://languagelog.ldc.upenn.edu/nll/?p=7094

Extraction of Action Implications

While Menzel framed the extraction of action implications as finding the practical applications of scientific knowledge, in a humanities context this could be seen as informing the general public through the exercise of expert knowledge and understanding of historical or cultural subjects. Humanities scholar's writing rarely garners a large audience or achieves measurable public impact (Greco and Wharton 2010). This can be especially true in jargon-heavy humanities disciplines. Many scholars want to write for the public, and some of them have turned to blogs as a way to reach a readership.

> I've met a lot of people on the fringes of academia (ex-academics, librarians working in the public sphere [not university libraries], book collectors), which has been amazing. It's forced me to think about how my work might speak to those who love books and history and digital tools who aren't researching or teaching for a living. (Whitney Trettien interview)

> It's really great to hear that people are reading and enjoying what you write, and it's a great reminder to make my work publicly accessible (one of my research foci is making the humanities more public, so it would be weird if my work on that wasn't itself public). (Amanda Visconti interview)

> We started the blog in the context of a course we were in together believing that it made more sense to write for both a public audience along with our professors. I don't want to just write for other academics—I want what I write to be accessible and available to whoever is interested in what I have to say. (Jason Heppler interview)

As discussed in the introduction, Matt Gold was writing his dissertation and living in Philadelphia in the mid 2000s. He was ABD, conflicted about his dissertation topic, and possessed by a strong desire to reach an audience beyond the purview of his committee; so he started a local political blog. From 2004 to 2006, Gold blogged every day, connecting with other bloggers, and cultivating a readership. The political blog was totally divorced from his academic work, contributing enormous amounts of personal anxiety about being "discovered" by hiring committees. "As [an academic who is] blogging about nonacademic topics, I was worried they wouldn't take [me] seriously as a candidate. It was really stressful" (Gold Interview).

The mid-2000s were a very different time for academic bloggers. An article in *Chronicle of Higher Education* by "Ivan Tribble" (a pseudonym used by a humanities professor at a small liberal-arts college in the Midwest) titled "Bloggers need not apply" cautioned blogging was detrimental to an academic career. The article set a tone for scholarly blogging for years. Blogs were seen within the academy as only platforms for personal, *extra-academic* writing. Over time, as shown on danah boyd's early and influential blog, the content of blogs shifted away from the personal towards the professional. Scholars, along with political writers, journalists, and technologists, have all turned to blogs as a form of serious writing and the general public has turned to blogs as a place for serious information.

Menzel's six functions of informal scholarly communication are as relevant today as they ever have been. His framework helps understand how blogs perform as (in)visible colleges; they are *informal* channels communication that are *public* and *visible* on the Open Web. Yet, informal communication is different today than in the late 1960s because it is *visible at scale* due the technical affordances of the Open Web. This openness at scale adds important dynamics, like the tension of writing *in* vs. *for* the public, and cultivates new configurations of knowledge production, like the post-publication peer review structures of DHNow, to handle scale. These dynamics relate to the wide circulation of content that is possible because of open access and open standards.

Visibility and Scale

While informality is a significant feature of an (in)visible college, visibility *at scale* is what distinguishes this conceptual model from existing models of scholarly communication. The visibility of blogs as an (in)visible college stems from the *open access* properties of the Open Web. The web is public (to those with access to the internet). There are no technical barriers that prevent access-to read or, more importantly, scraping-of the content blogs.

Scale is the other important dynamic. The data collection described in Chapter 4 resulted in the acquisition of **106,804** posts written from **1997 to 2013** on **396** blogs. This is a rough average of twenty-five blog posts per year. Now, many posts do not have substantial content, over sixty thousand posts had less than one hundred words. Furthermore, bloggers write at different paces. Some scholars post a single long, thoughtful article per year, others post a collection of links on a daily basis (*Openculture.com*, a group blog dedicated to free and open educational media had 3,581 posts in the dataset, although many of them are short links to other resources on the web so their presence in the topic model was minimal). Informality coupled with open publishing allows scholars to write and publish according to their personal style, although as shown above, digital humanists still practice restraint in content and form. This flexibility coupled with the lack of editorial control and peer review contributed to what Cronin (2001) calls *hyperauthorship*, albeit with a fundamentally different set of ethical implications that should be explored in future research.

The Open Web, as an infrastructure, enacts the sociotechnical conditions for new ways of studying communication and behavior. Blogs, as a native format of the Open Web, are composed from open standards, namely HTML, Javascript, CSS, and RSS. These open standards mean blogs are readily computable. This dynamic is what enables Google to index the web for their search engine. The PageRank algorithm leverages standardized HTML links, explicitly articulated in anchor tags and their *href* attributes. Similarly, working with Diffbot this research indexed 396 digital humanities blogs and extracted the text content from individual posts. Open access and open standards enabled the creation of an aggregate, large collection of informal discourse that had previously never been studied.

Circulation across Boundaries

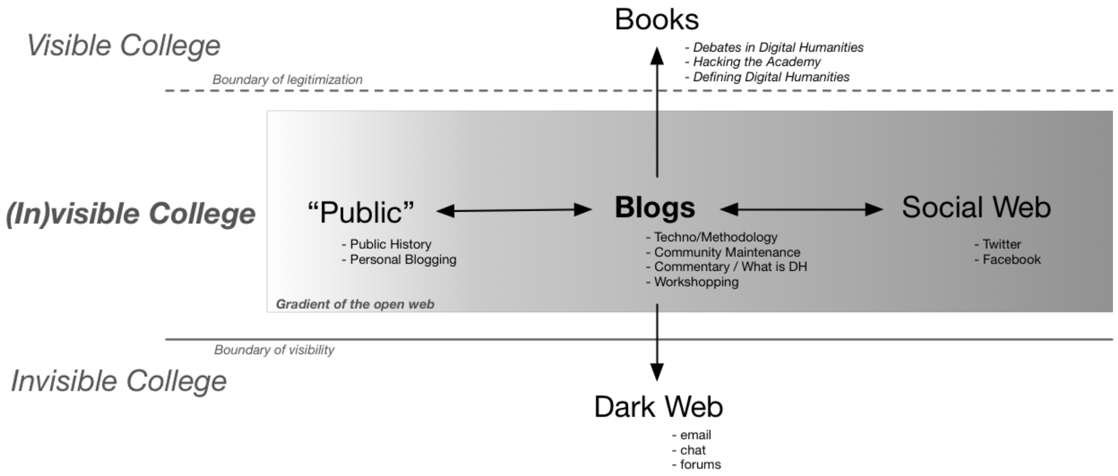


Figure (51). Spectrum of Circulation of Open Source Scholarship

One of dynamics around visibility is the circulation of content across the various boundaries that delineate (in)visibility. Writing that is posted on digital humanities blogs circulates across a variety of boundaries. In contrast to existing formulations of informal scholarly communication, blog writing is not only directed towards formal publications, it is written for the community (or simply just to specific authors).

Other blogs, such as Wonders & Marvels described in Chapter 5, are written with the express purpose of reaching a public audience. Wonders & Marvels circulates through the porous space between the problematic distinction between "academia" and the "public" and features fiction and scholarly authors who write about historical topics for a general audience (Holly Tucker Interview). This is the "open source" nature of blogs on the Open Web. As the diagram in Figure 52 shows, there are no structural boundaries and distinctions within an (in)visible college, instead there is a dynamic gradient.

The structural boundaries that do exist are not determiners of genre or disciplinary orientation, but rather cross the boundaries "up" into *legitimization* or "down" into *invisibility*. Moving "down" the diagram information and interaction enters the invisible college of scholarly communication. A space we can theorize about or study micro-sociologically, but is not observable at scale. Moving "up" in the diagram above means crossing into the scholarly record and into the lifecycles of legitimization through peer review, dissemination into journals or books, and preservation by libraries and archives (or publishers?). When blogs are transformed into journals or books, they are afforded new forms of visibility, in citation networks, in formal dissemination process, or into history via preservation.

Debates in the Digital Humanities

Debates in the Digital Humanities is an excellent case study for the alternative means by which blogs posts circulate "up" and cross the "boundary of legitimization." The volume was edited by Matthew Gold and published in print form by the University of Wisconsin Press in 2012. The book is important to the digital humanities because it brought together the extensive conversations and debates about the digital humanities into a single volume. Gold solicited essays about digital humanities from scholars, organized them under five headings, and did masterful editorial work by bringing together a conversation that had previously been percolating throughout the digital humanities community, informally and irregularly, as a dangerous undertow. *Debates* was novel and timely not only in content, but also in process. As Gold describes in his introduction, the book underwent a semi-public peer review whereby authors submitted their essays and reviewed essays from other contributors. *Debates* also innovated in the sources from which it selected entries. Beyond the solicited contributions, each section of the book included a selection of blog posts incorporated into the volume as individual chapters alongside the solicited essays. It is important to highlight that these entries were explicitly labeled as blog posts, not merely essays that had started out as blog posts and informally evolved into standard book chapters. The blog posts included in *Debates* underwent only minor editing and their provenance as blog posts is preserved, rather than erased.

In a blog post reflecting on the process, one of the contributors, Trevor Owens, talked about the strangeness of seeing “things I hadn’t intended for print in print.” The transition from web-centricity to print-centricity is a crossing of boundaries of expectation. The meaning of the post to Owens, the author, changed for the better, but was also unexpected.

> It is fun and neat to have a post end up in a book, but it is also a bit disorienting. On my blog it was part of a threaded run of posts about my teaching and writing. I like to think that everything I write here always remains a draft. Everything I write here is something I might return to and revise. Undoubtedly there will be typos in this post that someone will point out that I will fix. But now, reading the post on paper in this volume, it feels completely different. Instead of being my informal thinking out loud on my teaching it has become something much more enduring.[136]

Owens’s ideas and the texts circulated through a series of people, artifacts, and institutions starting with a post about course blogs and ending up within a print volume. Owens highlights the material transformation, the different “feeling,” of his words. He treats his blog as a first draft of ideas that he would “return to and revise,” but instead of the ideas silently and invisibly percolating into traditional print-centric work, debates rapidly elevated it to a new kind of documentary status of content published within the (in)visible college.

Implications of the Research

The implications of this research are broad. This section focuses on three areas, computational research, the digital humanities, and scholarly communication. For computational research, there are implications about *data fitness* or how data are shaped and re-shaped as part of computational or algorithmic processes. For the digital humanities, this is the first data-driven study of the community. The empirical results will provide fuel for ongoing debtates in digital humanities, but also the data reveal the importance of libraries and librarians as first-order members and contributors to research, teaching, and service to the community.

[136] Owens, Trevor. “Debating the Digital Humanities Gets Real”
http://www.trevorowens.org/2012/01/debating-the-digital-humanities-gets-real/

Finally, the implications of the (in)visible college upon scholarly communication is described.

Implications for Computational Research

Machine learning, data mining, information retrieval, and statistical learning processes require data to be of a certain shape. Chapter four described the data collection and preparation processes necessary to transforming blogs from the Open Web into "bags of word." I want to generalize this work of getting data into shape as *data fitness*. Data fitness is the process of transforming the shape and structure of data for computational analysis. Data fitness is an all encompassing term that includes the collection, cleaning, and transformation of data before, during, and after computational analysis. In this sense it is a broader conceptualization of data preparation because it includes the work that occurs before and after computational processing.

Data fitness emphasizes the structure or *shape* of data, but also the processes of *getting data into shape*. Gathering and assembling blogs on the Open Web invoked a series of tensions. Blogs as HTML pages are semi-structured but the semantics of an HTML document need to be mapped into the semantics of document-term vectors for topic modeling. This creates a tension because the decision about what to include and what to discard in the reshaping process affect the outcome of the computational analysis. Documentation of the data fitness was crucial in this regard because it frames whatever claims emerged from the interpretation.

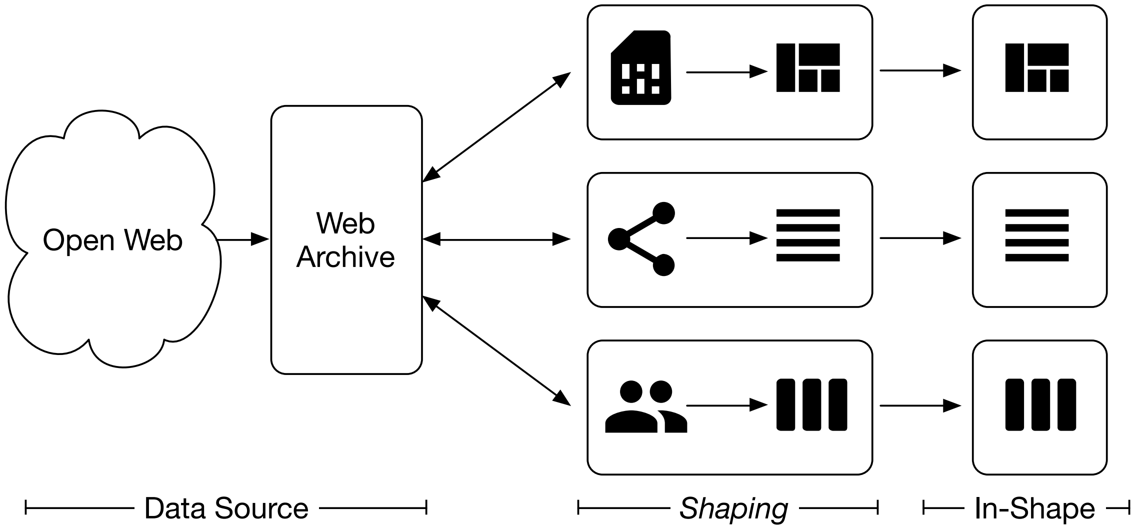


Figure (52). A data fitness program: The process of getting data from the web into algorithmic shape.

A *data fitness program* articulates the work of shaping data in preparation to work with different models and algorithms. For example, topic modeling expects data in a bag of words. Topic modeling blogs means extracting a set of word from the blogs. But which words? Blogs need their own data fitness program to specify which features should be selected. The algorithm and the data source are intertwined, the data fitness program is the wrapper around the whole process.

Data fitness highlights three areas:

1. The shape of the data source.
2. The re-shaping of data.
3. The shape of the data for the algorithm or model.

Data fitness is not only a conceptual model, it has the potential to be codified into an infrastructure for implementing data pipelines and machine learning algorithms. Once a standard set of shapes have been identified, generic tools and frameworks could be used to help facilitate the transformation of source data into shapes that can be easily plugged into different models and algorithms (depending on the shapes they expect). Furthermore, thinking about the shape of data provides a standard interface for the implementation of models and algorithms for machine learning researchers.

The use and application of machine learning algorithms is currently being held back by the lack of easy to use implementations. Topic modeling is popular because David Blei's LDA-C was posted publicly on the Internet. The popularity has only grown because MALLET is even easier to use and the user community is large enough to support itself. Data fitness could provide a technological and sociological frame around, which people and code could implement and use computational tools.

Implications for Digital Humanities

Much of the theorizing, postulating, and rhetoricians about the digital humanities has been anything but empirical. This study is a milestone in that it is the first systematic investigation to comprehensively compile data about the digital humanities and provide empirical, data-driven social studies of the digital humanities. What does it mean to have a *data-driven* understanding of the digital humanities? Here is what the data says:

> Digital humanities dh scholars research scholarship work projects tools scholarly humanists project field computing methods history technology studies academic text thatcamp literary scholar collaboration media traditional teaching analysis data questions neh building literature collaborative graduate humanist university tool disciplines pedagogy humanistic lab critical academy technologies students cultural institute center twitter.

The list of keywords above are the top twenty-five keywords from topic 75, the "what is DH" topic. If you were to poll the digital humanities community for a list of words that best describe the digital humanities, these terms would likely rise to the top. So what does it mean that an *unsupervised* machine learning algorithm was able to draw these terms together in a way that has high face-validity to the community under study?

First, it is an indicator that the method works. Distant reading has matured to a level where it can find semantically meaningful patterns within a diverse, multidimensional space. This "unreasonable effectiveness" (Wigner 1960) is simultaneously spooky, but also powerful because it allows scholars to ask questions previously impossible because of scope or scale (Jockers 2013).

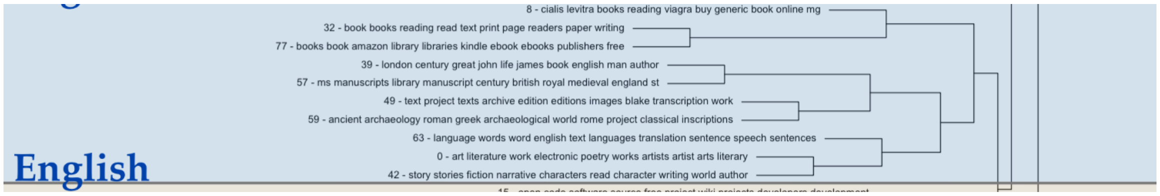

8 - cialis levitra books reading viagra buy generic book online mg
32 - book books reading read text print page readers paper writing
77 - books book amazon library libraries kindle ebook ebooks publishers free
39 - london century great john life james book english man author
57 - ms manuscripts library manuscript century british royal medieval england st
49 - text project texts archive edition editions images blake transcription work
59 - ancient archaeology roman greek archaeological world rome project classical inscriptions
63 - language words word english text languages translation sentence speech sentences
0 - art literature work electronic poetry works artists artist arts literary
42 - story stories fiction narrative characters read character writing world author
English

Figure (53). The English cluster of topics

The dendrogram of topics confirmed Kirschenbaum's argument that English dominates the digital humanities. Historical topics, another discipline with strong digital humanities involvement, did not pop out with as clear of a cluster; its topics were diffused throughout the model. For example, some historians showed up in the methodology topic, or there was a historical topic in the English cluster. This might indicate that discussions about historical themes are more diverse than discussions about English themes.

The findings of this study—that digital humanists are prolific bloggers who write about a variety of themes, who self-regulate, and for whom blogs are a vital component of community management and maintenance—are not, I suspect, surprising to digital humanists. The significance of blogs has been anecdotally understood among digital humanists for some time. Most of the reflective and reflexive accounts and inquiries into the digital humanities has been methodologically traditional. From a research perspective, DH places a strong emphasis on data and computational analysis, yet most of the research in the field has been rhetorical and anecdotal in nature. More importantly, this study is a digital humanities-style study of the digital humanities community.

"Where are the social studies of the digital humanities?" asks Borgman (2009). Does the digital humanities (or humanities more broadly) need a social science to study it? I would argue "yes," and this dissertation makes a case for the potential value of a social science-oriented investigation of humanities scholarship practices. The social significance of the sciences is partially a result of the history and philosophy of science. Their existence contributes to the social significance;

they are not merely a reflection of it. The social significance of the humanities and digital humanities needs to be more broadly affirmed. Studying them helps scholars and our broader society understand why the academic fields of history and English (whether digital or not) are worthy of intellectual exploration.

This project is exploratory. It has not sought to test assumptions about digital humanities or reveal, through distant reading, a previously unknown fact about the digital humanities. Such research is a possible next step for future social studies of the digital humanities. The methods and findings show a new way of knowing about digital humanities, a way driven by data and distant reading. The results reinforces the perspective that a traditional humanistic way of knowing about the digital humanities, through argument and rhetoric, actually works pretty well and is effective at characterizing the field. This finding might be unsatisfying to someone hoping the data-driven, quantitative methods would shatter commonly held beliefs.

Digital Humanities and Libraries

The *Compendium of Digital Humanities* has a very inclusive heuristic for membership. The list includes tenure-track faculty in "classic" digital humanities fields (i.e., English and history), but there are also journalists, such as Alexis Madrigal, who write thoughtful and well-reasoned articles about culture and technology. The list is very broad and it is a reflection of the breadth of the digital humanities community in its acceptance of non-traditional intellectuals. Yet, for all of this rhetoric of inclusiveness, the list is dominated by two kinds of scholars: traditional academics (professors, postdocs, and students) and librarians. This composition reflects an underlying tension that I don't fully understand, though I observed some traces of it in the topic model.

Librarians constitute a large portion of the digital humanities community. They are well represented in the digital humanities blogosphere, on Twitter, and at the annual digital humanities conference. The librarian portion of the digital humanities community instantiated their own version of the DHNow editorial model, DH+Lib. Librarians and academic libraries are just as much a part of the digital humanities as English and history.

Libraries have historically always been oriented toward the service and support of the humanities. The librarian's role is characterized as reactive and passive, serving the information needs of students and faculty in the course of their research, teaching, and learning. Librarians, mainly in academic libraries, see the digital humanities as a moment where this servile relationship can be reconfigured (Vandegrift 2012; Nowviskie 2013).

> What would happen if we saw our libraries' obligation to the DH community as being less about the provision of smooth and reliable services leading to the continuation of smooth and reliable services, and more about building on our own organizational and operational knowledge to model the digital humanities being done well? (Nowviskie 2013, 60)

Nowviskie argues that libraries, not the academic units, are best positioned to take advantage of the organizational and epistemological opportunities. Crucial to her vision is a library's organizational and operational advantage over digital humanities driven by faculty.

> [Libraries] understand the way that open source communities are cultivated and the benefits of investing in them. The digital humanities community pays a good deal of lip service to open source, but not many scholarly projects do it well. (Nowviskie 2013, 63)

One of the factors that helped the digital humanities to thrive is a willingness to talk, write, and openly share in-development concepts and ideas, meta-quandaries about the community, and technical knowledge *on blogs*. Knowledge, especially technical knowledge, is openly published and shared on blogs especially *by librarians*. Methodological themes, like those captured in topic 45 (, *data number results analysis numbers time year study survey report words*, have a mix of faculty and librarians in the top-ranked proportions. Librarians, unlike classically trained humanities faculty, have the technical knowledge necessary to do digital humanities work; they know how to collaborate (across disciplines), and they understand the *radical openness* upon with digital humanities thrives.

> Many of us sense that we are moving into a kind of alternative academic universe where long-held stereotypes of faculty and librarian personalities, research interests, devotions, inclinations, and native capacities break down. (Nowviskie 2013, 65)

This sentiment is reflected in the topics aggregating job postings, many of which are alt-ac, library, and other non-tenure stream positions associated with the digital humanities. As I noted in Chapter 5, only seven of the top twenty job postings identified in topic 10 were tenure stream. What are the larger implications of moving academic labor, first teaching and now research, out of the hands of tenured faculty and into contingent knowledge workers? Questions and anxieties about contingent teaching labor have haunted humanities academics for decades (witness the many editorials about "adjunctification"). Now, with the rise of alternative academic careers focused on applied research, grant writing, and innovative forms of scholarship, what does it mean to be a tenure-track professor? Does digital humanities need a professoriate?

Implications for Scholarly Communication

The impact of the web on scholarly communication has not gone unnoticed (Cronin et al. 1998; Ginsparg 2004; Kling and Callahan 2003; Kling and Covi 1995; Kling, McKim, and King 2003), but previous attempts have suffered from premature generalization or narrow proscription. The dynamic nature of the Open Web means the sociotechnical assemblages and configurations of people, practice, and technology are only partially stable. New information communications technologies, including especially Twitter, are actively and enthusiastically used by digital humanities scholars today, but in a couple years maybe a different platform, with its own particular set of affordances, will become popular.

The (in)visible college is a *descriptive* rather than *proscriptive* model for scholarly communication. The (in)visible college is a characterization and conceptualization of a scholarly communications ecosystem *that currently exists*. I am not attempting to characterize an ideal scholarly communications architecture that "starts from scratch" (Ginsparg 2004) because local scholarly communities are already experimenting and innovating within the accreted brambles and barnacles of the existing ecosystem. The digital humanities community still operates traditional print-centric journals (*Digital Scholarship in the Humanities*), open-access digital journals (*Digital Humanities Quarterly*), and publishes prolifically in traditional online monographs. We don't *need* any more proscriptive models to imagine an ecosystem that fully leverages the power of the Open Web. There is a "quiet revolution" ([Humanist 18.001] from Svensson 2009) in scholarly publishing and it is happening in and through (in)visible colleges.

Blogs are an undifferentiated mass. Unlike formal scholarly communication, there are no structural features to indicate quality. So the default position has been to categorize all scholarly communication on blogs as informal. This does not mean however that there is nothing of quality on blogs; I can't imagine digital humanists would spend their time and effort blogging like they do if it was all for nothing. There is value in blogs, but the ways in which it is foregrounded do not resemble the traditional systems of legitimization (i.e., peer review as it is normally conceived).

The technical affordance of open publishing, accompanied by the willingness to write about a variety of themes, means there are no constraints upon what gets published. Publishing in an (in)visible college has a different meaning than in formal scholarly communication. In traditional scholarly communication, "published" means peer reviewed, legitimated, and enmeshed within the infrastructures and practices of knowledge production in the academy. In the (in)visible college, it merely means to put something online whereby it becomes accessible. There are no processes of legitimization, there is no review. This means it is as easy to publish a transcript of your latest conference talk as it is your favorite recipe.

I found, in practice, that digital humanities bloggers as a whole write about a broad collection of themes, but as individuals they tend to focus on only a couple areas of interest. In the interviews (an admittedly biased sample), scholars reported that they tend to shy away from blogging about personal or private matters. Most of the scholars in my sample write about scholarly themes. They also write *a lot*. So the community faces a situation where there is a substantial amount of valuable and interesting writing occurring in an informal space.

In response to the need to identify valuable and relevant content, programs like Digital Humanities Now have emerged to help find, filter, and feature content in the (in)visible college. DHNow is a functioning model for legitimization of content published on the Open Web. It is also a system that scales, albeit at a cost, unlike the *lossless* traditional peer review, the DHNow model is *lossy*. Posting something on a blog is not the same as submitting it to a journal. With a journal submission, there is a guarantee that it will be read and reviewed; the same is not true in the DHNow model. There is a *faith* that the sharing and circulation of blog posts via RSS and Twitter would have the effect of naturally raising good content to the top, but this means good content is determined by the tyranny of the crowds.

While blogs are structurally undifferentiated, my analysis of blog content uncovered four distinct categories of scholarly communication. Blogs are not a uniform modality; they are used in multiple ways by multiple constituencies within the digital humanities. My analysis reveals a depth and complexity to scholarly blogging that has never before been explicated systematically or at scale. This work is significant because it opens the door to a richer understanding of the digital humanities, but to other scholarly communities that communicate and interact within (in)visible colleges. The four categories not only show variation within informal spaces of communication, but they also have an empirical basis.

Their empirical foundation is the most significant contribution of the four categories because they can initiate a data-driven conversation about the kinds of discursive work conducted within informal channels of scholarly communication. Until this analysis, informal scholarly communication could only be understood indirectly through latent structures in formal bibliographic networks or via narrow (but deep) qualitative inquiry. Access at scale changes the avenues to knowing about how scholarly communities interact. This has implications for the sociological (or sociotechnical) understanding of knowledge.

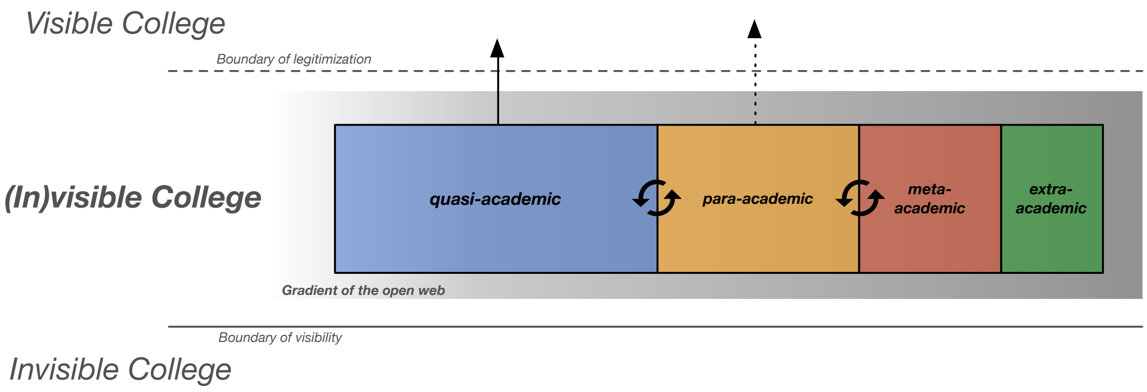


Figure (54). The four categories of informal scholarly communication. Each is proportional to their distribution in the topic model. The arrows indicate circulations.

Figure (54) above represents the relationship the four categories of informal scholarly communication have within the (in)visible college of the digital humanities. The size of each box represents the category's proportion across the topic model. I have added some arrows to indicate the relationship each category has with other modes of scholarly communication. The circular arrows are meant to show how information in these categories circulate in and through the (in)visible college. Projects like DHNow, and platforms like Twitter, through the processes of *selective switching* and *screening*, *evaluation*, *and synthesis* discussed above, facilitate dissemination inside the context of the Open Web. The missing arrow with the extra-academic category represents how personal blog posts are not circulated as widely as the other categories. In my interviews, scholars typically kept their blogs "strictly professional" and review systems like DHNow expressly filter out extra-academic content. While extra-academic content exists and circulates within the community, it appears to be something like background noise made possible because of open publishing. Digital humanities blogging is mainly focused on quasi, para, and meta-academic categories of discourse, understanding the specific significance of extra-academic conversations in this field is an open question.

Quasi- and para-academic categories have an outward facing orientation that has significance in two ways. First, they are more generally focused upon the circulation of knowledge (either new or previously published in formal venues). Scholars writing the quasi-academic category are writing about traditional humanities research themes or they are bringing their knowledge and expertise to bear in a public context. As we have seen with Debates in Digital Humanities, Hacking the academy, and Defining Digital Humanities, text-centric, quasi-academic blog posts can be brought into the formal channels and preserved as traditional texts.

Para-academic categories are circulating radically new forms of knowledge that should circulate up into the formal systems of scholarly publishing; they don't because of structural barriers. Para-academic discourse are not only differentiated by content, they also includes multi-modal productions, like Ted underwood and Andrew Goldstone's online appendix. This kind of para-academic discourse can't circulate up into the formal channel, because both the content may be epistemologically alien to formal channels or because the material form can't be separated from its infrastructural context. For example, blog posts and web publishing more generally can leverage a web browser's multi-modal, interactive, and non-linear capacities. This work is scholarly, but how can it be included in the permanent scholarly record?

As an infrastructure for the production and circulation of knowledge, blogs and the Open Web present a grand challenge for the preservation. Preserving the entire web, as per the Internet Archive and Common Crawl, is not the answer. Archiving the entire Open Web doesn't solve the key issue, differentiating the content with scholarly and epistemological value (quasi and para-academic categories) from the content that has local, personal, and non-scholarly value (meta and extra-academic categories). The infrastructures of formal scholarly publication have three general functions, *legitimization, dissemination,* and *preservation* (Borgman 2007). As blogs and the Open Web are increasing used as an alternative infrastructure for scholarly communication (Lagoze et al. 2015), these functions *must* be reconfigured.

Limitations of the Research

The corpus I have assembled is rich with potential insights about the digital humanities and scholarly communication. My exploration of topics has only revealed a small portion of what the data have to offer researchers interested in knowing more about digital humanities and scholarly communication. The processes by which I collected, cleaned, and prepared the data as well as the topic model I generated from the text content of blog posts shows a highly reified representation of the immeasurable richness and complexity of digital humanities blogs. Scalar devices bring certain dimensions of a phenomenon to fore, and occlude others. My choice to use topic modeling was informed by the question I wanted to answer, what are digital humanists writing about on their blogs?

There are other scalar devices that might reveal other interesting dynamics of the digital humanities or of informal scholarly communication in an (in)visible college. An obvious choice would be social network analysis, which might reveal some of the structural relations between and across the digital humanities community. Who interacts with whom? How much? Are their subcommunities? Are there scholars who should be included in the compendium that are not currently listed? Topic modeling focuses on text content, where network analysis focuses on structure (which constitutes different features to be extracted from the HTML structure of blogs). Each of these approaches foregrounds and backgrounds different aspects of blogs and neither is globally better or worse than the other. That said, each scalar device is better or worse at answering particular questions of a large-scale sociotechnical phenomenon.

The process of data collection, cleaning, and preparation also foregrounded particular aspects of the blogs at the expense of others. Foremost was my decision not to include comments in my text analysis. My decisions are admittedly as much practical as they are epistemological. In my select interviews with digital humanities bloggers, most indicated that comments are actually not a significant part of their discursive practice. Comments *are* important to the constitution of scholarly blogs as an (in)visible college, but the practical reality is that most posts didn't actually have any comments. I am not sure what percentage of the posts I collected had comments, and this lack of knowledge connects to the practical challenges of extracting features from web pages. Because there is no standardized HTML structure for comments, they can be difficult to reliably extract. I delegated the feature extraction process to Diffbot, which does have a heuristic for counting the number of comments on a blog post, but I found the results to be unreliable on a blog-by-blog basis. It was more practical to completely ignore comments, rather than have a biased selection based on those blogs where it was easier to extract and count comments.

A potentially important limitation in the analysis concerns readership. The focus of this research has been on publishing and sharing writing on the web, but I have paid little attention to how the digital humanities community *reads* blogs. Quantitatively measuring and tracking readership is a bit more difficult, but the data does potentially exist in blogger's Google analytics and WordPress statistics. Obtaining the data is much more difficult because it is not public. The current enthusiasm for alt-metrics by scholarly publishers is an indicator that this information may not remain hidden for long (at least to certain powerful actors).

Another approach would be to engage the questions of readership qualitatively through surveys and interviews. Understanding how scholars read and use blogs, and other social media (especially Twitter) is a rich area for future research. Blogs do not exist in isolation from other forms of communication. It has been a challenge in this project to ignore Twitter as a platform for scholarly communication in the digital humanities. If you think scholars are enthusiastic about blogs, you don't understand their insane delight for Twitter. Twitter has its own set of sociotechnical dynamics that are different from blogs and the Open

Web. The extent to which those dynamics have impacted scholarly communication remains to be studied.

Future Research

The corpus of blogs that I have assembled is rich with knowledge and insight, this project has chosen only one of many possible paths into the data. Many other questions could be asked about the digital humanities, about blogs, or about scholarly communication in general. The next steps for future research can, as with the findings, be broken down along methodological, empirical, and conceptual lines.

Methodological

Digital humanities is at the bleeding edge of research that blends the quantitative and qualitative, the computational and the interpretive, and close and distant reading. This methodological experiment, trace ethnography of scaling, contributes to those ongoing areas of research. When studying the web using computational methods it is absolutely necessary to document and be transparent about the scalar devices used to transform the web into data and data into insight. However, the infrastructures used in this project are not set up to support the diligent requirements of theoretically sound research. Diffbot, while an excellent service, does not create archives of the websites it scrapes nor are the algorithms employed openly documented. Infrastructure in the service of computational interpretive social science needs to be more open.

We need a Diffbot service for scholars. This service would, like the commercial service, manage the web crawling and scraping infrastructure. It would also, like the commercial service, provide information extraction and classification layers. And again like the commercial service, it would have user-friendly interfaces and APIs so scholars with all forms of technical expertise could use it. The Diffbot service for scholars would also:

- Create archives of all pages crawled and make them available both programmatically (via Memento and other APIs) and humanely (via browser and search)
- Partnerships with existing web archival institutions, both independent like the Internet Archive and academic like the California Digital Libraries to distribute the preservation and access processes
- Provide full algorithmic transparency for any classification or feature extraction processes[137]

Empirical

The *Compendium of the Digital Humanities* is a living document. In the several years since I first sampled the *Compendium* in 2013, the number of blogs listed has grown by several hundred. Clearly, as a community of practice, the digital humanities embody a growing and moving target. This is the challenge of doing social science on a subject that is ever changing (as opposed to historical subjects whose dynamism has different, though equally challenging, characteristics). There is an opportunity to add more blogs and more data to the corpus and train a new topic model to see if the digital humanities conversations have changed in the last couple years.

The *Compendium* is also not the definitive authority on digital humanities bloggers, it is merely a convenient sample. There are potentially many other blogs that could be included in an analysis, but it would require different techniques to find them. Network analysis, such as following linking behavior, blogrolls and comment practices might yield another set of blogs to include in future analyses. There is a challenge however, what is the selection criteria for a blog to be included as a scholarly or digital humanities blog and what is to be excluded? My study delegated that determination to the *Compendium*, which had its

[137] Steps two and four of the data preparation process describe in Chapter 4 were challenging because I was dealing with Diffbot's black box algorithm. In practice, I was able to remediate the data cleanliness problems, but theoretically I am reliant upon processes I cannot know and this there is not practical way to overcome this barrier.

drawbacks, but any attempt at casting a broader net will face the inevitable "who's in, who's out" problem that has plagued the community since its inception.

Network analysis could be used to find new blogs, but also as a scalar device for exploring the scope and scale of the *Compendium* in a way that highlights different features than topic modeling. Rather than focusing on content, network analysis would reveal the structural dynamics of the corpus and possibly find patterns of interaction that can only be seen from a distance. Are there bloggers who regularly interact? Is there a small group of elite bloggers who dominate? Is there a long tail of bloggers to whom no one pays any attention?

Conceptual

This project makes no claims regarding the generalizability of the (in)visible college as a conceptual model. In its current incarnation, the concept is derived from an examination of a single, small, scholarly community. This does not mean the concept has no value, but that it needs to be tested, expanded, and refined. Do the four categories of communication, quasi-academic, para-academic, meta-academic, and extra-academic exist in other scholarly communities? Do posts in these communities circulate towards public readership or into formal publishing processes?

There are many other scholarly communities using blogs, the (in)visible college can be used as a basis of comparative analysis. The Object Oriented Ontologists, a subcommunity of philosophy, have a unofficial reputation for using blogs as a platform for doing philosophical and rhetorical work. There are also other platforms on the Open Web, such as Facebook and Twitter, whose sociotechnical affordances, corporate politics, and ethnics have implications for how scholars interact online. The (in)visible college provides a conceptual template to be fleshed out with explorations of other platforms and infrastructures.

Conclusion

Blogs and other scholarly communication on the web should remain (in)visible. Blogs should not generally count as scholarship. The new bibliometric movement of alt-metrics, the measurement of scholarly activity in the web, must thread a very fine line between surfacing and rewarding important yet (in)visible knowledge work and corrupting such space with poorly designed incentive structures. What would happen if annual evaluations included an expectation of blogging and tweeting with the public to operationalize impact and outreach? These media come naturally to some scholars, but not others, and they are suited for some forms of information, but not others.

Digital humanities bloggers are, for the most part, blogging for each other. They are commenting upon, linking to, and responding to each other without explicit incentives; the open circulation of knowledge and information is part of the community's discursive practice. Professionalization, institutionalization, and legitimization would have a significant impact upon digital humanities bloggers. However, as scholarly blogging becomes a focal lens for greater audiences and attention the possibility for monetization and commercialization becomes reality. For-profit scholarly publishers are beginning to commercialize scholarly blogging as they have scholarly publishing.[138] This is worrisome.

Scholarly publishers, formal legitimization, and metrics will have an adverse effect upon blogging. Such involvement would create perverse incentives that would upset the advantages of the (in)visible college that digital humanities bloggers currently enjoy. Raising the *visibility* of blogs, from the purview of traditional scholarly publishing, would mirror the professionalization,

[138] EBSCOhost is launching a new blog, EBSCOPost, dedicated to librarianship. https://www.ebsco.com/blog/article/welcome-to-ebscopost

corporatization, and institutionalization of technology, political, and journalistic blogging.[139]

If blogs were professionalized and commoditized they might succumb to the pressure of the “least publishable unit“ or LPU that plagues academic conference papers. Digital humanities blog posts are usually *smaller* than an LPU (remember the average length was around three hundred words), but they are flexible because there are no formal expectations of length, breadth, or depth. Metrics, like alt-metrics, in conjunction with hiring, promotion, and tenure review processes would optimize/corrupt the system (like impact factor and h-index has for publishing). Blogging, as an (in)visible college, is valuable because it lives at a lower status than traditional forms of scholarly communication. This lower status enabled experiments in subject matter and in form; experiments that continue to be crucial to the further development of the digital humanities as a discipline.

Making blogs count would render visible a discursive space that thrived precisely because it was (in)visible from counting. This is why blogs, as a whole category, should not be considered a new form of academic scholarship. New web-centric forms of scholarly communication are needed, and blogs, as a category, should be considered an ancestor of the web-centric genres of scholarship that have yet to come. The genre and form of scholarly communication that is going to radically transform and disrupt scholarly publishing and perhaps even the academy as a whole does not exist, or if it does, we won’t know its historical importance for a while.

Blogs are not the future of scholarly communication; they are its present and recent past. They have been a crucial piece of infrastructure for the digital humanities as the community transitioned from the narrow focus of humanities computing to the big tent of digital humanities. The technical properties of the

[139] Andrew Sullivan, one of the first political bloggers, is leaving blogging, causing the community to engage in a reflexive analysis of what has happened to the format.
http://www.vox.com/2015/1/30/7948091/andrew-sullivan-leaving-blogging

Open Web (open standards, access, and publishing) created the conditions of possibility for the emergence an (in)visible college, which in turn enriched the digital humanities. With the technological and infrastructural stage set, it was the particular social and cultural enthusiasms for technology and experimentation that lead the digital humanities community towards blogs as a platform for scholarly communication. Open standards, open access, and open publishing allowed a range of themes and rapid interactions from a diverse community of interest.

The digital humanities blogging community is composed of many different kinds of scholars. This composition is a reflection of the ongoing achievement of the digital humanities' diversity. Tenured and junior faculty, graduate students, administrators, librarians, technologists, and many others are constituents of the (in)visible college of digital humanities blogs. This diversity creates a productive tension yielding a vibrant online dimension to the broader digital humanities community around the world. How the digital humanities will develop in the future and what role blogs might take in other scholarly communities are open questions. Fortunately, the nature of both of digital humanities and blogs means we, as participants in and scholars of these phenomena, can continue to observe them.

Appendix A — Topic Modeling

Topic modeling is a catchall term for a group of computational techniques that, at a very high level, find patterns of co-occurrence in data (broadly conceived). In many cases, but not always, the data in question are words. More specifically, the frequency of words in documents. In natural language processing this is often called a "bag-of-words" model. A bag-of-words model has the effect of simplifying the complex structure of natural language by ignoring syntax and grammar and focusing on the frequency of words within documents. So instead of a properly ordered, grammatically correct sentence, the bag-of-words approach slices and dices text into a table of words and frequency counts.

You might wonder, *How can we find meaning without structure? Without order the meaning is lost!* Yes, significant context is lost by only counting words in documents. Such concerns are absolutely correct, but counting words is still quite effective.[140] My purpose here is not to engage in a prolonged argument about the epistemic validity of topic modeling's underlying assumptions; I merely want to describe them because I don't think they have been well articulated in other introductions to topic modeling. It is my hope as scholars from the humanities and interpretive social sciences learn more about topic modeling, text mining, and natural language processing, that their knowledge of language and writing

[140] Ted Underwood points out that while word counts are simplistic, they are still extremely powerful. The full richness of words themselves, he argues, are still not a fully utilized feature for machine learning algorithms. In the comments Ryan Shaw points to another blog post by Brendan O'Conner which succinctly and brilliantly observes: "Words are already a massive dimension reduction of the space of human experiences." http://tedunderwood.com/2013/02/20/wordcounts-are-amazing/

will inform the state-of-the-art of text and language models.

To understand and interpret topic models, it is important to have a solid understanding of how topic models work. Topic models have been described from a variety of perspectives, ranging from the metaphorical, like Jocker's LDA Buffet,[141] to the rigorously mathematical, like Blei, Ng, and Jordan's article introducing LDA in the *Journal of Machine Learning Research*,[142] to the pragmatic, like Brett's introduction in the *Journal of Digital Humanities*.[143] My goal is to describe topic modeling by complementing existing introductions to topic modeling and filling some important bits of information they have left out.

The following treatise has three parts. First, a brief jaunt into what I mean when I say "model." Second, a deeper discussion into what I mean by *word*, *document*, and *topic*. Third, a non-mathy description of topic models by tracing the evolving complexity of four generative language models. Not everything I cover here is directly related to topic modeling, but I think much of what I cover are assumptions and information generally left out of most topic modeling conversations. It is difficult to understand how topic modeling works if you don't understand natural language processing concepts like tokenization and stemming. Additionally, I think the distinction between a topic model's generative process, and the estimation of a topic model's parameters is an important detail left out of most discussions on topic modeling. Scholars interested in topic modeling need to know this stuff, so I have done my best to assemble it all together in one place.

[141] http://www.matthewjockers.net/macroanalysisbook/lda/

[142] Blei, David M.,, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3 (2003): 993–1022.

[143] http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/

On Models and Reality

The topic models I discuss here are known as *generative* topic models. Generative models try to represent, in computational abstraction, a process by which documents in a corpus could be authored. It is important to recognize such computational models are not claiming *this is how these documents were actually authored,* rather they are probabilistic approximations of the document creation process. In the *Companion to Digital Humanities*,[144] Willard McCarty explores what the term modeling means in a computational context.

> Two effects of computing make the distinction between "idea" or other sort of mental construct on the one hand, and on the other "model" in the sense we require: first, the demand for computational tractability, i.e., for complete explicitness and absolute consistency; second, the manipulability that a computational representation provides.

These two effects of computation that McCarty points out are crucial for understanding how topic modeling works and the kinds of knowledge it produces. First, *computational tractability* might be, for someone not trained in computer science or programming, a somewhat alien concept. To help illustrate this tractability problem, I want to share a wonderful anecdote from early pioneer of computational art, Frieder Nake. In the wonderful documentary, *Hello World! Processing,*[145] Nake tells the story of an interaction between another early computational artist Georg Nees and the painter Hans Drucker at a 1965 exhibition of Nees's computational art.

> The leadings fine artist, the painter Hans Drucker, raised his hand and said, "young man" addressing Georg Nees, "all said very well, what you told us, but you know what, could you make your machine draw the way I draw?" and Nees pondered for a moment and said, "you know what, if you tell me how you draw I can make my machine do it."

[144] McCarty, Willard. "Modeling: a study in words and meanings." *A Companion to Digital Hdigital umanities* (2004): 254–70270. http://nora.lis.uiuc.edu:3030/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-7&toc.depth=1&toc.id=ss1-3-7&brand=9781405103213_brand

[145] Hello World! Processing. http://vimeo.com/60731302#t=1800

Nake explains how Drucker and Nees were both correct, Drucker assumed Nees's answer would be "no,“ because there is no way a machine could replicate the richness and complexity of a human's artistic talent. Nees, however, pointed out that a machine can do anything *if you can describe how to do it*. Computational tractability requires models to be expressed using the explicit and precise mathematical language of algorithms. The challenge is not that a computer cannot produce (or replicate) art, it is that art that defies reification into a formalized set of steps. Much of human action and understanding lives in what Michael Polanyi calls the *tacit dimension*,[146] which is best articulated by Polanyi's famous aphorism "we know more than we can tell." Indeed, we know more than we can tell computers.

The second effect of computing as described by McCarty involves an understanding of representational manipulability. When we describe a model and make it computationally tractable, we make it material (in a manner of speaking) and subject to, and the arbiter of, mechanical/computational manipulation. There is a deep sense of movement and change associated with computation; when we make our models tractable, we articulate a series of steps, an algorithm, for the computer. I sometimes like to jokingly think of computation as *math with motion*. But what is crucial to understand with respect to movement and manipulability is we do not *know* what will come out of a computational process *until it occurs*. McCarty connects this sense of movement to emergent understanding and knowledge. Models are, in McCarty's words, "*temporary states in a process of coming to know* rather than fixed structures of knowledge."[147]

[146] Polyani, Michael. "The tacit dimension." (1966).

[147] Emphasis in the original. McCarty, Willard. "Modeling: a study in words and meanings." A companion to digital humanities (2004): 254–70270 http://nora.lis.uiuc.edu:3030/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-7&toc.depth=1&toc.id=ss1-3-7&brand=9781405103213_brand

A World from a Topic Model's Perspective

First and foremost, it is important to understand the strange meanings *word, document,* and *topic* assume in the world of language modeling.

At the start of any text mining adventure, the natural sequences of words, the sentences and paragraphs of written documents are broken up via a process called *tokenization*. Individual words become *unigrams* or individually unique tokens. Tokens are not always equivalent to words because the tokenization process may count two or more words together as a single token, creating what are called *bigrams* or *ngrams*. For example, the words "digital humanities" could be a bigram or two individual unigrams, "digital" and "humanities." Tokenization is more of an art than a science; it requires subjective decisions as well as domain understanding of the texts being processed.

There are typically two additional preprocessing steps applied to tokenized text before we can partake in the joy of topic modeling. The first involves the removal of *stop words* and the second is *stemming*. I should note, there are other flavors of preprocessing, such as parts-of-speech tagging and removal, but I won't be covering them here.

Once the beautiful prosaic text has been sliced and diced, it contains tokens like "and," "but," or "or." These *stop words* are a wrench in the gears of bag-of-words language modeling producing incomprehensible or low-value output. Stop words lose their meaning once they have been decontextualized from their positions in the sequential order of the original texts. Stop words lists are often part of text mining or natural language processing software packages, posted on the web, or passed around from researcher to researcher. Alternatively, they might be generated for a specific corpus using techniques like *term frequency-inverse document frequency*[148] ranking, a technique that ranks a word's prevalence in individual documents against their prevalence across a corpus of documents.

[148] http://en.wikipedia.org/wiki/Tf-idf

This has the effect of giving words common across all documents, like stop words, a low ranking enabling the possibility of creating a corpus or domain specific stop word list by selecting all words with a score less than some specified value.

Stopwords illustrate a couple interesting, and sometimes problematic, assumptions in the preprocessing of texts. Consider the string of words "to be or not to be." This famous sequence of words is pregnant with meaning and implications, but in the eyes of a textual preprocessor it is completely mangled. When the phrase is tokenized and counted we end up with the following representation devoid of its original meaning: ["to":2, "be":2, "or":1, "not":1, "to"1]. Further, when we filter stop words, every word in that famous phrase is completely removed.

Once the stop words have been removed, there are still morphological problems with word tokens to be overcome. Basic tokenization and term frequency is going to count "model" and "models" as separate tokens. This can be a problem because we *want* these tokens to be counted together. Stemming is a process that trims word tokens down to their morphological roots. Different algorithms stem more or less aggressively. A lightweight stemmer might remove pluralization or other suffixes; a more aggressive stemmer cuts words back to their lexicographical root. One very popular and aggressive algorithm used in fulltext search and information retrieval, the Porter stemmer,[149] trims words to incomprehension; "example" becomes "exampl" and "courage" becomes "courag." Such aggressive stemming is generally not very useful for topic modeling because the topics become difficult to interpret because word's morphological roots may have different meanings.

Documents, in this strange ontological space, are not a sequence of words and punctuation as we might expect. Instead, documents are more like a word census; a sum totals of the number of times each word occurs in the original, natural document. Choosing exactly what unit of text will come to represent an individual

[149] Porter, Martin F. "An algorithm for suffix stripping."suffix stripping." 14.3 (1980): 130–37137.

document is a bit of an art form in topic modeling. Text’s natural partitions do not always yield the best results. For example, if you are modeling books, you might treat individual chapters or sections of a chapter as an individual document rather than the entire book. As usual, it is important to understand the nature of the text you are topic modeling to determine the appropriate unit of analysis.

To briefly recap, *words* are not words, as we typically understand them. Words in topic modeling are *unigrams, bigrams,* or *ngrams* that have been *tokenized, filtered, stemmed* and *counted*. A collection of word counts, that is, the *term frequencies*, represent individual *documents*. Collections of documents, a *corpus,* are seen from the perspective of the model not as a collection of text files filled with sequences of words, but rather as a *term-document matrix*.

Table 7: An example term-document matrix

Vocabulary	Document 1	Document 2	Document 3
humanities	8	4	0
digital	8	12	4
model	0	0	14
…	…	…	…

In the term-document matrix, each row represents a word token resulting in one row for every word in the corpus. This collection of words is called the *vocabulary*. Each column in the matrix represents a single document, as represented by a set of frequencies of the term in a particular row. Often it is the case the term-document matrix contains a lot of zero entries, that is, there are terms in the vocabulary that only show up in some documents but not others (and visa versa). Such a term-document matrix is considered to be *sparse*.

The term-document matrix is a data structure, a computationally tractable (to use McCarty's term) representation of the texts able to be modeled by a computational process. These preprocessing steps transforms a human readable sequence of words into a long list of word tokens, which are then counted for each individual document and (essentially) recorded in an Excel spreadsheet. Once the texts are represented as a matrix of numbers and all the messy human bits have been eliminated, the fun part, topic modeling, can begin!

Such processing is a boring and, I argue, a taken-for-granted assumption overlooked in many tutorials and introductions to topic modeling. Text preprocessing is an *infrastructural* process, vitally important, but also completely *ordinary* within the topic modeling *community of practice*.[150] Members of these spaces have already internalized and *normalized* many crucial knowledge practices making the process of socializing new members difficult. Such process need to be explicitly articulated, creating opportunities for what Lave and Wenger termed *legitimate peripheral participation* by new members of the practice. The practices of preprocessing text can, at first glance, seem alien to a humanities scholar versed in close reading, but as I will describe below, even bags-of-words can be used to find interesting patterns within texts.

OK, now we can talk about topics

Perhaps the most confusing aspect of topic modeling to a newcomer is the term "topic." Topic does *not* mean "a matter dealt with in text, discourse, or conversation" or "a subject" or anything a reasonable person might consider a "topic" if you asked them on the street. A topic, in the domain of language models, means a *probability distribution* over a vocabulary of words. This means, given a list of words, each has a specific value between zero and one (or

[150] Lave, Jean. "Situating learning in communities of practice." *Perspectives on socially shared cognition* 63 (1991): 82.

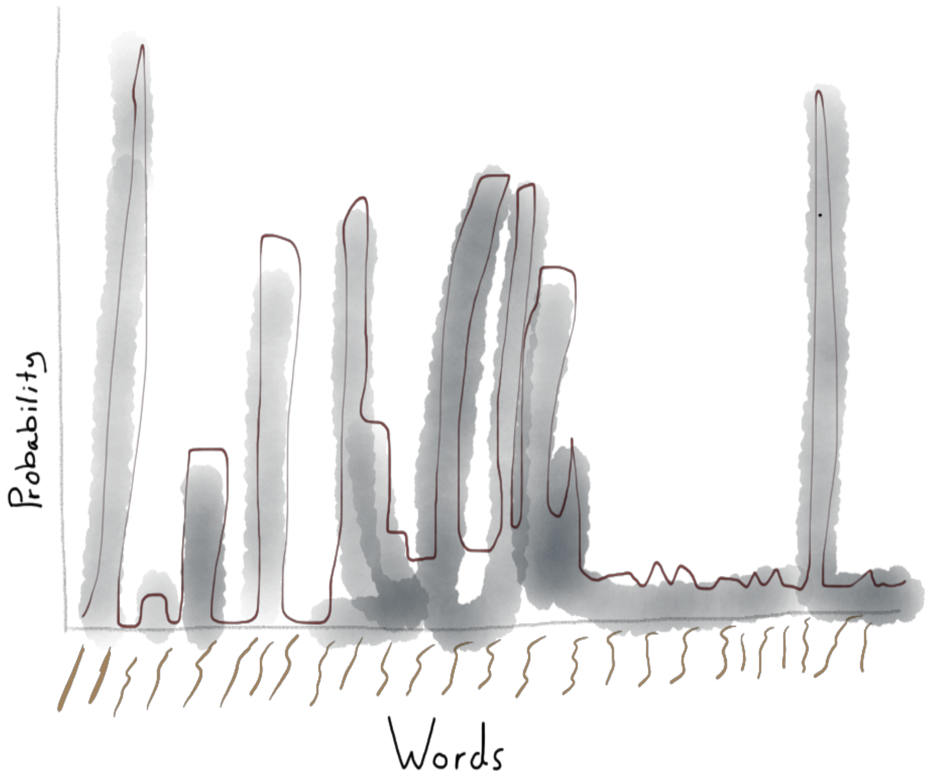
alternatively, 0 percent to 100 percent) associated with that word. The list of values represents an individual topic and different topics will (hopefully) have different values associated with each word.

One simplistic way to think about topic distributions would be as bags of words containing some varying allotment words. When I reach into the bag and pull out a word the likelihood I will pull out any particular word depends upon the allotment of words in the bag. However, the exact word you choose is unknown until you actually reach in the bag.

Table 8: An example topic (word distribution)

Word	Probability
`humanities`	`0.01`
`unigram`	`0.0004`
`digital`	`0.03`
`model`	`0.02`
`...`	`...`

For example, in the distribution shown in the table above, I would have a 1 percent likelihood of selecting the word `humanities`, a 3 percent probability of selecting the word `digital`, a 2 percent chance of selecting `model` and miniscule (.04 percent) chance of selecting `unigram`. Also important, and not necessarily intuitive, is that each selection of a word is *independent* so my selections do not affect any subsequent selections, *even of the same word*. This would mean if my topic distribution assigned 99 percent probability to the word `computer` I will most likely select the word `computer` every time I draw from the distribution.



In this non-artist's rendition of a topic, the brown squiggles along the bottom represent a vocabulary of words and the grey peaks represent individual word's *probability density*. I should note, it is very unlikely you might find a topic like the one above, with such dramatic peaks and valleys, in the wild. In my (limited) experience the topic distributions are relatively flat with some small clusters of words having a bit more weight than others. The list of *top words*, words that are "heavy" with probabilistic mass, are the interesting group of words; they are the co-occurring words in a topic distribution.

Now that you (hopefully) understand what words, documents, and topic means from the perspective of a topic model it is time to discuss the generative models themselves.

A Brief History of Generative Topic Models

One of the best ways to understand the assumptions of generative language models is to start with simplistic models and then work up to modern topic modeling techniques like LDA. I am drawing heavily here upon Blei et al.[151] and section four of the original LDA paper, but instead of contrasting these models with LDA, I want to build up our understanding of each model through the innovations they introduced. Starting with the simple *unigram model*, to the *mixture of unigrams*, to *probabilistic latent semantic analysis*, to *latent dirichlet allocation*. Each model rests upon a complex mathematical foundation; I am going to gloss over the math and focus more upon intuitive, but not overly simplistic, descriptions of each model's assumptions.

As I discussed above, these are *generative* models. Each represents generative process that repeats on a loop, selecting word tokens from a probabilistic bag-of-words (topics) and generating unique documents from increasingly complex combinations and mixtures of these bags. The models generate words, topics, and documents as I have just explained above, not the infinitely rich structures of writing and language you are reading right now.[152] With each model I am including a representation in plate notation, a way of visually representing graphical models, and a description of the generative procedure in pseudocode. In the plate notation, a square means a looping, repeating process and a square within a square means nested loops. Circles represent variables, the shaded circles are *observed* variables (things we have) and the white circles are *latent* (things we assume are there). Topics are always latent variables, white circles, because in these models assume the existence of topics and make them a set of variables to estimate. In both cases I have attempted to simplify these representations to make them slightly less intimidating to someone unfamiliar with such forms of notation. I have used English descriptions of variables instead of Greek characters to reduce complexity.

[151] Blei, David M, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation." *the Journal of Machine Learning Research* 3 (2003): 993–1022.

[152] This of course assumes my writing is "infinitely rich" and "complex."

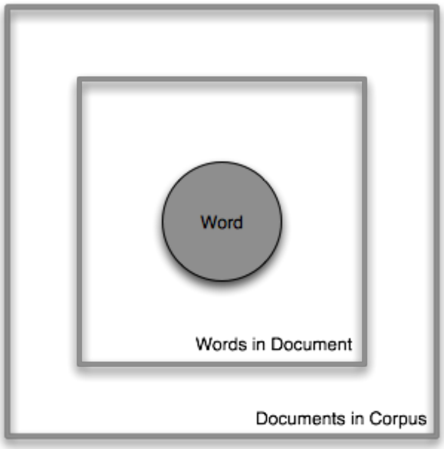
Unigram

One of the most simplistic language models, although not always considered a "topic model," is the *unigram language model*. This model uses a single topic in the entire corpus. Each document in the corpora is composed of some number of words selected from a single topic distribution for the entire corpus.

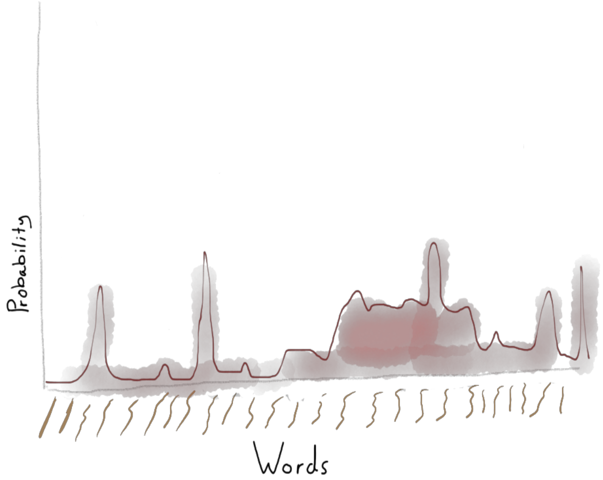
The generative process of the unigram model is described in pseudo code below:

```
For each document in the corpus do the following:
    For each word in the document do the following:
        Select a word from the word distribution.
```

Intuitively, the model generates a document by repeatedly selecting words from a single word distribution, i.e., topic. Each word selection is independent from the words selected before and after, which means, given a word distribution where one word is highly likely, that word will frequently show up in any generated document. For example, if your word distribution, your bag of words, is about food it might assign more weight to the word "pizza." When you draw however many words from the distribution you want to "compose" your document, chances are you will draw several instances of the word "pizza." Because I only have one topic distribution, that is, one bag of words, the kinds documents and corpora I can represent with this model are, probabilistically, not very diverse. The chances I will be able to generate a document about "automobiles," is, probabilistically, less likely if I am given a word distribution weighted in favor of food.



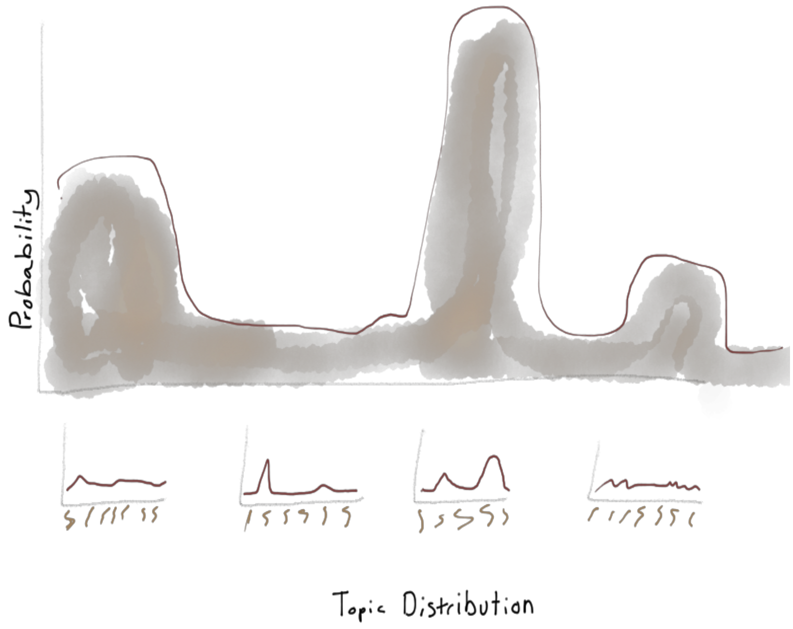
The image above describes the model using plate notation. The outer square represents iteration over every single document. The inner square represents iteration over every word for each document. The grey circle in the middle represents the *observed* variable; in this case the words in each document. The shaded circle is the observed word token we select from the word distribution, a.ka. topic (*below)*, and it is encapsulated by two squares meaning it is nested within two loops. The outer square loops over every document in the corpus and the inner square loops over every word in a document.



In the unigram model, the bag never changes as we select word after word and compose document after document. Only accommodating a single topic distribution limits the unigram model's capacity to effectively model the complexity and richness of many human authored corpora. This is not to say the unigram model is not useful, it has been used to great effect in information retrieval, but its effectiveness as a topic model is low.

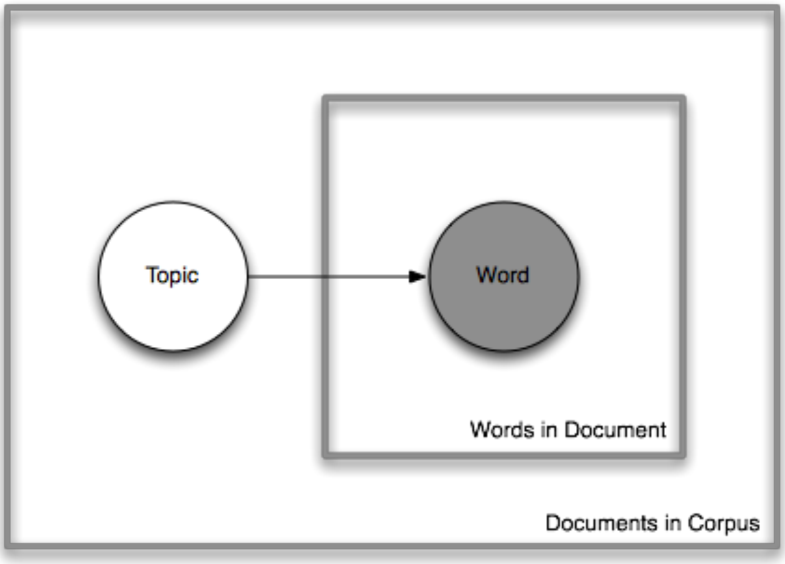
Mixture of Unigrams

The *mixture of unigrams* model introduces the possibility of multiple *topics,* that is, more than one bag from which to draw words when generating documents. The mixture of unigram model introduces a new distribution, a distribution of *topics*, from which we draw a new distribution of words for each document.



The generative process of the mixture of unigrams model is described below:

```
For each document in the corpus do the following:
       Select a distribution of words (topic) from a distribution of topics.
       For each word in the document do the following:
              Select a word from that distribution of words.
```



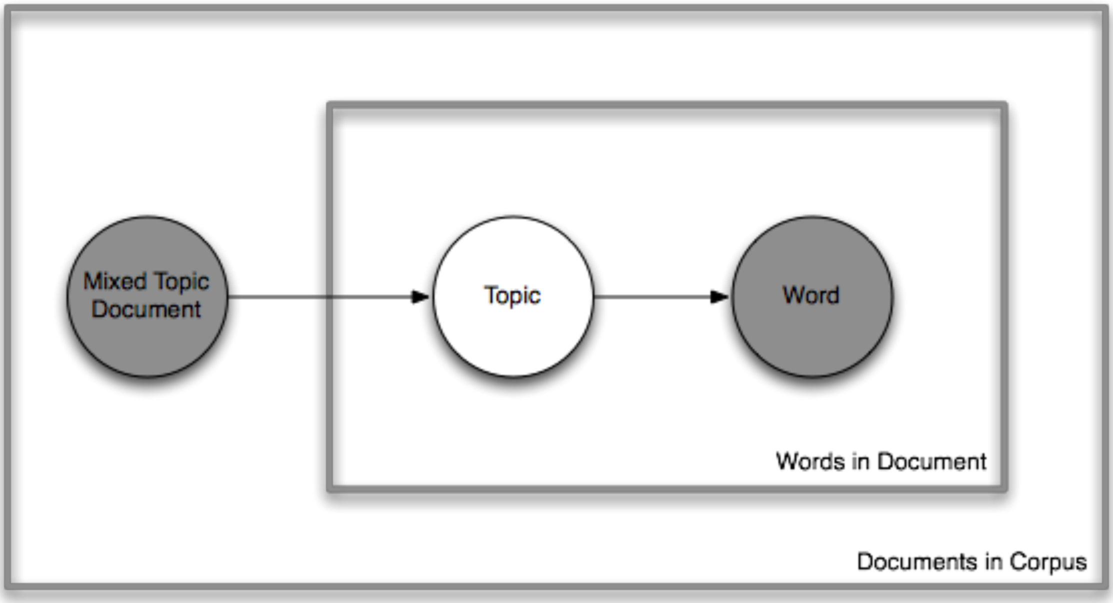
The mixture of unigrams model is represented in plate notation above. The mixture of unigrams adds a new *latent,* or unobserved, variable to the model that represents the topic, the word distribution, from which each document will be drawing words. Because the latent variable is outside the inner square, there is only one topic per document. This is better than the unigram model, which only allows one topic per corpus. This adds a bit more diversity to the model of the corpus, but not necessarily much diversity to any individual document. This means, while the corpus might be about food and books, a single document is about either food or books, but not both.

Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) also called Latent Semantic Indexing, introduced in 1999 by Thomas Hofmann, was one of the early and popular topic models. Hofmann's model introduced several novel innovations over the simplistic techniques I described above. Like the mixture of unigrams model, PLSA models multiple topics or word distributions in the corpus, but, unlike the mixture of unigrams, PLSA allows individual documents to be composed of multiple topics. PLSA does this by sampling a distribution of topics each time we draw a word, instead of each time we create a document. The generative process of the PLSA model is described below:

```
For each document in the corpus do the following:
        For each word in the document do the following:
                Select a distribution of words from the distribution of topics.
                        Select a word from that distribution of words.
```

Notice in the plate notation below how the inner square, the "words in document" iteration, has expanded to encompass the latent topic variable. The arrows indicate a dependency; before a word is drawn from a topic, a new topic must be drawn from a distribution. Each document has it's own unique distribution or mixture of topics. This allows individual documents to be composed of words drawn from multiple topics; a more plausible model of a document's reality.



However, as Blei et al. point out, the ways in which the document mixtures are created are prone to *overfitting*, that is, the mode by which an individual document's topic mixture is established is not robust enough to handle the addition of new documents to the corpus after the model has been generated, or *trained* in machine learning terms. Overfitting can be a real problem if you are using topic models to work with new documents, for example, using topic models to generate recommendations in a scholarly journal database. If you initially train your PLSA topic model on the articles you have, as you receive new articles the recommendations will get progressively worse unless you retrain using the entire updated corpus. For very large corpora, this can be prohibitively expensive computationally.

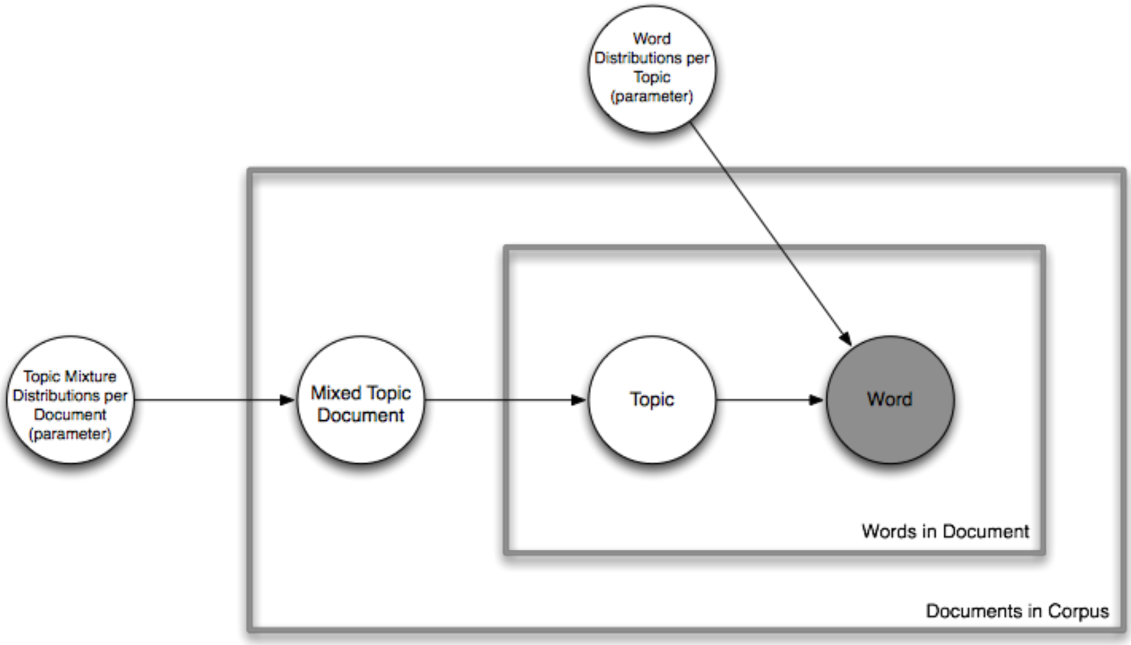
The Overfitting Problem

The problem of overfitting marks an interesting distinction between how computer scientists and digital humanists might use topic modeling. One of the benefits of LDA over PLSA is, as I describe below, a robust method for generating a document's topic mixture. This feature allows a model trained on an existing corpus to identify the topic mixture of new documents without retraining the entire corpus. When, as is trendy in computer science these days, you start talking about “big data,” that is, massive corpora such as the Google Books dataset, training a model becomes computationally expensive.

However, in the digital humanities, our corpora are often (but not always) meso-scale, or as I like to put it, “bigger than a laptop smaller than a large hadron collider.” Furthermore, it is often the case there will *never* be any additional documents in our corpora. There is never going to be any *new* nineteenth-century British and American literature. I acknowledge this is a grossly simplistic assumption about literary history and the complexities of digitization, but once a historical collection has been fully digitized it should be reasonable not to expect new documents in the corpus. Thus, if the text model you are generating is exclusively for the purposes of exploring a fixed corpus, is overfitting a problem?

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is very similar to PLSA. If you look at the the plate notation LDA there are only two additional, though very important, latent variables added to the model.


Word Distributions per Topic (parameter)
Topic Mixture Distributions per Document (parameter)
Mixed Topic Document
Topic
Word
Words in Document
Documents in Corpus

These two corpus-level parameters introduce a Bayesian method for sampling the mixture of topics within each document. Essentially this means random sampling but not sampling just any old tea leaves or turtle shells. LDA draws randomly from a parameterized Dirichlet distribution producing, through the magic of mathematics, robust topic mixtures and word distributions able to overcome the overfitting problems of PLSA. Additionally there are fewer parameters to estimate, which is important when training the model. The generative process of the LDA model is described below:

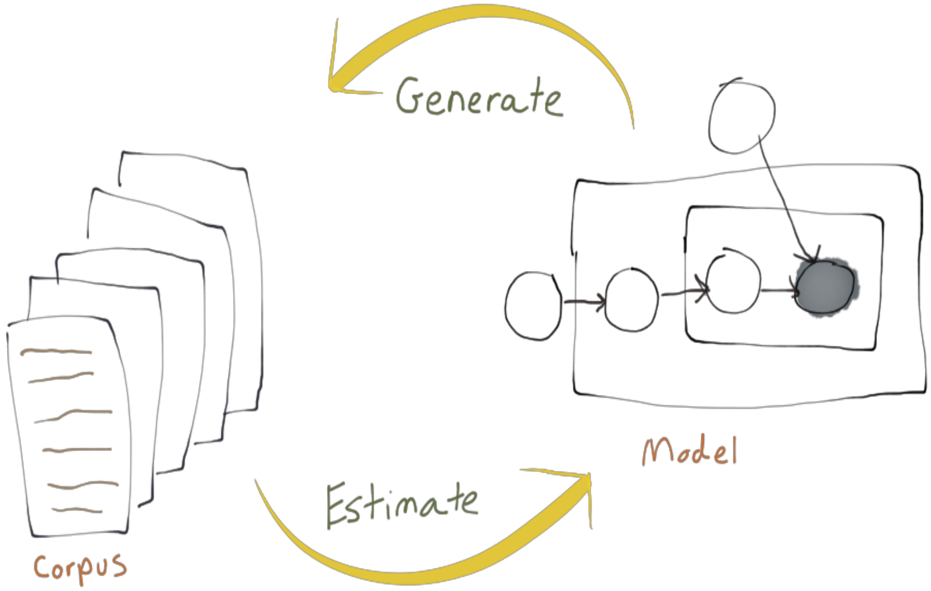
```
For each document in the corpus do the following steps:
      Select a topic mixture distribution from a Dirichlet distribution.
      For each word in the document do the following steps:
            Select a topic from the topic mixture distribution.
            Select a word from the word distribution selected above.
```

LDA describes a generative process whereby, given a Dirichlet *conditioned* bag filled with topic distributions for each document, we draw a topic mixture from this bag. Then, we repeatedly draw both a topic and then a word from that topic to generate the words in that document. Voila, we have a generative model that represents the process by which a corpus of documents might be created. What if we already have a corpus of documents?

Parameter Estimation

Everything I have described so far, about the structure and underlying assumptions of generative topic models is how topic modeling works *in theory*. When we use topic modeling to model a corpus of texts, what we are practically trying to do is *estimate* the parameters of the model as I have described. That is, we are trying to find a model whose parameters have a high likelihood of generating the corpus *if* we were to use a generative process to create the corpus we have.

This is a *very* important distinction between topic modeling as theoretically understood and topic modeling in implementation (practice). The models I have described, in as plain of english as I can muster, are theoretical articulations of a generative process. Given a set of parameters, for LDA this would be word distributions (topics) and a topic mixtures, the process would repeatedly sample these distributions to generate a term-frequency matrix, i.e., a corpus. However, this is not how we use topic modeling in practice. Instead of having the parameters for these various distributions a priori we have, after text preprocessing, a corpus that has been generated by some topic model. The goal of topic modeling in practice is to find the model, that is, find the document topic mixtures and word distributions that generated the corpus you have.



The secret to successful topic modeling is *estimating* the distributions from the set of all possible distributions (an *extremely* large space impossible to fully enumerate) that best fits the corpus of documents at hand. This process, called parameter estimation, is where much of the mathematical complexity in topic modeling lives. There are many ways to estimate the parameters; the original LDA paper used a process called *variational inference* and the MALLET toolkit uses a process called *Gibbs Sampling*. David Mimno's talk at the MITH Topic

Modeling workshop[153] is an excellent discussion on how exactly he uses Gibbs sampling to estimate the parameters of an LDA topic model in the MALLET toolkit.[154]

My goal here was simply to unpack, in detail, a non-mathematical description of the LDA generative model in hopes that others will better understand how it is exactly that we can use these techniques to explore and understand bodies of text too large to simply read by hand. By understanding these underlying assumptions of generative language models we can first and foremost be better informed about the kinds of claims we make when we use them, but also potentially contribute in making even more robust and pragmatically useful language models for future digital humanists.

What Can Topic Modeling Tell Us?

To ground this discussion, I provide some example output from the MALLET toolkit. Listed below are the top ten words from four of the ten topics I estimated based on a corpus of blog posts from Digital Humanities Now's Editor Choice selections

```
Topic

0: students education cr learning student free courses class university higher

1: library access digital content public libraries future google art impact

2: data visualization information objects mining http heritage open april big

3: knowledge thinking history historical human point kind understanding place
creating
```

I fed MALLET a set of text files, a number of iterations, and a number of topics. MALLET tokenized my text, removed stop words (the toolkit does not perform stemming for reasons articulated by the author on MALLET mailing list[155]), and

[153] http://journalofdigitalhumanities.org/2-1/the-details-by-david-mimno/

[154] http://MALLET.cs.umass.edu/topics.php

[155] http://comments.gmane.org/gmane.comp.ai.MALLET.devel/1724

estimated the word distribution for ten topics and the topic mixtures for each document in the corpus.

While mainly a science, topic modeling has aspects of an art form. There are several parameters that we must specify before estimating the model. The most significant of these parameters is the number of topics. In the example above, I have selected ten topics. The number of topics is a subjective selection dependent upon the size and shape of the corpus.

Each document is associated to each topic by some proportion. Just as every topic has a ranked probabilities of words, every document has a ranked probability of topics. Thus, while every document might have some trace of every topic, generally we are only interested in the top one, two or three topics associated with each document. It is fairly common, when analyzing the models, to set some frequency threshold for the document/topic relation (say, 10 percent) so that you attend to the topics best represented in a document (or conversely, the top documents in the topic in question).

```
Document: file1.txt

Topic Probability

3      0.3986013986013986

4      0.12665112665112666

1      0.11888111888111888

2      0.07459207459207459

7      0.07381507381507381

6      0.05439005439005439

8      0.05128205128205128

5      0.04895104895104895

0      0.041181041181041184

9      0.011655011655011656
```

In the example above, the topic proportions for file1.txt are ranked from highest to lowest. Topic 3 is the most prominent with a proportion of 39 percent, followed by topic 4 at 12 percent and topic 1 with 11 percent. Topic 0 and topic 9 are the lowest with 4 percent and 1 percent respectively. The document in question is a blog post by Peter Organisciak, a graduate student at the University of Illinois who was actually one of the founders of the "Day of DH" blogging project.[156] The post begins with this self-reported summary of its content:

> Last month, I gave a presentation about paid crowdsourcing in the humanities at SDH-SEMI. Below are my notes.[157]

So the post is about "paid crowdsourcing in the humanities" and according to the topic model, the topic with the highest proportion, topic 3, contains these top words:

```
books time people texts make terms research don work things simply
sense ways fact change early process read human
```

Given a very cursory analysis of this unrefined model, I think there is some sense to be made from this topic. Crowdsourcing is all about taking advantage of *people*'s free *time* to do certain kinds of *work,* generally *simple* tasks, often for the purposes of *research*. Obviously, to do this analysis justice I would want to go back and see how these terms are used in the original text. Additionally, I would probably want to tweak my model to include more or less topics depending on the dynamics of the corpus.

There are several implementations of the Latent Dirichlet Allocation available to a researcher interested in topic modeling. LDA-C is perhaps one of the most widely known; it was implemented by David Blei using the C programming language. Perhaps the other most popular implementation of LDA is part of a java toolkit, MALLET, maintained by Andrew McCallum at the University of Massachusetts Amherst.

Given that topics are merely lists of words, any topic modeling exercise requires some *interpretive* effort to discern if the model is a reasonable representation of

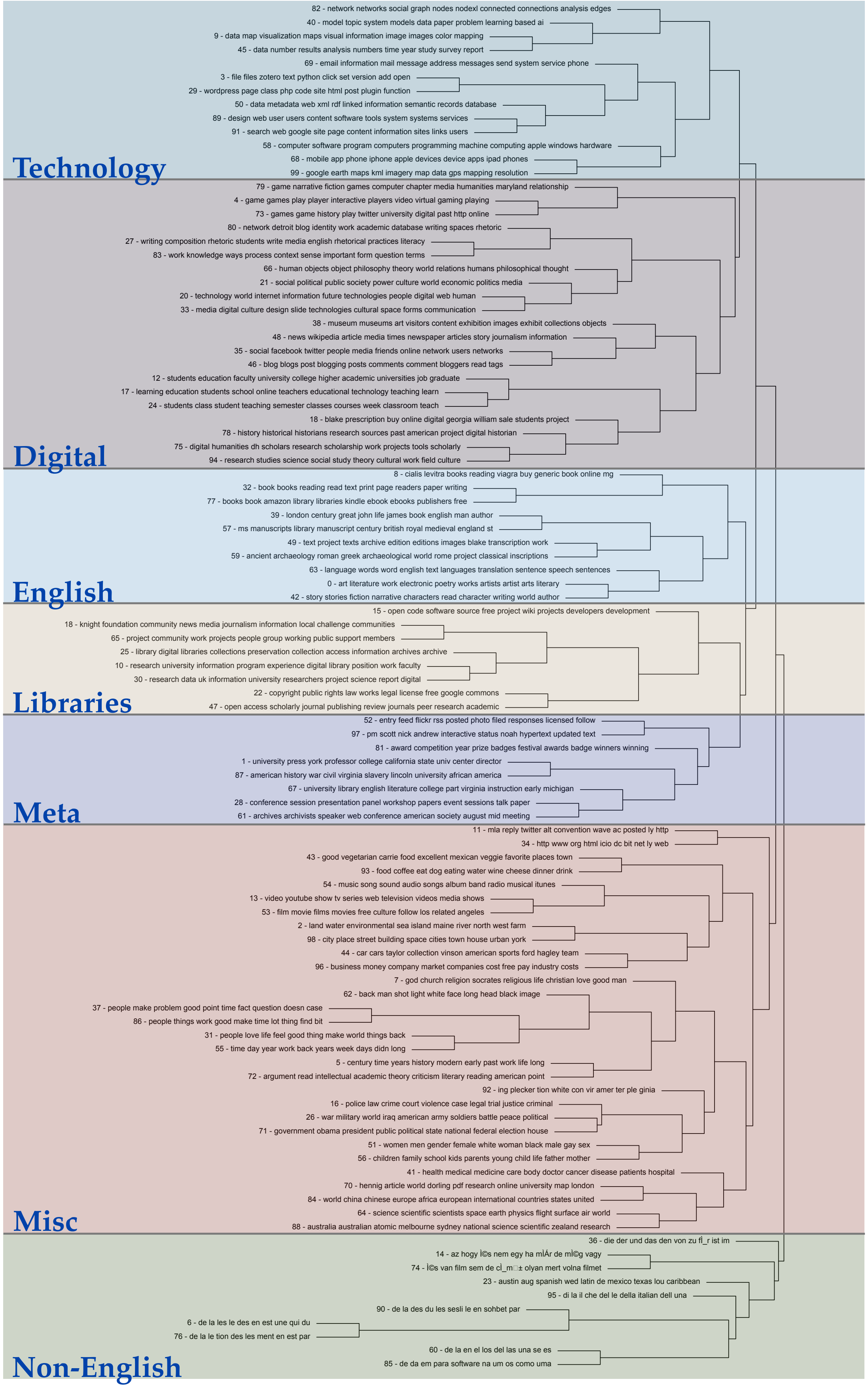
[156] http://tapor.ualberta.ca/taporwiki/index.php/Day_in_the_Life_of_the_Digital_Humanities

[157] http://www.porganized.com/blog/a-modest-payment-for-a-modern-proposal

the corpus and what that representation means. Thus, a close and careful reading of the relationship between topics and documents is necessary to fully understand and contextualize what the words of a topic really *mean*.

Topic modeling clusters sets of documents according to latent themes and provides a set of keywords associated with that theme. Reading topic models then is an exercise in reading the documents with high proportions for each topic paying special attention to how the set of keywords are used both within and across those high-proportion documents. In a sense we might think about reading these documents as editorialized selections, but unfortunately, the editor who put them together has been mysteriously struck with amnesia and all we have are a list of underlined words in each document. With this information we must engage in a semi-hermeneutic exercise of constructing the latent meanings beneath the surface of this scraps of information.

Appendix B — Topic Clusters


82 - network networks social graph nodes nodexl connected connections analysis edges
40 - model topic system models data paper problem learning based ai
9 - data map visualization maps visual information image images color mapping
45 - data number results analysis numbers time year study survey report
69 - email information mail message address messages send system service phone
3 - file files zotero text python click set version add open
29 - wordpress page class php code site html post plugin function
50 - data metadata web xml rdf linked information semantic records database
89 - design web user users content software tools system systems services
91 - search web google site page content information sites links users
58 - computer software program computers programming machine computing apple windows hardware
68 - mobile app phone iphone apple devices device apps ipad phones
99 - google earth maps kml imagery map data gps mapping resolution
Technology
79 - game narrative fiction games computer chapter media humanities maryland relationship
4 - game games play player interactive players video virtual gaming playing
73 - games game history play twitter university digital past http online
80 - network detroit blog identity work academic database writing spaces rhetoric
27 - writing composition rhetoric students write media english rhetorical practices literacy
83 - work knowledge ways process context sense important form question terms
66 - human objects object philosophy theory world relations humans philosophical thought
21 - social political public society power culture world economic politics media
20 - technology world internet information future technologies people digital web human
33 - media digital culture design slide technologies cultural space forms communication
38 - museum museums art visitors content exhibition images exhibit collections objects
48 - news wikipedia article media times newspaper articles story journalism information
35 - social facebook twitter people media friends online network users networks
46 - blog blogs post blogging posts comments comment bloggers read tags
12 - students education faculty university college higher academic universities job graduate
17 - learning education students school online teachers educational technology teaching learn
24 - students class student teaching semester classes courses week classroom teach
18 - blake prescription buy online digital georgia william sale students project
78 - history historical historians research sources past american project digital historian
75 - digital humanities dh scholars research scholarship work projects tools scholarly
94 - research studies science social study theory cultural work field culture
Digital
8 - cialis levitra books reading viagra buy generic book online mg
32 - book books reading read text print page readers paper writing
77 - books book amazon library libraries kindle ebook ebooks publishers free
39 - london century great john life james book english man author
57 - ms manuscripts library manuscript century british royal medieval england st
49 - text project texts archive edition editions images blake transcription work
59 - ancient archaeology roman greek archaeological world rome project classical inscriptions
63 - language words word english text languages translation sentence speech sentences
0 - art literature work electronic poetry works artists artist arts literary
42 - story stories fiction narrative characters read character writing world author
English
15 - open code software source free project wiki projects developers development
18 - knight foundation community news media journalism information local challenge communities
65 - project community work projects people group working public support members
25 - library digital libraries collections preservation collection access information archives archive
10 - research university information program experience digital library position work faculty
30 - research data uk information university researchers project science report digital
22 - copyright public rights law works legal license free google commons
47 - open access scholarly journal publishing review journals peer research academic
Libraries
52 - entry feed flickr rss posted photo filed responses licensed follow
97 - pm scott nick andrew interactive status noah hypertext updated text
81 - award competition year prize badges festival awards badge winners winning
1 - university press york professor college california state univ center director
87 - american history war civil virginia slavery lincoln university african america
67 - university library english literature college part virginia instruction early michigan
28 - conference session presentation panel workshop papers event sessions talk paper
61 - archives archivists speaker web conference american society august mid meeting
Meta
11 - mla reply twitter alt convention wave ac posted ly http
34 - http www org html icio dc bit net ly web
43 - good vegetarian carrie food excellent mexican veggie favorite places town
93 - food coffee eat dog eating water wine cheese dinner drink
54 - music song sound audio songs album band radio musical itunes
13 - video youtube show tv series web television videos media shows
53 - film movie films movies free culture follow los related angeles
2 - land water environmental sea island maine river north west farm
98 - city place street building space cities town house urban york
44 - car cars taylor collection vinson american sports ford hagley team
96 - business money company market companies cost free pay industry costs
7 - god church religion socrates religious life christian love good man
62 - back man shot light white face long head black image
37 - people make problem good point time fact question doesn case
86 - people things work good make time lot thing find bit
31 - people love life feel good thing make world things back
55 - time day year work back years week days didn long
5 - century time years history modern early past work life long
72 - argument read intellectual academic theory criticism literary reading american point
92 - ing plecker tion white con vir amer ter ple ginia
16 - police law crime court violence case legal trial justice criminal
26 - war military world iraq american army soldiers battle peace political
71 - government obama president public political state national federal election house
51 - women men gender female white woman black male gay sex
56 - children family school kids parents young child life father mother
41 - health medical medicine care body doctor cancer disease patients hospital
70 - hennig article world dorling pdf research online university map london
84 - world china chinese europe africa european international countries states united
64 - science scientific scientists space earth physics flight surface air world
88 - australia australian atomic melbourne sydney national science scientific zealand research
Misc
36 - die der und das den von zu fÌ_r ist im
14 - az hogy Ì©s nem egy ha mÌÁr de mÌ©g vagy
74 - Ì©s van film sem de cÌ_m□± olyan mert volna filmet
23 - austin aug spanish wed latin de mexico texas lou caribbean
95 - di la il che del le della italian dell una
90 - de la des du les sesli le en sohbet par
6 - de la les le des en est une qui du
76 - de la le tion des les ment en est par
60 - de la en el los del las una se es
85 - de da em para software na um os como uma
Non-English

Appendix C — Topic Labels

Topic labels were assigned through qualitative coding the topic's keywords.

Topic	*Label*
0	Digital Media
1	Conferences
2	Environmental History
3	How To Tools
4	Games
5	Time
6	non-english
7	Religion
8	spam
9	Data Visualization
10	Job Postings
11	Online Activity
12	Meta-Humanities
13	Media Studies
14	non-english
15	Library Tech
16	Justice
17	Learning
18	spam
18	Journalism
20	junk
21	Social Theory
22	Intellectual Property
23	Latin Studies
24	Managing Classroom
25	Digital Preservation at LOC
26	War
27	Writing
28	CFPs
29	Wordpress
30	Science
31	Personal
32	Books
33	Urban Design
34	Library Tech
35	Social Media
36	non-english
37	People
38	Museums
39	Brit Lit
40	Humanities Computing
41	Health
42	Stories
43	Personal
44	Car Museum
45	Humanities Computing
46	Blogs
47	Open Access
48	Journalism
49	William Blake Archive
50	Library Tech
51	Identity Politics
52	junk
53	Film Studies
54	Sound Studies
55	Personal
56	Family
57	British manuscripts
58	Computing
59	Ancient History

60	non-english
61	Archives
62	Film Studies
63	junk
64	Pop Science
65	Project Management
66	OOO & Philosophy
67	junk
68	junk
69	junk
71	junk
72	Critical Theory
73	Games
74	non-english
75	What is DH
75	non-english
76	non-english
77	eBooks
78	junk
79	junk
80	junk
81	Competitions
82	Network Analysis
83	Knowledge Work
84	American History
85	non-english
86	Knowledge Work
87	History
88	junk
89	UxD
90	non-english
91	SEO
92	junk
93	Food
94	Humanities & Social Science
95	non-english
96	Business
97	Online Activity
98	Personal
99	Google Earth

Works Cited

Abbate, Janet. 2000. *Inventing the Internet*. MIT Press. http://books.google.com/books?id=E2BdY6WQo4AC.

Al-Ani, Ban, Gloria Mark, Justin Chung, and Jennifer Jones. 2012. “The Egyptian Blogosphere: A Counter-Narrative of the Revolution.” In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 17–26. CSCW ’12. New York, NY, USA: ACM. doi:10.1145/2145204.2145213.

Arun, R., V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. 2010. “On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations.” In *Advances in Knowledge Discovery and Data Mining*, edited by Mohammed J. Zaki, Jeffrey Xu Yu, B. Ravindran, and Vikram Pudi, 391–402. Lecture Notes in Computer Science 6118. Springer Berlin Heidelberg.

Avery, Paul.. 2007. “Open Science Grid: Building and Sustaining General Cyberinfrastrucutre Using a Collaborative Approach.”

Bamman, David, and Noah A. Smith. 2013. “Learning Latent Personas of Film Characters.” ACL, 2013.

Beaulieu, Anne. 2004. “Mediating Ethnography: Objectivity and the Making of Ethnographies of the Internet.” *Social Epistemology* 18 (2-3): 139–63. doi:10.1080/0269172042000249264.

———. 2010. “Research Note: From Co-Location to Co-Presence: Shifts in the Use of Ethnography for the Study of Knowledge.” *Social Studies of Science* 40 (3): 453–70. doi:10.1177/0306312709359219.

Berners-Lee, Tim. 1989. *Information Management: A Proposal.* http://www.w3.org/History/1989/proposal.html.

———. 2000. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. 1 edition. San Francisco: HarperBusiness.

Berry, David M., ed. 2012. *Understanding Digital Humanities*. Palgrave Macmillan.

Bijker, Wiebe E. 1997. *Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change*. The MIT Press.

Binder, Jeffrey M., and Collin Jennings. 2014. “Visibility and Meaning in Topic Models and 18th-Century Subject Indexes.” *Literary and Linguistic Computing*, May, fqu017. doi:10.1093/llc/fqu017.

Blei, David M., and J. D. Lafferty. 2007. “A Correlated Topic Model of Science.” *The Annals of Applied Statistics*, 17–35. http://www.jstor.org/stable/10.2307/4537420.

Blei, David M., A. Y. Ng, and M. I. Jordan. 2003. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research* 3: 993–1022. http://dl.acm.org/citation.cfm?id=944937.

Blei, David M. 2013. “Topic Modeling and Digital Humanities.” *Journal of Digital Humanities*. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/.

Blei, David M., and John D. Lafferty. 2006. “Dynamic Topic Models.” In, 113–20. ACM Press. doi:10.1145/1143844.1143859.

Blevins, Cameron. 2010. “Topic Modeling Martha Ballard’s Diary Cameron Blevins.” http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/.

Boellstorff, Tom, Bonnie Nardi, Celia Pearce, T. L. Taylor, and George E. Marcus. 2012. *Ethnography and Virtual Worlds: A Handbook of Method*. Princeton: Princeton University Press.

Bogost, Ian, and Nick Montfort. “Platform Studies: Frequently Questioned Answers.” *Digital Arts and Culture 2009*, December 12, 2009. http://escholarship.org/uc/item/01r0k9br.

Bonilla, Tabitha, and Justin Grimmer. 2013. “Elevated Threat Levels and Decreased Expectations: How Democracy Handles Terrorist Threats.” *Poetics*, Topic Models and the Cultural Sciences, 41 (6): 650–69. doi:10.1016/j.poetic.2013.06.003.

Borgman, Christine L. 2009. “The Digital Future Is Now: A Call to Action for the Humanities.” *Digital Humanities Quarterly* 3 (4). http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html.

Borgman, Christine L. 1989. “Bibliometrics and Scholarly Communication Editor’s Introduction.” *Communication Research* 16 (5): 583–99. doi:10.1177/009365089016005002.

———. 2000. “Digital Libraries and the Continuum of Scholarly Communication.” *Journal of Documentation* 56 (4): 412–30. doi:10.1108/EUM0000000007121.

———. 2007. *Scholarship in the Digital Age*. MIT Press. http://books.google.com/books?id=ZDDu3CuzDdMC.

Borgman, Christine L., and Jonathan Furner. 2002. “Scholarly Communication and Bibliometrics.” *Annual Review of Information Science and Technology* 36 (1): 2–72. doi:10.1002/aris.1440360102.

Bowker, Geoffrey C., and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. The MIT Press.

Bowker, Geoffrey C., Karen Baker, Florence Millerand, David Ribes, Jeremy Hunsinger, Lisbeth Klastrup, and Matthew Allen. 2010. "Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment." In *International Handbook of Internet Research*, 97–117. Springer Netherlands. http://link.springer.com/chapter/10.1007/978-1-4020-9789-8_5.

Bowman, Timothy D., Bradford Demarest, Scott B. Weingart, Grant Leyton Simpson, Vincent Lariviere, Mike Thelwall, and Cassidy R. Sugimoto. 2013. "Mapping DH Through Heterogeneous Communicative Practices." *Digital Humanities*. http://did.ils.indiana.edu/dh/pdf/DH2013.MappingDH.pdf.

Brand, S. (1994). *How Buildings Learn*. New York: Viking.

Brett, Megan R. 2013. "Topic Modeling: A Basic Introduction." *Journal of Digital Humanities*. http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/.

Brown, J. S, and P. Duguid. 1996. "The Social Life of Documents." *First Monday* 1 (1-6).

Brown, John Seely, and Paul Duguid. 2002. *The Social Life of Information*. 1st ed. Harvard Business School Press.

Bruns, Axel, and Joanne Jacobs, eds. 2006. *Uses of Blogs*. New York: Peter Lang International Academic Publishers.

Buckland, Michael K. 1991. "Information as Thing." *JASIS* 42 (5): 351–60. http://skat.ihmc.us/rid=1KR7VC4CQ-SLX5RG-5T39/BUCKLAND(1991)-informationasthing.pdf.

Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp. 2012. *Digital_Humanities*. The MIT Press.

Busa, R. 2004. "Foreword: Perspectives on the Digital Humanities." In *A Companion to Digital Humanities*, 1–19.

Castells, Manuel. 2011. *Communication Power*. 2 edition. Oxford: Oxford University Press.

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

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Advances in Neural Information Processing Systems*, 288–96. http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2009_0125.pdf.

Chun, Wendy Hui Kyong, and Lisa Marie Rhody. 2014. “Working the Digital Humanities: Uncovering Shadows Between the Dark and the Light.” *Differences* 25 (1): 1–25. doi:10.1215/10407391-2419985.

Clark, Vicki L. Plano, and John W. Creswell. 2008. *The Mixed Methods Reader*. SAGE.

Collins, H. M. 1974. “The TEA Set: Tacit Knowledge and Scientific Networks.” *Science Studies* 4 (2): 165–85.

Crane, Diana. 1969. “Social Structure in a Group of Scientists: A Test of the ‘Invisible College’ Hypothesis.” *American Sociological Review* 34 (3): 335–52. doi:10.2307/2092499.

———. 1972. “The Invisible College.” *Chicago: Univ. of Chicago Press* 12.

Creswell, John W. 2009. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE.

Creswell, John W., and Vicki L. Plano Clark. 2010. *Designing and Conducting Mixed Methods Research*. SAGE Publications.

Cronin, Blaise. 2003. “Scholarly Communication and Epistemic Cultures.” *New Review of Academic Librarianship* 9 (1): 1–24. http://www.tandfonline.com/doi/abs/10.1080/13614530410001692004#.Us8CxGRDuoU.

———. 2001a. “Bibliometrics and Beyond: Some Thoughts on Web-Based Citation Analysis.” *Journal of Information Science* 27 (1): 1–7. doi:10.1177/016555150102700101.

———. 2001b. “Hyperauthorship: A Postmodern Perversion or Evidence of a Structural Shift in Scholarly Communication Practices?” *Journal of the American Society for Information Science and Technology* 52 (7): 558–69. doi:10.1002/asi.1097.

Cronin, Blaise, Herbert W. Snyder, Howard Rosenbaum, Anna Martinson, and Ewa Callahan. 1998. “Invoked on the Web.” *Journal of the American Society for Information Science* 49 (14): 1319–28.

Çelik, Tantek. 2010. “What Is the Open Web.” http://tantek.com/2010/281/b1/what-is-the-open-web.

Dasu, Tamraparni, and Theodore Johnson. 2003. “Data Quality.” In *Exploratory Data Mining and Data Cleaning*, 99–137. John Wiley & Sons, Inc.

Davidson, Cathy N. 2008. “Humanities 2.0: Promise, Perils, Predictions.” *PMLA* 123 (3): 707–17.

De Solla Price, D.J., and D. Beaver. 1966. “Collaboration in an Invisible College.” *American Psychologist* 21 (11): 1011.

Dean, Jodi. 2010. *Blog Theory: Feedback and Capture in the Circuits of Drive*. 1 edition. Cambridge, UK; Malden, MA: Polity.

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. "Indexing by Latent Semantic Analysis." *JASIS* 41 (6): 391–407. http://www.cob.unt.edu/itds/faculty/evangelopoulos/dsci5910/LSA_Deerwester1990.pdf.

Dietz, Laura, Steffen Bickel, and Tobias Scheffer. 2007. "Unsupervised Prediction of Citation Influences." In *Proceedings of the 24th International Conference on Machine Learning*, 233–40. ICML '07. New York, NY, USA: ACM. doi:10.1145/1273496.1273526.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics*, Topic Models and the Cultural Sciences, 41 (6): 570–606. doi:10.1016/j.poetic.2013.08.004.

Duda, Richard O., and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis*. 1 edition. New York: Wiley.

Edwards, Paul N, Steven J Jackson, Geoffrey C Bowker, and Cory P Knobel. 2007. *Understanding Infrastructure: Dynamics, Tensions, and Design*. Working Paper. http://deepblue.lib.umich.edu/handle/2027.42/49353.

Egyedi, Tineke. 2001. "Infrastructure Flexibility Created by Standardized Gateways: The Cases of XML and the ISO Container." *Knowledge, Technology & Policy* 14 (3): 41–54. doi:10.1007/s12130-001-1015-4.

Fitzpatrick, Kathleen. 2010. "Peer–to–peer Review and the Future of Scholarly Authority." *Social Epistemology* 24 (3): 161–79. doi:10.1080/02691728.2010.498929.

Fortnow, Lance. 2009. "Viewpoint: Time for Computer Science to Grow up." *Communications of the ACM* 52 (8): 33. doi:10.1145/1536616.1536631.

Galison, Peter Louis. 1997. *Image and Logic: A Material Culture of Microphysics*. 1st ed. University Of Chicago Press.

Garfield, Eugene. 1955. "Citation Indexes for Science: A New Dimension in Documentation Through Association of Ideas." *Science* 122 (3159): 108–11. doi:10.1126/science.122.3159.108.

———. 1979. *Citation Indexing Its Theory and Application in Science, Technology and Humanities*. 1St Edition edition. New York: John Wiley & Sons Inc.

Garfinkel, H., M. Lynch, and E. Livingston. 1981. "The Work of a Discovering Science Construed with Materials from the Optically Discovered Pulsar." *Philosophy of the Social Sciences* 11 (2): 131–58.

Garvey, William D. 1979. *Communication, the Essence of Science: Facilitating Information Exchange Among Librarians, Scientists, Engineers, and Students*. Oxford ; New York: Pergamon Pr.

Garvey, William D., and Belver C. Griffith. 1964. "Scientific Information Exchange in Psychology The Immediate Dissemination of Research Findings Is Described for One." *Science* 146 (3652): 1655–59. doi:10.1126/science.146.3652.1655.

———. 1968. "Informal Channels of Communication in the Behavioral Sciences: Their Relevance in the Structuring of Formal or Bibliographic Communication." In *The Foundations of Access to Knowledge*, 129–51. [Syracuse, N.Y.]Division of Summer Sessions, Syracuse University.

Geiger, R. Stuart, and David Ribes. 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*, 117–26. CSCW '10. New York, NY, USA: ACM. doi:10.1145/1718918.1718941.

———. 2011. "Trace Ethnography: Following Coordination Through Documentary Practices." In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, 1–10.

Gerrish, Sean, and David M. Blei. 2010. "A Language Based Approach to Measuring Scholarly Impact." In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 375–82. http://machinelearning.wustl.edu/mlpapers/paper_files/icml2010_GerrishB10.pdf.

Gillies, James, and Robert Cailliau. 2000. *How the Web Was Born: The Story of the World Wide Web*. Oxford: Oxford University Press.

Ginsparg, P. 2004. "Scholarly Information Architecture, 1989 - 2015." *Data Science Journal* 3: 29–37. doi:10.2481/dsj.3.29.

Ginsparg, Paul. 1994. "First Steps Towards Electronic Research Communication." *Computers in Physics* 8 (4): 390–96. doi:10.1063/1.4823313.

———. 2007. "Next-Generation Implications of Open Access." *CTWatch Quarterly* 3 (3). http://www.ctwatch.org/quarterly/articles/2007/08/next-generation-implications-of-open-access/.

Gitelman, Lisa, ed. 2013. *"Raw Data" Is an Oxymoron*. Cambridge, Massachusetts ; London, England: The MIT Press.

Goddard, Stephen B. 1996. *Getting There: The Epic Struggle Between Road and Rail in the American Century*. University of Chicago Press.

Gold, Matthew K. 2012. *Debates in the Digital Humanities*. U of Minnesota Press.

Goldstone, Andrew, and Ted Underwood. 2013. “What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship?” *Journal of Digital Humanities*. http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/.

———. 2014. “The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us.” *New Literary History*, Forthcoming.

Greco, A.N., and R.M. Wharton. “The Market Demand for University Press Books.” *Journal of Scholarly Publishing* 42, no. 1 (2010): 1–15.

Griffiths, T. L., and M. Steyvers. 2004. “Finding Scientific Topics.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (Suppl 1): 5228–35. http://www.pnas.org/content/101/suppl.1/5228.short.

Grimmer, Justin. 2010. “A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases.” *Political Analysis* 18 (1): 1–35. doi:10.1093/pan/mpp034.

Grusin, Richard. 2014. “The Dark Side of Digital Humanities: Dispatches from Two Recent Mla Conventions.” *Differences* 25 (1): 79–92. doi:10.1215/10407391-2420009.

Halavais, A. 2006. “Scholarly Blogging: Moving Toward the Visible College.” In *Uses of Blogs*, 117–26. Peter Lang International Academic Publishers.

Hall, David, Daniel Jurafsky, and Christopher D. Manning. 2008. “Studying the History of Ideas Using Topic Models.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 363–71. EMNLP ’08. Stroudsburg, PA, USA: Association for Computational Linguistics.

Hank, Carolyn. 2011. “Scholars and their Blogs: Characteristics, Preferences, and Perceptions Impacting Digital Preservation.” PhD thesis, University of North Carolina. http://ils.unc.edu/~wildem/ASIST2011/Hank-diss.pdf.

———. 2013. “Communications in Blogademia: An Assessment of Scholar Blogs’ Attributes and Functions.” *New Review of Information Networking* 18 (2): 51–69. doi:10.1080/13614576.2013.802179.

Heuser, Ryan, and Long Le-Khac. 2012. “A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method.” *Pamphlets of the Stanford Literary Lab*, no. 4.

Hine, Dr Christine M. 2000. *Virtual Ethnography*. 1st ed. Sage Publications Ltd.

Hirsch, J. E. 2005. “An Index to Quantify an Individual’s Scientific Research Output.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (46): 16569–72. doi:10.1073/pnas.0507655102.

Hockey, S. 2004. “The History of Humanities Computing.” *A Companion to Digital Humanities*, 1–19.

Hofmann, Thomas. 1999. "Probabilistic Latent Semantic Indexing." In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57. SIGIR '99. New York, NY, USA: ACM. doi:10.1145/312624.312649.

Howard, Philip N. 2002. "Network Ethnography and the Hypermedia Organization: New Media, New Organizations, New Methods." *New Media & Society* 4 (4): 550–74. doi:10.1177/146144402321466813.

Hughes, Thomas Parker. 1983. *Networks of Power: Electrification in Western Society, 1880-1930*. Johns Hopkins University Press. http://books.google.com/books?hl=en&lr=&id=g07Q9M4agp4C&oi=fnd&pg=PR9&dq=networks+of+power&ots=BBBIoAopMZ&sig=WS2H1l5qXc1_oYJTUaK2qnBVyMA.

Hurd, Julie M. 2000. "The Transformation of Scientific Communication: A Model for 2020." *Journal of the American Society for Information Science* 51 (14): 1279–83. doi:10.1002/1097-4571(2000)9999:9999<::AID-ASI1044>3.0.CO;2-1.

Jensen, Michael. 2007. "Authority 3.0: Friend or Foe to Scholars?" *Journal of Scholarly Publishing* 39 (1): 297–307. doi:10.1353/scp.2007.0027.

Jockers, Matthew L. 2013. *Macroanalysis : Digital Methods and Literary History*. 1st Edition. University of Illinois Press.

Johnson, R. Burke, Anthony J. Onwuegbuzie, and Lisa A. Turner. 2007. "Toward a Definition of Mixed Methods Research." *Journal of Mixed Methods Research* 1 (2): 112–33. doi:10.1177/1558689806298224.

Jones, Steven E. 2013. *The Emergence of the Digital Humanities*. 1 edition. New York: Routledge.

Juola, Patrick. 2008. "Killer Applications in Digital Humanities." *Literary and Linguistic Computing* 23 (1): 73–83. doi:10.1093/llc/fqm042.

Kirschenbaum, Matthew. 2010. "What Is Digital Humanities and What's It Doing in English Departments?" *ADE Bulletin* 150: 55–61. http://wip.cch.kcl.ac.uk/wp-content/uploads/2012/01/kirschenbaum_ade150.pdf.

Kjellberg, Sara. 2010. "I Am a Blogging Researcher: Motivations for Blogging in a Scholarly Context." *First Monday* 15 (8). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2962.

Klein, Julie Thompson. 2014. *Interdisciplining Digital Humanities: Boundary Work in an Emerging Field*. University of Michigan Press. http://hdl.handle.net/2027/spo.12869322.0001.001.

Kling, Rob, and Ewa Callahan. 2003. “Electronic Journals, the Internet, and Scholarly Communication.” *Annual Review of Information Science and Technology* 37 (1): 127–77. doi:10.1002/aris.1440370105.

Kling, Rob, and Lisa Covi. 1995. “Electronic Journals and Legitimate Media in the Systems of Scholarly Communication.” *The Information Society* 11 (4): 261–71. http://www.tandfonline.com/doi/abs/10.1080/01972243.1995.9960200.

Kling, Rob, Geoffrey McKim, and Adam King. 2003. “A Bit More to It: Scholarly Communication Forums as Socio-Technical Interaction Networks.” *Journal of the American Society for Information Science and Technology* 54 (1): 47–67. doi:10.1002/asi.10154.

Korth, Henry F., Philip A. Bernstein, Mary Fernandez, Le Gruenwald, Phokion G. Kolaitis, Kathryn McKinley, and Tamer Ozsu. 2008. “Paper and Proposal Reviews: Is the Process Flawed?” *SIGMOD Rec.* 37 (3): 36–39. doi:10.1145/1462571.1462581.

Krippendorff, Klaus H. 2012. *Content Analysis: An Introduction to Its Methodology*. Third Edition edition. Los Angeles ; London: SAGE Publications, Inc.

Kronick, David A. 2001. “The Commerce of Letters: Networks and ‘Invisible Colleges’ in Seventeenth- and Eighteenth-Century Europe.” *The Library Quarterly* 71 (1): 28–43. doi:10.2307/4309484.

Lagoze, Carl, Paul Edwards, Christian Sandvig, and Jean-Christophe Plantin. “Should I Stay or Should I Go? Alternative Infrastructures in Scholarly Publishing.” *International Journal of Communication* 9, no. 0 (March 30, 2015): 20.

Lampland, Martha, and Susan Leigh Star. 2009. *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*. Cornell University Press.

Latour, Bruno. 1988. *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press.

———. 1991. “Technology Is Society Made Durable.” In *A Sociology of Monsters Essays on Power, Technology and Domination*, 103–32. Sociological Review Monograph 38. http://www.bruno-latour.fr/node/263.

———. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, USA.

Latour, Bruno, and Steve Woolgar. 1986. *Laboratory Life*. Princeton University Press.

Laudun, John, and Jonathan Goodwin. 2013. “Computing Folklore Studies: Mapping over a Century of Scholarly Production Through Topics.” *Journal of American Folklore* 126 (502): 455–75.

http://muse.jhu.edu.proxy.lib.umich.edu/journals/journal_of_american_folklore/v126/126.502.laudun.html.

Levinson, Marc. 2010. *The Box: How the Shipping Container Made the World Smaller and the World Economy Bigger (New in Paper)*. Princeton University Press. http://books.google.com/books?hl=en&lr=&id=ljHq6-HnW1QC&oi=fnd&pg=PR12&dq=the+box:+shipping+container&ots=P_W1YhKoT_&sig=FVluKKW4ePSQ6eQIIJOETq29J48.

Levy, D.M. 2007. "No Time to Think: Reflections on Information Technology and Contemplative Scholarship." *Ethics and Information Technology* 9 (4): 237–49.

Lievrouw, Leah A. 1988. "Four Programs of Research in Scientific Communication." *Knowledge in Society* 1 (2): 6–22. http://link.springer.com/article/10.1007/BF02687210.

———. 1989. "The Invisible College Reconsidered Bibliometrics and the Development of Scientific Communication Theory." *Communication Research* 16 (5): 615–28. doi:10.1177/009365089016005004.

Lin, Jimmy, Kari Kraus, and Ricardo Punzalan. 2014. "Supporting 'Distant Reading' for Web Archives." In. Lausanne Switzerland. http://dharchive.org/paper/DH2014/Paper-886.xml.

Linmans, A. J. M. 2009. "Why with Bibliometrics the Humanities Does Not Need to Be the Weakest Link." *Scientometrics* 83 (2): 337–54. doi:10.1007/s11192-009-0088-9.

Liu, Alan. 2013. "The Meaning of the Digital Humanities." *PMLA* 128 (2): 409–23. doi:10.1632/pmla.2013.128.2.409.

Luzón, María José. 2009. "Scholarly Hyperwriting: The Function of Links in Academic Weblogs." *Journal of the American Society for Information Science and Technology* 60 (1): 75–89. http://onlinelibrary.wiley.com/doi/10.1002/asi.20937/full.

Lynch, Clifford. 2010. "Imagining a University Press System to Support Scholarship in the Digital Age." *Journal of Electronic Publishing* 13 (2). doi:10.3998/3336451.0013.207.

Lynch, Michael. 1997. *Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science*. Cambridge University Press.

Mark, Gloria, Mossaab Bagdouri, Leysia Palen, James Martin, Ban Al-Ani, and Kenneth Anderson. 2012. "Blogs As a Collective War Diary." In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 37–46. CSCW '12. New York, NY, USA: ACM. doi:10.1145/2145204.2145215.

Mccallum, AK. 2002. "MALLET: A Machine Learning for Language Toolkit."

McCarty, Willard. 2012. “A Telescope for the Mind?".” In *Debates in the Digital Humanities*, 113–23. http://books.google.com/books?hl=en&lr=&id=_6mo2tApzQQC&oi=fnd&pg=PA113&dq=A%22telescope%22for%22the%22mind&ots=9XujosK3Nx&sig=9wpKhdfNf3N24X9Woj2bDvF5cTo.

McFarland, Daniel A., Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D. Manning, and Daniel Jurafsky. 2013. “Differentiating Language Usage Through Topic Models.” *Poetics*, Topic Models and the Cultural Sciences, 41 (6): 607–25. doi:10.1016/j.poetic.2013.06.004.

McPherson, Tara. 2009. “Introduction: Media Studies and the Digital Humanities.” *Cinema Journal* 48 (2): 119–23.

———. 2012. “Why Are the Digital Humanities So White? Or Thinking the Histories of Race and Computation.” *Debates in the Digital Humanities*, 139–60.

Meadows, A. J. 1997. *Communicating Research*. Emerald Group Publishing Limited.

Mei, Qiaozhu, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. “Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs.” In Proceedings of the 16th International Conference on World Wide Web, 171–80. WWW ’07. New York, NY, USA: ACM, 2007. doi:10.1145/1242572.1242596.

Mei, Qiaozhu, Chao Liu, Hang Su, and ChengXiang Zhai. “A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs.” In Proceedings of the 15th International Conference on World Wide Web, 533–42. WWW ’06. New York, NY, USA: ACM, 2006. doi:10.1145/1135777.1135857.

Menzel, Herbert. 1968. “Informal Communication in Science: Its Advantages and Its Formal Analogues.” In *The Foundations of Access to Knowledge*, 153–63. [Syracuse, N.Y.]Division of Summer Sessions, Syracuse University.

Mimno, David. 2012. “Computational Historiography: Data Mining in a Century of Classics Journals.” *Journal on Computing and Cultural Heritage* 5 (1): 1–19. doi:10.1145/2160165.2160168.

Mohr, John W., and Petko Bogdanov. 2013. “Introduction—Topic Models: What They Are and Why They Matter.” *Poetics*, Topic Models and the Cultural Sciences, 41 (6): 545–69. doi:10.1016/j.poetic.2013.10.001.

Mohr, John W., Robin Wagner-Pacifici, Ronald L. Breiger, and Petko Bogdanov. 2013. “Graphing the Grammar of Motives in National Security Strategies: Cultural Interpretation, Automated Text Analysis and the Drama of Global Politics.” *Poetics*, Topic Models and the Cultural Sciences, 41 (6): 670–700. doi:10.1016/j.poetic.2013.08.003.

Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models For A Literary History*. Verso.

———. 2013. "'Operationalizing'." *New Left Review*, II, no. 84 (December): 103–119.

Nallapati, R., and W. Cohen. 2008. "Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs." In *International Conference for Weblogs and Social Media*. https://www.aaai.org/Papers/ICWSM/2008/ICWSM08-018.pdf.

Nowviskie, Bethany. 2013. "Skunks in the Library: A Path to Production for Scholarly R&D." *Journal of Library Administration* 53 (1): 53–66. doi:10.1080/01930826.2013.756698.

Pagano, Dennis, and Walid Maalej. 2011. "How Do Developers Blog?: An Exploratory Study." In *Proceedings of the 8th Working Conference on Mining Software Repositories*, 123–32. MSR '11. New York, NY, USA: ACM. doi:10.1145/1985441.1985461.

Pearl, Lisa, and Mark Steyvers. 2012. "Detecting Authorship Deception: A Supervised Machine Learning Approach Using Author Writeprints." *Literary and Linguistic Computing* 27 (2): 183–96. doi:10.1093/llc/fqs003.

Price, Derek J. de solla. 1985. "Little Science, Big Science. and Beyond." *New York*.

Price, Derek J. de Solla. 1963. *Little Science Big Science*. 1st edition. Columbia University Press.

PuschmannI, Cornelius, and Merja MahrtII. 2012. "Scholarly Blogging: A New Form of Publishing or Science Journalism 2.0?" In *Science and the Internet*. Düsseldorf: University Press. http://files.ynada.com/papers/cosci12.pdf.

Ramsay, Stephen. 2011. *Reading Machines: Toward and Algorithmic Criticism*. 1st Edition. University of Illinois Press.

Reagle, Joseph Michael. 2015. *Reading the Comments: Likers, Haters, and Manipulators at the Bottom of the Web*. Cambridge, Massachusetts: MIT Press.

Rettberg, Jill Walker. 2013. *Blogging*. Polity.

Rhody, Lisa M. 2013. "Topic Model Data for Topic Modeling and Figurative Language." *Journal of Digital Humanities*. http://journalofdigitalhumanities.org/2-1/topic-model-data-for-topic-modeling-and-figurative-language-by-lisa-m-rhody/.

Ribes, David. 2014. "Ethnography of Scaling, or, How to a Fit a National Research Infrastructure in the Room." In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, 158–70. CSCW '14. New York, NY, USA: ACM. doi:10.1145/2531602.2531624.

Ribes, David, and Geoffrey C. Bowker. 2009. "Between Meaning and Machine: Learning to Represent the Knowledge of Communities." *Inf. Organ.* 19 (4): 199–217. doi:10.1016/j.infoandorg.2009.04.001.

Ribes, David, Steven Jackson, Stuart Geiger, Matthew Burton, and Thomas Finholt. 2013. "Artifacts That Organize: Delegation in the Distributed Organization." *Information and Organization* 23 (1): 1–14. doi:10.1016/j.infoandorg.2012.08.001.

Riddell, Allen Beye. 2014. "How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models." In *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*. Rochester, New York: Camden House. https://ariddell.org/static/wustl-german-journals-unrevised-proof.pdf.

Rodzvilla, John. 2002. *We've Got Blog: How Weblogs Are Changing Our Culture*. Basic Books.

Rosenberg, Scott. 2010. *Say Everything: How Blogging Began, What It's Becoming, and Why It Matters*. Three Rivers Press.

Sandvig, Christian. "The Internet as Infrastructure." In *The Oxford Handbook of Internet Studies*, 86–106. Oxford University Press, 2013. http://books.google.com/books?hl=en&lr=&id=zXryMw4uVPoC&oi=fnd&pg=PA86&dq=the+internet+as+infrastructure&ots=S6AoLyZlco&sig=SHc-uQmIoNKIBRHZtPvF_YDqzmE.

Saper, C. 2006. "Blogademia." *Reconstruction* 6 (4). http://www.citeulike.org/group/1736/article/1108357.

Schmidt, Benjamin M. 2013. "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities* 2 (1). http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/.

Serfaty, Viviane. 2004. *The Mirror and the Veil: An Overview of American Online Diaries and Blogs*. Rodopi.

Seroussi, Yanir, Ingrid Zukerman, and Fabian Bohnert. 2011. "Authorship Attribution with Latent Dirichlet Allocation." In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 181–89. CoNLL '11. Stroudsburg, PA, USA: Association for Computational Linguistics.

Siles, Ignacio. 2011. "From Online Filter to Web Format: Articulating Materiality and Meaning in the Early History of Blogs." *Social Studies of Science* 41 (5): 737–58.

———. 2012. "Web Technologies of the Self: The Arising of the 'Blogger' Identity." *Journal of Computer-Mediated Communication* 17 (4): 408–21.

Singh, V.K., P. Waila, R. Sadat, R. Piryani, and A Uddin. 2013. “Computational Analysis of Thematic Blog Data for Sociological Inference Mining.” In *2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, 293–98. doi:10.1109/SACI.2013.6608985.

Star, Susan Leigh. 1999a. “It’s Infrastructure All the Way down.” *American Behavioural Scientist* 43: 377–91. ftp://ftp10.us.freebsd.org/users/azhang/disc/disc01/cd1/out/papers/dl/p271-star.pdf.

———. 1999b. “The Ethnography of Infrastructure.” *American Behavioral Scientist* 43 (3): 377–91. doi:10.1177/00027649921955326.

Star, Susan Leigh, and Karen Ruhleder. 1996. “Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces.” *INFORMATION SYSTEMS RESEARCH* 7: 111–34. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.88.3121.

Star, Susan Leigh, and Anselm Strauss. 1999. “Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work.” *Computer Supported Cooperative Work (CSCW)* 8 (1-2): 9–30. doi:10.1023/A:1008651105359.

Suber, Peter. 2012. *Open Access*. MIT Press Essential Knowledge Series. Cambridge, Mass: MIT Press. http://mitpress.mit.edu/sites/default/files/titles/content/openaccess/Suber_05_toc.html.

Sula, Chris Alen, and Matthew Miller. 2014. “Citations, Contexts, and Humanistic Discourse: Toward Automatic Extraction and Classification.” *Literary and Linguistic Computing* 29 (3): 452–64. doi:10.1093/llc/fqu019.

Svensson, Patrik. 2010. “The Landscape of Digital Humanities.” *Digital Humanities Quarterly* 4 (1). http://digitalhumanities.org:8080/dhq/vol/4/1/000080/000080.html.

———. 2009. “Humanities Computing as Digital Humanities” 3 (3). http://www.digitalhumanities.org/dhq/vol/3/3/000065/000065.html.

Tang, Jian, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. “Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis.” In, 190–98. http://jmlr.org/proceedings/papers/v32/tang14.html.

Teddlie, Charles. 2009. *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. SAGE.

Templeton, Clay, Travis Brown, Sayan Battacharyya, and Jordan Boyd-Graber. 2011. “Mining the Dispatch Under Supervision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Cor.” In *Chicago Colloquium on*

Digital Humanities and Computer Science. http://www.umiacs.umd.edu/~jbg/docs/slda_civil_war.pdf.

Terras, Melissa, Julianne Nyhan, and Edward Vanhoutte, eds. 2013. *Defining Digital Humanities: A Reader*. New edition edition. Farnham, Surrey, England : Burlington, VT: Ashgate Pub Co.

Thelwall, Mike, and Paul Wouters. 2005. "What's the Deal with the Web/Blogs/the Next Big Technology: A Key Role for Information Science in E-Social Science Research?" In *Context: Nature, Impact, and Role*, edited by Fabio Crestani and Ian Ruthven, 187–99. Lecture Notes in Computer Science 3507. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/11495222_15.

Underwood, Ted. 2014. "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago." *Representations* 127 (1): 64–72. doi:10.1525/rep.2014.127.1.64.

Vandegrift, Micah. 2012. "What Is Digital Humanities and What's It Doing in the Library?" *In the Library with the Lead Pipe*, June. http://www.inthelibrarywiththeleadpipe.org/2012/dhandthelib/.

Vardi, Moshe Y. 2009. "Conferences Vs. Journals in Computing Research." *Communications of the ACM* 52 (5): 5. doi:10.1145/1506409.1506410.

———. 2010. "Revisiting the Publication Culture in Computing Research." *Communications of the ACM* 53 (3): 5–5. doi:10.1145/1666420.1666421.

Velden, Theresa, and Carl Lagoze. 2008. "The Transformation of Scientific Communication Systems in the Digital Age – Towards a Methodology for Comparing Scientific Communication Cultures."

———. 2013. "The Extraction of Community Structures from Publication Networks to Support Ethnographic Observations of Field Differences in Scientific Communication." *Journal of the American Society for Information Science and Technology* 64 (12): 2405–27. doi:10.1002/asi.22929.

Velden, Theresa, Asif-ul Haque, and Carl Lagoze. 2010. "A New Approach to Analyzing Patterns of Collaboration in Co-Authorship Networks: Mesoscopic Analysis and Interpretation." *Scientometrics* 85 (1): 219–42. doi:10.1007/s11192-010-0224-6.

Waila, P., V.K. Singh, and M.K. Singh. 2013. "Blog Text Analysis Using Topic Modeling, Named Entity Recognition and Sentiment Classifier Combine." In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1166–71. doi:10.1109/ICACCI.2013.6637342.

Waters, Lindsay. 2004. *Enemies of Promise: Publishing, Perishing and the Eclipse of Scholarship*. Prickly Paradigm Press, LLC.

Welshons, M. 2009. “Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences.”

Wickham, Hadley. 2014. “Tidy Data.” *Under Review*. http://courses.had.co.nz.s3-website-us-east-1.amazonaws.com/12-rice-bdsi/slides/07-tidy-data.pdf.

Wigner, Eugene P. 1960. “The Unreasonable Effectiveness of Mathematics in the Natural Sciences. Richard Courant Lecture in Mathematical Sciences Delivered at New York University, May 11, 1959.” *Communications on Pure and Applied Mathematics* 13 (1): 1–14. doi:10.1002/cpa.3160130102.

Wilkens, Matthew. “Canons, Close Reading, and the Evolution of Method.” *Debates in the Digital Humanities*, 2012, 249–58.

Wilkinson, David, Gareth Harries, Mike Thelwall, and Liz Price. 2003. “Motivations for Academic Web Site Interlinking: Evidence for the Web as a Novel Source of Information on Informal Scholarly Communication.” *Journal of Information Science* 29 (1): 49–56. doi:10.1177/016555150302900105.

Williams, Robin, and Neil Pollock. 2011. “Research Commentary—Moving Beyond the Single Site Implementation Study: How (and Why) We Should Study the Biography of Packaged Enterprise Solutions.” *Information Systems Research* 23 (1): 1–22. doi:10.1287/isre.1110.0352.

Wouters, Paul. 1999. “The Creation of the SCI.” In *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, edited by Mary Ellen Bowden, Trudi Bellardo Hahn, and Robert Virgil Williams, 127–36. ASIS Monograph Series. Medford, NJ: Published for the American Society for Information Science; the Chemical Heritage Foundation by Information Today.

Wouters, Paul, and Repke de Vries. 2004. “Formally Citing the Web.” *Journal of the American Society for Information Science and Technology* 55 (14): 1250–60. doi:10.1002/asi.20080.

Wu, Tim. 2003. *Network Neutrality, Broadband Discrimination*. SSRN Scholarly Paper ID 388863. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=388863.

Yang, Tze-I, Andrew J. Torget, and Rada Mihalcea. 2011. “Topic Modeling on Historical Newspapers.” In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 96–104. LaTeCH ’11. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org.proxy.lib.umich.edu/citation.cfm?id=2107636.2107649.

Yano, Tae, William W. Cohen, and Noah A. Smith. 2009. “Predicting Response to Political Blog Posts with Topic Models.” In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of* *the Association for Computational Linguistics*, 477–85. NAACL ’09. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org.proxy.lib.umich.edu/citation.cfm?id=1620754.1620824.

Zimmerman, Ann S. 2008. “New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data.” *Science Technology Human Values* 33 (5): 631–52. doi:10.1177/0162243907306704