DHQ: Digital Humanities Quarterly
2020
Volume 14 Number 1

The Push and Pull of Digital Humanities: Topic Modeling the "What is digital humanities?" Genre

Elizabeth Callaway <elizabeth_dot_callaway_at_utah_dot_edu>, University of Utah
Jeffrey Turner <jeff_dot_turner_at_utah_dot_edu>, University of Utah
Heather Stone <heather_at_askthestones_dot_com>, TETON Sports
Adam Halstrom <adam_dot_halstrom_at_utah_dot_edu>, University of Utah

Abstract

In this paper we run a topic model on over 300 article-length pieces from the extended bibliography of Melissa Terras, Juliette Nyhan, and Edward Vanhoutte's edited collection *Defining Digital Humanities*. We use this topic model as a way to think through entry into the digital humanities as a negotiation between warm invitation and gatekeeping, the "pull" and "push" of digital humanities. We then analyze the metadata we collected about these pieces to explore how the push and pull manifest themselves unevenly across different demographics.

Introduction

Digital humanities has a definition addiction. In the decades since the formal inauguration of the term, rather than settle on a shared set of contours that encompass the field, definitions about what digital humanities is (and is not) have only proliferated. By the time Matthew Kirschenbaum wrote his definition in 2010, he had noted that pieces like his own were already an established genre [Kirschenbaum 2010b]. One notable collection of definitions is a recent volume edited by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte, *Defining Digital Humanities: A Reader*, which contains about twenty foundational definitions of the field [Terras, Nyhan, and Vanhoutte n.d. 2013]. In a useful and extensive addition to their book Terras, Nyhan, and Vanhoutte maintain an online bibliography of "Further Reading", which is an ongoing list of definitions for digital humanities [Terras, Nyhan, and Vanhoutte n.d. 2013]. We used this bibliography as a starting point to build a corpus of 334 definitions which we topic modeled and visualized in order to get a meta-view of the discipline and its members. 1

We were particularly interested in taking this approach in order to analyze the welcome-mat/trapdoor dynamic that Ted Underwood introduced at his 2018 MLA presentation. Underwood describes entry into his sub-field of cultural analytics as encountering "a gentle welcome mat followed by a trapdoor" [Underwood 2018]. He argues that there is insufficient academic infrastructure to help new scholars become competent in cultural analytics. This lack of support has resulted in a field that is ostensibly open but is actually only available to those who have prior extracurricular experience in coding and statistics (which, he notes, is unevenly distributed). Underwood carefully kept his critique to cultural analytics, which he does not consider synonymous with digital humanities. However, even in areas of digital humanities that do not involve the "massive cultural datasets" [Manovich 2016] of cultural analytics, his argument seems to apply. 2

Stephen Ramsay's well-known MLA talk from 2011 titled "Who's in and Who's out," exemplifies Underwood's welcome mat and trapdoor analogy. Ramsay's talk, which is often read as an example of coding-as-gatekeeping in digital humanities, is more of an example of a simultaneous trapdoor *and* welcome mat. In the most infamous section, Ramsay states, "Do you have to know how to code? I'm a tenured professor of digital humanities, and I say 'yes'" [Ramsay 2011]. Even though he follows this statement by acknowledging that different advisors might have different answers to this question, this excerpt functions and feels like a trapdoor because it refuses admittance to those who do not program. Elsewhere in the same short piece, Ramsay takes an almost opposite stance, admitting that "the coding 3

question is, for me, a canard." What he thinks sets digital humanities apart is not coding, but building. He is "willing to entertain highly expansive definitions of what it means to build something," which is a statement that functions more like a welcome mat by inviting people to try building of nearly any sort to enter the ranks of "who's in." Even in the most oft-quoted example of gatekeeping, then, one experiences not just a line drawn around what defines digital humanities for one person, but whiplash between a warm invitation to attempt new ways of building and cold exclusion for not knowing the correct tools to be a digital humanist.

4 This simultaneous push and pull, this invitation-in and guarding-of digital humanities exists in both field-wide definitions and individual methods. One technique in which this dynamic is evident is topic modeling. In the past few years, topic modeling has gone from being a "hot," new method to existing solidly within the mainstream digital humanities toolkit. Topic modeling is so much associated with digital humanities that, at times, it seems like digital humanities is topic modeling. Scott Weingart and Elijah Meeks use it as a "synecdoche of digital humanities" in their special-issue introduction to topic modeling [Weingart and Meeks 2012]. Christopher Schöch has stated that "Topic Modeling has proven immensely popular in Digital Humanities" [Schoch 2017]. Stephen Robertson even notes that "text mining and topic modeling are the predominant practices" within digital literary studies [Robertson 2014]. A myriad of informal introductions walks the novice through what topic modeling is and how one can do it. Explaining what topic modeling is has even become something of another sub-genre. There are many excellent and accessible summaries by prominent digital humanists detailing what the process entails and assumes [Weingart 2012] [Posner 2012b] [Jockers 2011] [Underwood 2012]. These definitions and guides invite newcomers to use the method.

5 Despite the abundant and inviting walk-throughs, amateurs are warned away from this method as vigorously as they are invited to try it. Ben Schmidt warns that "simplifying topic models for humanists who will not (and should not) study the underlying algorithms creates an enormous potential for groundless — or even misleading — 'insights'" [Schmidt 2013]. Andrew Goldstone warns of easy-entry digital humanities tools in general when he writes: "DH should be wary of promises of ease: in prepackaged tools, in well-meaning introductory tutorials and workshops that necessarily stop short of what a researcher would need to draw conclusions, in rationalizations of inconclusive arguments as exploration, play, or productive failure" [Goldstone 2017].

6 This simultaneous invitation and warning away is a confounding feature of precisely those areas in digital humanities which established scholars have used to draw lines around the field. One has to learn a little command line, R, or Python to run and visualize the results of a topic model. With all the manuals and textbooks available, topic modeling could be a relatively painless way to earn one's entry into digital humanities and satisfy any remaining naysayers who demand or imply that digital humanists know how to code. However, approaching it also puts newcomers in the position of the magician's apprentice: is this going to be a tool that gets out of our control and wreaks havoc on our scholarship and our reputation in the very field we are trying to enter?

7 This article explores a series of topic models run by relative newcomers to digital humanities who encountered both a welcome mat and potentially a trap door. At the time of carrying out this project, we were a group of one postdoc and three graduate students in various humanities fields who learned topic modeling together through Matthew Jockers' introductory book [Jockers 2014] and a DHSI session on topic modeling. Starting from Melissa Terras' bibliography on "Defining Digital Humanities" and adding non-redundant entries from Elijah Meeks' conceptual map of DH [Meeks 2011], we collected and curated a corpus of full-text digital humanities definitions as .txt files as well as 15 different metadata fields such as department, career stage, institution, etc. for each definition's authors. This set of texts is not a comprehensive list of definitions within digital humanities. It does, however, include a breadth of the disciplines involved in digital humanities scholarship, it contains varying forms of media involved in defining digital humanities, and it is large enough for topic modeling to be useful, though it is on the small side of topic modeling corpora. One limitation of the corpus is that the texts we topic modeled were all in English, which might skew the overall corpus' representation toward North American and English-speaking European scholarship. To clean the texts we removed special (non-UTF-8) characters, front matter, and headers and footers. We then topic modeled this 334-definition corpus using the R package "mallet" by David Mimno [Mimno 2013]. We then prepared the topic modeling data for analysis in three ways. First, we produced word clouds representing the top 100 words associated with each topic with size representing the preponderance of each word. Second, we produced timeline graphs of the average presence of each topic through time

(sum of all the topic presences for each year in a topic-document matrix divided by the number of documents in each year). Third, we mapped the location of the institutions of the first author of each document at the time the document was published. We color-coded the resulting dots to make heatmaps of each topic's presence in space. Despite the attempt to look at the data in various ways, very few of the timelines and not one of the heatmaps showed discernible trends of topics through time or space. What ended up being more revealing was the metadata we collected about each author at the time of their piece's publication: academic rank (or job title), department, institution, location, gender, media of publication (blog, article, etc.) and whether or not the piece was co-authored.

On one level, this project is a straightforward analysis of the results of our topic models and our metadata. We are interested in what types of things people talk about when they define digital humanities. Are there trends through time in how people define digital humanities? Are there detectable national or regional differences in topic distribution? Is there gender or academic rank disparity in who is writing definitions of the field? Most centrally, we wanted to know if a topic model of definitions combined with metadata about the authors of said definitions could tell us about the push and pull of digital humanities and its methods, especially as they are experienced across gender lines: is the push and pull of digital humanities detectable in the topics we produce? Are male and female authors represented proportionally in topics, or are there topics that are dominated by one gender? 8

On another level, this is an experiment in learning a quantitative method in digital humanities. Using only an introductory book, can a group of interested humanists run topic models that produce novel insights? What is all the fuss about topic modeling? Is it powerful? Fun? Dangerous? Misleading? Revealing? Or even predictable, merely confirming what we already know? This article serves as an analysis of our own admittance into the field. Writing as junior scholars just entering our disciplines, we feel both the push and pull of digital humanities intensely. We hope we can add to the discussion of the phenomenon by analyzing it from the other side of the disciplinary door from scholars like Underwood. 9

How is digital humanities defined?

While we ran models with a wide range of parameters on this corpus, we finally decided on a run with 55 topics. Briefly, topics are a group of words that tend to co-occur in the corpus, and working with them can feel uncomfortably arbitrary. However, after playing with the inputs for a while, 55 topics seemed to produce topics that were granular enough to be meaningful, but not so narrow as to be overlapping. 10

Some of the topics that emerged from this model run were what one might expect. Words about different kinds of tools tended to co-occur in certain entries. We had a "tool, tools, web, topic, mapping" topic that included papers discussing the implications of Web 2.0 and web tools in the digital research environment [Cohen 2008], surveys of projects using digital tools [Paradise 2015], and articles that claim to "focus on conceptual issues rather than particular tools or projects" but that nonetheless mention tool/s 72 times, for example [Robertson 2016]. Some other topics focused on digital humanities as coming out of humanities computing ("computing, mccarty, question, experimental, software"), discipline-specific history-department approaches ("history, historian, historians, philosophy, narrative"), and library roles in digital humanities ("libraries, librarians, service, library, work"). Others seemed to be words that co-occur when definitions explore digital humanities as it uses, relates to, or critiques social media ("twitter, social, network, users, scholars") or digital humanities in the classroom ("students, tools, teaching, college, pedagogy"). These topics fit the kinds of definitions we expected to find. 11

While these topics help to generally outline and categorize the range of subjects people address when they define digital humanities, there were four topics that, taken together, seem to encapsulate the push and pull of entry into digital humanities. These topics engage with coding in the context of jobs, digital humanities as a community with shared values, distant reading, and diversity in digital humanities 12

Figure 1. Word clouds of the four topics we examine in detail.

.

Code

Given our interest in the gatekeeping around coding and technical skills more broadly, we were drawn to inspecting the topic that emerged with the top words: “job, field, scholars, students, degree, code, programming, software, technologists.” The documents that are most highly associated with this topic indeed all address code and coding, especially in relation to the question of jobs. So, what is the consensus? Does one have to code to be a digital humanist or at least to get a job? There is no answer as such within the top documents associated with this topic. Instead, these documents together form more of a discussion of coding itself. The most favorable endorsements of coding range from memoir-like recollections of personal experiences of learning to code [Ramsay 2016] to advise on how and where to learn Ruby [Cordell 2011]. Others argue that coding is potentially the least interesting aspect of what digital humanities practitioners do and should be downplayed on the job market [Davidson 2011] or not required at all in job postings [Gailey and Poerter 2011]. Others came at coding from the opposite angle, arguing that developers and technologists should be recognized not only for their coding but also for their intellectual contributions to projects [Ridge 2013]. This topic may look like it would be comprised of scholars arguing that digital humanities is about coding, or that one needs to code to get a job. However, there is a range of different views towards coding even among the documents which have a high probability of these words co-occurring. While coding may be an ongoing conversation in the field, there was no consensus that coding was a bar that had to be passed to enter the field. The articles associated with this topic usually brought up questions about coding and digital humanities rather than taking a hard stance about coding’s role in the field. In this case, what appeared at first to be an instance of gatekeeping was actually mostly a discussion about gatekeeping (from both directions) and how to overcome it.

13

Community

In opposition to gatekeeping around to coding, which we thought we might find in the documents associated with the coding topic but did not, another topic seemed to be “about” defining digital humanities as an inclusive community with shared values. If we expected coding to be the trap door in Underwood’s metaphor, the definition of a community with shared values might be the welcome mat that invites people in. A topic in our model was comprised of the co-occurring, positive words “values, community, open, collaboration, openness, ideas.” Documents with a high probability of these words co-occurring included Lisa Spiro’s call for creating a core values statement in a collaborative, open way where everyone can edit and contribute [Spiro 2012] as well as a follow up on Spiro’s initial call proposing no overarching final statement as an outcome but rather a perpetual place to gather ideas on values [Hawk 2013]. Other documents included Scheinfeldt’s definition of digital humanities as a “set of overlapping personal communities” with shared interests and shared values like open access and collaboration, among others [Scheinfeldt 2010]. This topic shows the largest spike in 2012, the year after Ramsay’s MLA 2011 “Who’s in Who’s out” talk (Figure 2), which, as we mentioned in the introduction to this article, galvanized a host of responses.

14

Overall, investigation of the documents most highly associated with this topic confirms our first impressions of the word cloud. When first faced with the “coding” cloud of words, however, it was evident that the words themselves failed to divulge the meaning which we eventually drew from them. The “coding” and “community” topics, taken together,

15

comprise two ends of a spectrum of topic transparency and underscore the importance of going back to the documents themselves to inform the interpretation of any topic. We had read about the necessity of toggling scales and perspectives in reading [Koller 2015] [Jänicke et al. 2015], but were surprised in how much topics varied in the distance between what they seemed to be "about" and what ideas the documents were actually addressing.

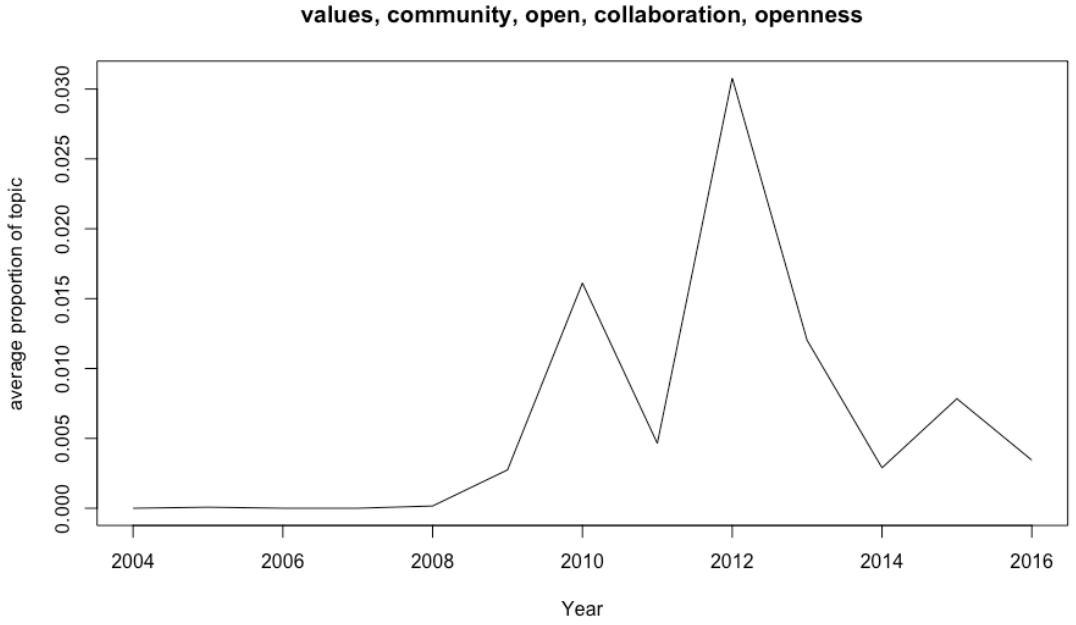

Figure 2. The average proportion of the "community and values" topic since 2004 (the first year in which there consistently is more than one document per year).

Distant Reading

Among the 334 definitions of digital humanities that we analyzed, there was a discernible "distant reading" topic. Words like: "literary, reading, text, texts, literature, Moretti, criticism, distant, patterns, close" comprised this topic. The texts most associated with this topic tend to mention Franco Moretti. They seem to either equate the digital humanities with text mining [Fish 2012] or let text mining stand as a synecdoche for the whole field [Turan 2016] [Benzon 2014b] [Benzon 2014a]. After all, literary analysis using text mining and distant reading is a major sub-field that sometimes seems to eclipse other forms of digital humanities. In response, Stephen Robertson has written about the disparate genealogy of digital history coming out of oral history, folk studies, and public histories in part to counter the predominant narrative of text analysis in humanities computing being *the* origin story of all of digital humanities [Robertson 2014]. 16

What is most striking about this topic, however, is not its robust presence in definitions of digital humanities, but the fact that the documents most highly associated with this topic are predominantly written by men (Figure 3). In fact, 19 of the top 20 documents associated with this topic are written by male first-authors.[1] There is no other topic in our run of this topic model that so disproportionately features male authors.[2] This might not come as a surprise since distant reading has been recognized as a field that has been unreceptive to women and which can replicate the most simple of stereotypes about women writers [Klein 2018]. For example, in *Macroanalysis*, Matthew Jockers separates 19th-century novels by gender and looks at differences in topics addressed by male and female authors. He writes that "The gender data from this corpus are a ringing confirmation of virtually all of our stereotypes about gender. Smack at the top of the list of themes most indicative of female authorship is 'Female fashion.' 'Fashion' is followed by 'Children,' 'Flowers,' 'Sewing,' and a series of themes associated with strong emotions" [Jockers 2013]. The decision to describe the gendered distributions of themes with the unfortunate word choice of "confirming" gender stereotypes rather than a more felicitous term like "reflecting" or "echoing" gender stereotypes of the era leaves Jockers open to critique. But an 17

issue with distant reading that is far more pervasive and important than a captious quibble with phrasing is the tendency to note a gendered (or other socially and historically constructed complex category) trend and move on to the next significant result rather than to sit with the uncomfortable result and think critically about it. A difference in the themes picked up by male and female writers of the 19th century can be the impetus for a discussion of what types of activities were open to women in the first place and the ways that laws, the marketplace, and social norms might restrict themes available to women writers. More often than not, in texts like *Macroanalysis*, the difference in topics is merely noted without critical reflection before the discussion moves on to listing other interesting trends that were also detected. In observing a pattern but skipping further critical examination, the trends uncovered by distant reading methods can, unfortunately, buttress stereotypes that women love fashion, for example. These instances of noting trends without unpacking them can affect how welcoming a field is to other voices. Women entering a sub-field heavily reliant on math, statistics, and rationality, such as text mining, might not feel especially welcome if the field is producing descriptions of women writers as emotional, fashion-obsessed mothers.

In addition to the reiteration of stereotypes about women writers, Franco Moretti, the founding figure of distant reading, is one of the men publicly named as a result of the #metoo movement. While some have excavated the female-led genealogy of distant reading, offering an alternative to Moretti [Buurma and Heffernan 2018], others, like Lauren Klein, have argued that this revelation about Moretti merely confirms the critiques that have been leveled against the practice of distant reading for years. Namely, that distant reading fails when it comes to dealing with gender conceptually, rhetorically, and in its models (as well as in the representation of women in the field) [Klein 2018]. 18

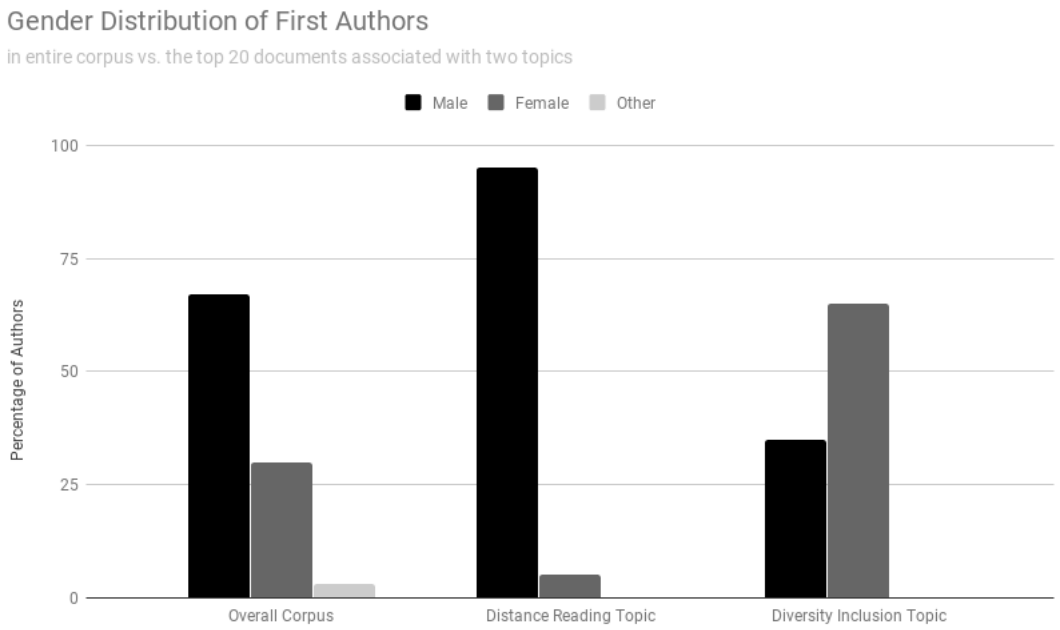

Figure 3. Gender Distribution of First Authors: The gender distribution of first authors in the entire corpus of DH definitions and the gender distribution of first authors in the top 20 documents associated with the distant reading topic and the diversity/inclusion topic (discussed next).

Diversity and Inclusion

The final topic we examine in-depth ties together a constellation of essays that turn the lens of critique back on digital humanities itself. In the documents most highly associated with this topic, authors address diversity and inclusion in digital humanities. The topic's top words are "race, project, projects, gender, women, feminist, transformdh, studies, color, black, koh, postcolonial, queer," and the documents with the highest probability of this topic engage vigorously with the question of diversity and representation in digital humanities. These definitions directly address the lack of diversity in digital humanities [Barnett 2014], propose taking a more proactive and transformative approach toward inviting diverse members in [Bailey 2012] [Lothian and Phillips 2013], advocate for recognizing the work that women and 19

people of color have always already been doing in digital humanities [Perez 2016], describe the feelings of being a woman of color in DH spaces [Cong-Huyen 2013] [Perez 2016], and survey projects that are at the cutting edge of examining race, privilege, and power and the digital humanities [Cong-Huyen 2013]. Women made up the majority of the authors of these pieces (Figure 3) with thirteen of the top twenty documents associated with this topic being authored by women.[3] It may not be surprising that when defining digital humanities, women are leading in bringing up issues of diversity and inclusion. This demonstrates that one of the ways inclusiveness benefits a field is to introduce new areas of research and writing that otherwise may not have been present. In this capacity, it shows that one of the benefits that women have already brought to the field of digital humanities is a critical conversation around diversity and inclusion.

The texts associated with this topic also pertain to the central theme of this article: the push and pull of the digital humanities, especially in regard to gender and race. This work is critically important in a field that embraces its reputation of niceness and inclusion while historically failing to foster the careers of diverse scholars. For example, the Digital Humanities 2011 conference theme of “Big Tent Digital Humanities” has been critiqued for simultaneously trumpeting the inclusiveness of the field while failing to recognize that the field might actually fail by many inclusivity metrics. Melissa Terras writes of DH 2011, “It is all very well saying that DH is open and welcoming and encourages participation – but despite open platforms such as DH answers, and the DIY approach, it is still a very rich, very western academic field with a limited number of job openings” [Terras 2011]. The conference also featured four male plenary and prize speakers complemented by male chairs for each of these most prestigious events. The 2011 conference may have provoked an ongoing response from the digital humanities community (in addition to that of Terras), as this diversity and inclusion topic increased over time, detectable in the corpus from 2011 onward (Figure 4). 20

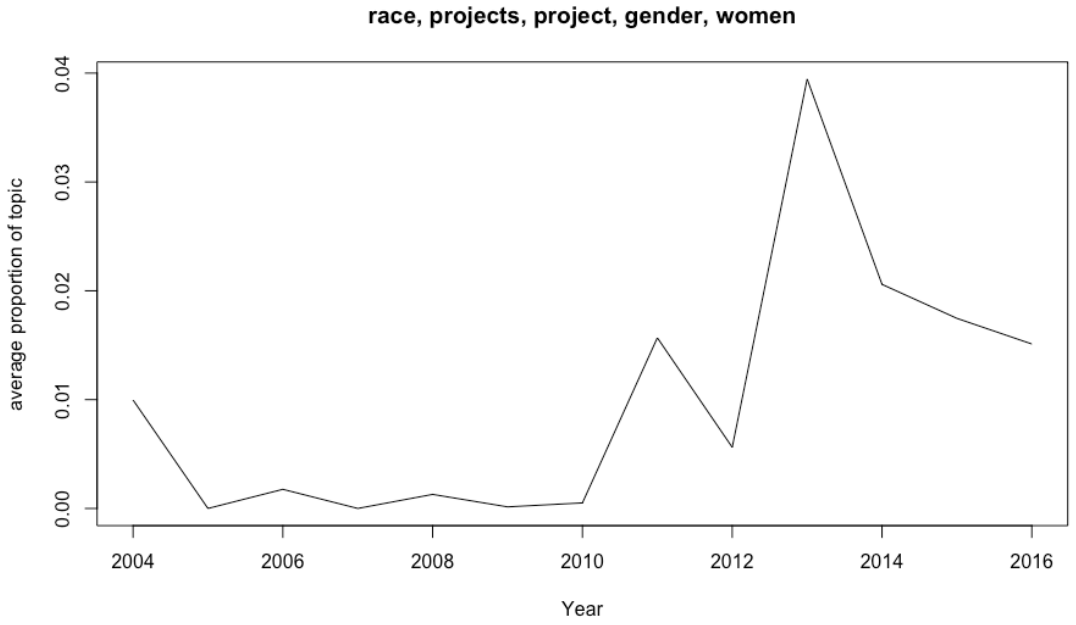

Figure 4. The average proportion of the “race, project/s, gender” topic since 2004 (the first year in which there consistently is more than one document per year).

At this point, it bears thinking through what we have been doing so far in this article. Are our objectives and method more aligned with the “distant reading” topic or the “diversity and inclusion” topic? Using topic modeling in this paper is most definitely a form of distant reading. Although we have a corpus in the hundreds, not thousands or millions, we are using a computer to put the corpus together and read it in aggregate in order to create new meaning. We also, of course, employ a similar method to the one we critiqued Matthew Jockers for in the previous section; we look at topics and examine the different ratios of male and female-authored papers associated with them. We even learned topic modeling in the first place from a different book by Jockers, the empowering *Text Analysis with R for Students of* 21

Literature which expands the field by teaching those who may not have picked up coding via “extracurricular” activities how to use R to perform all sorts of analyses on texts. The difference is, at this moment, we are attempting to *use* distant reading to *critique* distant reading. While we may question whether we can critique something using its own tools, we also think that the discomfort in being both a distant reader and the object of distant reading is a productive one. As Nickoal Eichmann, Jeana Jorgensen, and Scott Weingart write in their study on diversity and inclusion in the annual digital humanities conference, “by turning our ‘macroscopes’ on ourselves, we offer a critique of our culture, and hopefully inspire fruitful discomfort in DH practitioners who apply often-dehumanizing tools to their subjects, but have not themselves fallen under the same distant gaze” [Weingart 2016]. What we have found most productive in this exercise is exploring the discomfort of turning our gaze on our own work and writing into it. Instead of hiding the discomfort of dividing authors based on gender and looking for differences, or simply noting a gendered difference in topics and moving on, we discussed it, debated it, and ultimately wrote about it in this article (see footnote 2 for a summary of our debate).

Who is defining digital humanities?

As is evident from our explorations via topic modeling, we are not only interested in getting a general sense of *how* people define the digital humanities, but *who* is doing the defining. To that end, we collected data on the pieces and their authors. Did, for example, these definitions hold up to digital humanities’ reputation for being a collaborative endeavor? In what ways might those who are writing definitions be a homogeneous or diverse group? 22

It turns out that definitional pieces might be one of the types of digital humanities work that is most closely aligned with traditional academic writing, at least with respect to collaboration. Only 12.6 percent of the definitions in our corpus were co-authored. Department of author is another area that is not evenly distributed. While there are a wide variety of departments represented in the corpus, most definitions are written by authors with positions in English departments. English department definitions are nearly twice as numerous as the next most common department-of-origin: digital humanities departments/centers. Out of the people writing definitions from digital humanities centers, half of those have PhDs in English. Library and Information Sciences authors follow, then definitions from History departments. Next-most-numerous are definitions written by people who have no departmental affiliation: journalists, independent scholars, deans, software engineers, etc. (Figure 5). 23

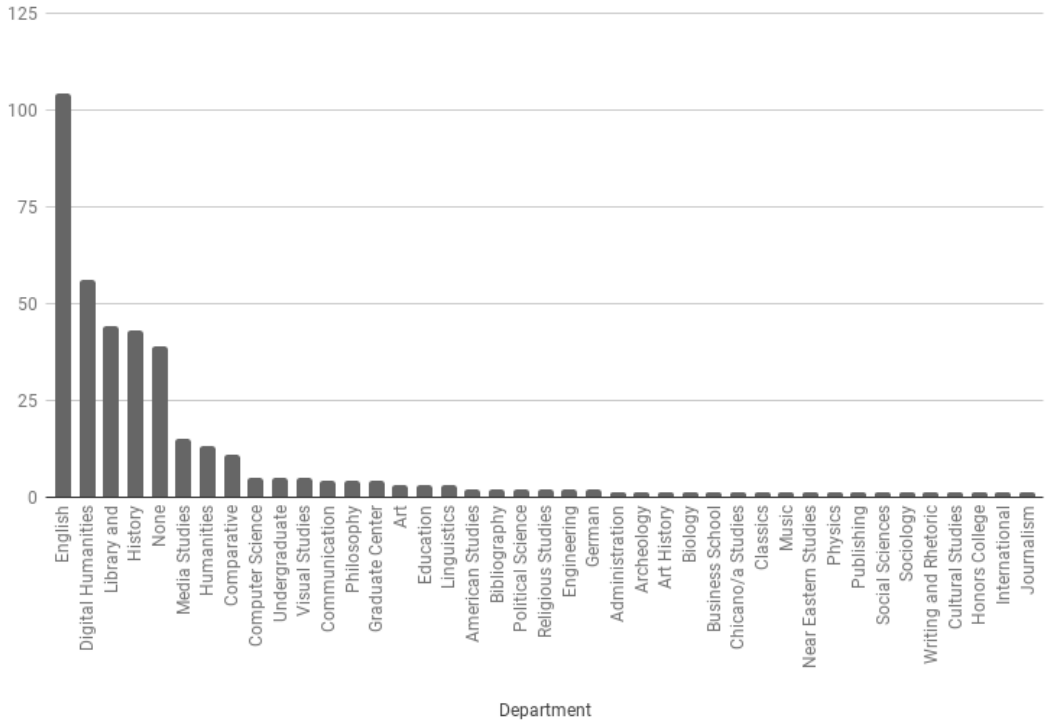

Figure 5. Number of definitions in corpus by department of author.

Even though about half of humanities PhDs go to women [Gender Distribution of Advanced Degrees in the Humanities] men wrote about 63 percent of the definitions in our corpus. Even more surprising is the way the male-female percentages break down by job title within academia (Figure 6). Male authors are overrepresented in the more senior academic positions. There are about four times as many definitional pieces written by male full professors than female full professors. Similarly, there are about twice as many definitions written by male Associate Professors, Assistant Professors, and Directors than female. On the other hand, in the pool of definitions written by junior academics, female authors outnumber male. There are more definitions written by female graduate students, postdocs, visiting scholars, and librarians than male. Overall, male full professors are the most prolific definers of the field, suggesting that while definitions are diverse in terms of academic position, they could be more equitable in their distribution among voices.

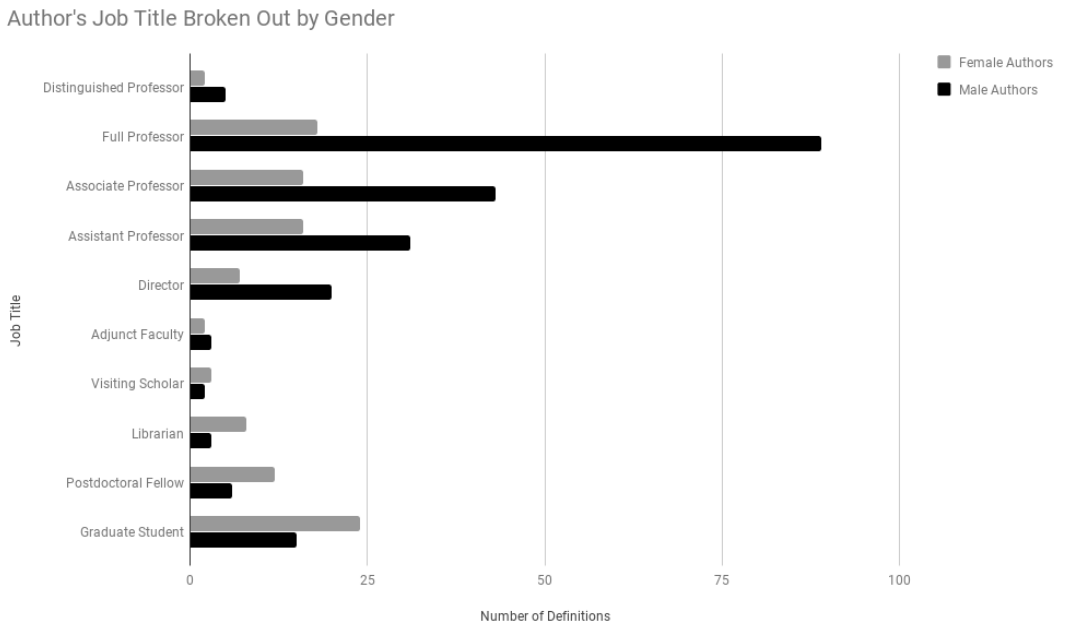

Figure 6. The job title/position of first authors broken out by gender.

In our analysis of the Terras/Nyhan/Vanhoutte bibliography, we only included short-form definitions (no full-length books). In our corpus of 334 definitions, over 200 were blog posts, followed distantly by 60+ academic journal articles. While there are some manifestos, popular articles, journal editorials, and book chapters, it seems that much work in defining digital humanities goes on in the gray literature of blog post, conference papers, and other informal communication. Interestingly, the more informal publication mode of blog posts, was just as uneven in terms of gender distribution of authors as the more formal peer-reviewed academic journal articles. There were 61 women and 138 men authoring blog posts in the corpus, while 17 women and 39 men were first authors in academic journal articles (in both cases, about 30% of the pieces had women as their first or only author). 25

This connects with Tara Thomson's warning that informal and nontraditional formats do not necessarily equate with egalitarianism. Her narrative of her experience at an unconference shows that those with less disciplinary knowledge may feel more exposed and that informal events tend to get led by those with more professional confidence [Thomson 2015]. Those that published definitions on the digital humanities on their blogs felt entitled to take a stab at defining the field on their own, without the feedback of peer review. Blogs might not be more egalitarian a means of publishing digital humanities definitions, at least in terms of gender. 26

Finally, our corpus was extremely homogeneous in terms of geographic location of the institution from which the authors were writing. Only three of the authors in the entire corpus came from cities outside of Europe, the United States, or Canada. Our corpus featured a definition by a scholar writing from The Centre for Internet and Society in Delhi, Universidad Nacional Autonoma de Mexico in Mexico City, and American University of Beirut in Beirut. As mentioned earlier, this lack of diversity in geographical origin of these definitions may partially stem from the fact that our corpus was entirely in English, but it also speaks to Melissa Terras' earlier critique that self-praise for inclusivity misses the fact digital humanities remains a "very rich, very western academic field" [Terras 2011]. Apart from the observation that nearly all the definitions were from the United States, Canada, and European nations, regional data did not lead to novel insights. While we analyzed topics by region of origin, no topic showed a preponderance of authors from a single nation or continent. We were expecting to find regional "flavors" of digital humanities, perhaps North American and European DH, respectively reflecting speculative vs. scientific modes of digital humanities, but mostly topics featured a mix of authors from North American and European institutions. If styles of DH vary by geographical region, our topic model was not a good instrument to detect this. 27

Conclusion

Defining digital humanities is an activity that shows no signs of slowing down. While proliferating definitions can be a good offset to gate-keeping, they can also lead to a sense of whiplash for those interested in entering the field. For our group of early-career academic co-authors, we experienced both excitement and worry at the prospect of using topic modeling to read a corpus of digital humanities definitions. We were excited to learn the technique of topic modeling, but the lack of consensus on how to use and interpret topics left us in a liminal space: we had successfully coded, and yet we had no idea if our efforts yielded meaningful results. It took interpretive work and insider DH knowledge to describe any topic's connection to the field. To us, the act of interpreting topics mirrored the push and pull that we see within the field, indicating that perhaps less emphasis should be placed on digital competencies and more emphasis on the step of interpretation that comes after the use of a digital tool. If established scholars not only provided manuals on how to use digital tools but also explained how they interpreted their results as meaningful, that would be an immensely useful step. 28

The topic model itself proved the least interesting and least difficult aspect of our work. Though the topic model indicated multiple routes that people tend to take when defining digital humanities, the really interesting stories emerged when we combined the topic model outputs with our painstakingly-gathered metadata or after hours of focused reading of the individual documents in the corpus. The model's breadth of topics did confirm our initial feeling of disorientation — not knowing if we were being welcomed or warned away from topic modeling and digital humanities in general. With definitions that ranged from advice on learning to code to documents that defined the field according to utopian values like "openness," the topic model did represent quite a range of ways the field is defined and encapsulated topics that covered both the push and the pull of digital humanities. While some define the field in relation to the subset of work that uses distant reading and text mining, others define the field as a community of people who self-identify as members based on shared values. Still others challenge the field to become more diverse even as they define it. Although temporal and spatial analyses of our 55 topics showed few interpretable trends in definitions across time and none across space, our analysis did produce some novel insights. We detected possible gender trends in the "distant reading" and "diversity" topics, indicating the push and pull of digital humanities may be felt disproportionately across gender differences, especially when it comes to particular methods or perspectives on what constitutes DH. In addition, the overall corpus of definitions contained gender and class imbalance, especially in the number of definitions written by tenured men. These insights are important to a field invested in inclusivity, yet grappling with a long history of failing on this front. If definitions of digital humanities are one of the genres that interested parties first encounter when beginning to research the field, then advancing definitions that both include and foster diverse perspectives is of critical importance. Knowing in which types of definitions women's voices are most underrepresented (and conversely, most robustly present) can help the field address problem areas. 29

In terms of technical competencies "required" for entry into DH versus opening up the field to individuals with a wide variety of experience-levels, it was refreshing that the most useful part of this study for ourselves was not the model itself, with its requisite R proficiency, but the part that one could do with a simple spreadsheet and google charts. What produced the most compelling insights was not the technical wizardry of Latent Dirichlet allocation but our communal spreadsheet. We considered producing the demographic graphs in this article in R as a signal of disciplinary belonging but ultimately decided that the fact that the most interesting figures were made not with a statistical programming language, but with built-in spreadsheet capabilities was an important aspect of what we wanted to say in this paper. As an exercise in learning a much-talked-about digital humanities technique as a foray into the field, we concluded that perhaps both the promise of the power of topic modeling and its dangers might have loomed larger in our imagination than in reality. The push and the pull of digital humanities can feel strong to early scholars interested in the field, but once you dive in, you might realize the waters are both less clear and less perilous than they seem from the shore. 30

Notes

[1] It is important to remember here that the topics and topic distributions spit out by a topic model are not stable entities that exist as concrete reality. These top twenty documents are not a final tally of the people who define digital humanities using distant reading. When choosing a different number of topics for the model, there might still be a recognizable "distant reading" topic, but the probability that each document features that topic will be different. So, for example, when running our topic model with 70 topics, we found there was still indeed a topic that

was recognizable as similarly “about” distant reading, but the top twenty documents associated with this “distant reading” topic featured 17 texts written by men and 3 texts written by women. There’s a danger in taking topics and the documents associated with them too much as reflections of reality, and, though any “distant reading” topic produced *may* be dominated by male authors, we are far from having shown that.

[2] We discussed, debated, and deliberated on how and whether to incorporate gender into our metadata. Most of the metadata we compiled was a snapshot at the time of publication of the particular piece. So while someone may be a distinguished professor now, if, at the time of their definition, they were a graduate student, then we recorded “graduate student” as their occupation. Gender is fluid too, in motion and on a spectrum, so we considered recognizing that gender identity can change by recording gender at the time of publication. But, while we wanted to recognize that gender is fluid, we wanted to avoid dead-gendering people. Also, we discussed how we could ethically determine someone else’s gender identity. Especially if the identity changed later, who is to say it hadn’t already changed in most circumstances at the time of publication? We decided to determine gender based on people’s own, current websites. Most of the time we were lucky in that the professional websites people would create for themselves, in their own voice, had bio-blurbs written in the third person. From this, we could get the author’s preferred pronouns, and use those as a key to gender identity. But when this was not available we did base our gender determination on names, which relies on stereotypes of what is a “male” vs. a “female” name, and inevitably led to some mistakes. We also debated not including gender at all at the risk of presuming someone’s gender identity or forcefully assigning gender to people. But on the other hand, our interest in definitions of digital humanities sprung from the disorienting push and pull we felt from the field, and a key point of that argument is that when bars are set for entry, they are felt especially by women. Ultimately, we feel it was important to include gender precisely for the results presented here: that more men mention distant reading when defining digital humanities than women. So despite our own discomfort, we have included a consideration of author gender when discussing these pieces.

[3] When we ran the topic model with 70 topics, we also looked at a topic similar to this one. The top twenty documents associated with the analogous topic of the 70 topic run had the same gender distribution: 13 female-authored and 7 male-authored pieces, but keep in mind the warning of the previous footnote — that topics are not stable entities that reflect the “reality” of the corpus, but are rather a heuristic which which to think.

Works Cited

Bailey 2012 Bailey, Moya Z. 2012. “All the Digital Humanists Are White, All the Nerds Are Men, but Some of Us Are Brave.” *Journal of Digital Humanities* 1 (1).http://journalofdigitalhumanities.org/1-1/all-the-digital-humanists-are-white-all-the-nerds-are-men-but-some-of-us-are-brave-by-moya-z-bailey/.

Barnett 2014 Barnett, Fiona M. 2014. “The Brave Side of Digital Humanities.” *Differences* 25 (1): 64–78. https://doi.org/10.1215/10407391-2420003.

Benzon 2014a Benzon, William. 2014a. “The Only Game in Town: Digital Criticism Comes of Age.” 3 Quarks Daily. May 5, 2014. https://www.3quarksdaily.com/3quarksdaily/2014/05/the-only-game-in-town-digital-criticism-comes-of-age.html.

Benzon 2014b . Benzon, William. 2014b. “Beyond Quantification: Digital Criticism and the Search for Patterns.” https://doi.org/10.13140/2.1.1812.0320.

Buurma and Heffernan 2018 Buurma, Rachel Sagner, and Laura Heffernan. 2018. “Search and Replace: Josephine Miles and the Origins of Distant Reading.” *Modernism/Modernity* 3 (1). https://modernismmodernity.org/forums/posts/search-and-replace.

Cohen 2008 Cohen, Daniel J. 2008. “Creating Scholarly Tools and Resources For the Digital Ecosystem: Building Connections in the Zotero Project.” *First Monday* 13 (8). http://firstmonday.org/ojs/index.php/fm/rt/printerFriendly/2233/2017.

Cong-Huyen 2013 Cong-Huyen, Anne. 2013. “#CESA2013: Race in DH – Transformative Asian/American Digital Humanities.” *Anne Cong-Huyen* (blog). September 24, 2013. https://anitaconchita.wordpress.com/2013/09/24/cesa2013-race-in-dh-transformative-asianamerican-digital-humanities/.

Cordell 2011 Cordell, Ryan. 2011. “More Hackety Hack, Less Yackety Yack: Ruby for Humanists.” “The Chronicle of Higher Education Blogs: ProfHacker” (blog). February 1, 2011. https://www.chronicle.com/blogs/profhacker/more-hackety-hack-less-yackety-yack-ruby-for-humanists/30175.

Davidson 2011 Davidson, Cathy. 2011. “Advice to DigHum Job Candidates: Don’t Lead With HTML.” *HASTAC* (blog). January 13, 2011. https://www.hastac.org/blogs/cathy-davidson/2011/01/13/advice-dighum-job-candidates-dont-lead-html.

Fish 2012 Fish, Stanley. 2012. “Mind Your P’s and B’s: The Digital Humanities and Interpretation.” *New York Times Opinionator* (blog). January 23, 2012. https://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-

humanities-and-interpretation/.

Gailey and Poerter 2011 Gailey, Amanda, and Dot Poerter. 2011. “Credential Creep in the Digital Humanities.” #alt-Academy: Alternative Academic Careers. May 6, 2011. http://mediacommons.futureofthebook.org/alt-ac/pieces/credential-creep-digital-humanities.

Gender Distribution of Advanced Degrees in the Humanities “Gender Distribution of Advanced Degrees in the Humanities.” 2017. Accessed May 29, 2018. https://www.humanitiesindicators.org/content/indicatordoc.aspx?i=47.

Goldstone 2017 Goldstone, Andrew. 2017. “Teaching Literary Data: What Makes It Hard · Preprint.” Andrew Goldstone. January 3, 2017. https://andrewgoldstone.com/blog/ddh2018preprint/.

Hawk 2013 Hawk, Brandon W. 2013. “DH Values Statement Planning.” Brandon W. Hawk. October 16, 2013. https://brandonwhawk.net/2013/10/16/dh-values-statement-planning/.

Jockers 2011 Jockers, Matthew. 2011. “The LDA Buffet: A Topic Modeling Fable.” *Matthew L. Jockers* (blog). September 29, 2011. http://www.matthewjockers.net/macroanalysisbook/lda/.

Jockers 2013 Jockers, Matthew. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, Urbana (2013). 152-3.

Jockers 2014 Jockers, Matthew. 2014. *Text Analysis with R for Students of Literature*. New York: Springer.

Jänicke et al. 2015 Jänicke, S., Franzini, G., Cheema, M.F., Scheuermann, G.. 2015. “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges.” *Eurographics Conference on Visualization: State of the Art Report*.

Kirschenbaum 2010a Kirschenbaum, Matthew G. 2010a. “Hello Worlds (Why Humanities Students Should Learn to Program).” *Matthew G. Kirschenbaum* (blog). May 24, 2010. https://mkirschenbaum.wordpress.com/2010/05/23/hello-worlds/.

Kirschenbaum 2010b Kirschenbaum, Matthew G. 2010b. “What Is Digital Humanities and What’s It Doing in English Departments?” *ADE Bulletin* 150: 55–61. https://mkirschenbaum.files.wordpress.com/2011/03/ade-final.pdf.

Klein 2018 Klein, Lauren. 2018. “Distant Reading after Moretti.” *Lauren F. Klein* (blog). January 10, 2018. http://lklein.com/2018/01/distant-reading-after-moretti/.

Koller 2015 Koller, Guido. 2015. “Pamphlet No 6 – Between Distant and Close Reading.” *We Think History* (blog). February 13, 2015. https://wethink.hypotheses.org/2085.

Lothian and Phillips 2013 Lothian, Alexis, and Amanda Phillips. 2013. “Can Digital Humanities Mean Transformative Critique?” *Journal of E-Media Studies* 3 (1). https://doi.org/10.1349/PS1.1938-6060.A.425.

Manovich 2016 Manovich, Lev. 2016. “The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics.” *Journal of Cultural Analytics*, May. https://doi.org/10.22148/16.004.

Mauri et al. 2017 Mauri, Michele, Tommaso Elli, Giorgio Caviglia, Giorgio Uboldi, and Matteo Azzi. 2017. “RAWGraphs: A Visualisation Platform to Create Open Outputs.” *In Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*, 28:1–28:5. CHItaly ’17. New York, NY, USA: ACM. https://doi.org/10.1145/3125571.3125585.

Meeks 2011 Meeks, Elijah. 2011. *Documents. Digital Humanities Specialist* (blog). February 1, 2011. https://dhs.stanford.edu/comprehending-the-digital-humanities/documents/.

Mimno 2013 Mimno,David. 2013. “A Wrapper around the Java Machine Learning Tool MALLET. Reference Manual” “CRAN”, https://cran.r-project.org/web/packages/mallet/mallet.pdf.

Paradise 2015 Paradise, Laurin. 2015. “‘When You Find Out What Digital Humanities Is, Will You Tell Me?’” *The Serials Librarian* 69 (2): 194–203. https://doi.org/10.1080/0361526X.2015.1036198.

Perez 2016 Perez,Annemarie. 2016. “Lowriding Through the Digital Humanities.” *Disrupting the Digital Humanities*. January 6, 2016. http://www.disruptingdh.com/lowriding-through-the-digital-humanities/.

Posner 2012a Posner, Miriam. 2012a. “Some Things to Think about before You Exhort Everyone to Code.” *Miriam Posner’s Blog* (blog). February 29, 2012. http://miriamposner.com/blog/some-things-to-think-about-before-you-exhort-everyone-to-code/.

Posner 2012b Posner, Miriam. 2012b. “Very Basic Strategies for Interpreting Results from the Topic Modeling Tool.” *Miriam Posner’s Blog* (blog). October 29, 2012. http://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-

the-topic-modeling-tool/.

Ramsay 2011 Ramsay, Stephen. 2011. “Who’s In and Who’s Out.” January 8, 2011. https://web.archive.org/web/20170426170232/http://stephenramsay.us:80/text/2011/01/08/whos-in-and-whos-out.

Ramsay 2016 Ramsay, Stephen.2016. “The Digital Naif.” Stephen Ramsay. November 5, 2016. https://web.archive.org/web/20161105014437/http://stephenramsay.us/2015/11/19/the-digital-naif/.

Ridge 2013 Ridge, Mia. 2013. “Beyond Code Monkeys: Recognising Technologists’ Intellectual Contributions.” *Open Objects* (blog). August 25, 2013. http://www.openobjects.org.uk/2013/08/beyond-code-monkeys-recognising-technologists-intellectual-contributions/.

Robertson 2014 Robertson, Stephen. 2014. “The Differences between Digital History and Digital Humanities.” *Dr Stephen Robertson* (blog). May 23, 2014. http://drstephenrobertson.com/blog-post/the-differences-between-digital-history-and-digital-humanities/.

Robertson 2016 Robertson, Stephen.2016. “Finding Questions As Well As Answers: Conceptualizing Digital Humanities Research.” *Dr Stephen Robertson* (blog). May 2, 2016. http://drstephenrobertson.com/blog-post/finding-questions/.

Scheinfeldt 2010 Scheinfeldt, Tom. 2010. “Stuff Digital Humanists Like: Defining Digital Humanities by Its Values.” *Found History*. December 2, 2010. https://foundhistory.org/2010/12/stuff-digital-humanists-like/.

Schmidt 2013 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/.

Schoch 2017 Schoch, Christopher. 2017. “Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama.” *DHQ: Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Spiro 2012 Spiro, Lisa. 2012. “‘This Is Why We Fight’: Defining the Values of the Digital Humanities.” In *Debates in the Digital Humanities*, edited by Matthew K. Gold. http://dhdebates.gc.cuny.edu/debates/text/13.

Terras 2011 Terras, Melissa. 2011. “Peering Inside the Big Tent: Digital Humanities and the Crisis of Inclusion.” *Melissa Terras’ Blog* (blog). http://melissaterras.blogspot.com/2011/07/peering-inside-big-tent-digital.html.

Terras, Nyhan, and Vanhoutte n.d. 2013 Terras, Melissa, Julia Nyhan, and Edward Vanhoutte. 2013. *Defining Digital Humanities: A Reader*. Terras, Melissa, Julia Nyhan, and Edward Vanhoutte. n.d. “Further Reading.” *UCL Defining Digital Humanities* (blog). Accessed May 21, 2018. https://blogs.ucl.ac.uk/definingdh/further-reading/.

Thomson 2015 Thomson, Tara. 2015. “What Is the Difference between ‘doing Digital Humanities’ and Using Digital Tools for Research?” “London School of Economics Impact Blog” (blog). February 11, 2015. http://blogs.lse.ac.uk/impactofsocialsciences/2015/02/11/digital-humanities-unconference-exclusion-access/.

Turan 2016 Turan, Julia. 2016. “How Exactly Do the ‘digital Humanities’ Mix Science with the Arts?” *Our Cells Our Selves*. February 9, 2016. https://juliaturan.com/2016/02/09/how-exactly-do-the-digital-humanities-mix-with-the-arts/.

Underwood 2012 Underwood, Ted. 2012. “Topic Modeling Made Just Simple Enough.” *The Stone and The Shell* (blog). April 7, 2012. https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

Underwood 2018 Underwood, Ted. 2018. “A Broader Purpose.” January 4, 2018. https://tedunderwood.com/2018/01/.

Weingart 2012 Weingart, Scott. 2012. “Topic Modeling for Humanists: A Guided Tour.” *The Scotbott Irregular*. July 25, 2012. http://www.scottbot.net/HIAL/?p=19113.

Weingart 2016 Weingart, Scott B. 2016. “Representation at Digital Humanities Conferences (2000-2015).” *The Scottbot Irregular*. March 22, 2016. http://scottbot.net/representation-at-digital-humanities-conferences-2000-2015/.

Weingart and Meeks 2012 Weingart, Scott, and Elijah Meeks. 2012. “The Digital Humanities Contribution to Topic Modeling.” *Journal of Digital Humanities* 2 (1). http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/.

Whitson 2011 Whitson, Roger. 2011. “Python for Humanists 1: Why Learn Python?” *Roger Whitson* (blog). December 7, 2011. http://www.rogerwhitson.net/?p=1260.