DHQ: Digital Humanities Quarterly

Preview
2025
Volume 19 Number 3

Slow, Painful and Expensive: Current Challenges in Text-Mining Corpus Construction for the Digital Humanities

Matt Warner <mattgw_at_stanford_dot_edu>, Stanford University https://orcid.org/0000-0003-4469-481X
Nichole Nomura <nnomura_at_uwyo_dot_edu>, University of Wyoming https://orcid.org/0000-0002-5624-9011
Carmen Thong <tcarmen_at_stanford_dot_edu>, Stanford University https://orcid.org/0000-0002-4576-1415
Alix Keener <alixkee_at_stanford_dot_edu>, Stanford University https://orcid.org/0000-0002-5606-9176
Alexander Sherman <alexander_dot_sherman_at_austin_dot_utexas_dot_edu>, University of Texas, Austin https://orcid.org/0000-0002-6811-1685
Gabi Birch <gab03_at_stanford_dot_edu>, Stanford University https://orcid.org/0009-0001-4417-4024
Maciej Kurzynski <maciejkurzynski_at_ln_dot_edu_dot_hk>, Lingnan University https://orcid.org/0009-0006-5466-2423
Mark Algee-Hewitt <malgeehe_at_stanford_dot_edu>, Stanford University https://orcid.org/0000-0002-2628-588X

Abstract

The process of assembling corpora for text-mining-based Digital Humanities projects is a crucial and yet frequently overlooked aspect of the research process. Often complicated by text availability and cost, as well as legal restrictions on in-copyright text, DH scholars frequently resort to “found” corpora marketed to libraries by publishing companies or questionably sourced corpora that inhabit legal grey areas. While such corpora have led to methodological developments in the field, there is a general sense that the biases of these corpora and the inability to share their raw data have made them imperfect vehicles for large-scale critical claims in the humanities. Recent developments, however, suggest that this situation may be changing. In the United States, the 2021 text and data-mining exemption to the Digital Millennium Copyright Act (DMCA) has promised to improve the viability of bespoke corpora for text-mining research. In this paper, we put these improvements to the test, reporting on our efforts to source a relatively small corpus of literary theory monographs. Focusing primarily on born-digital works and operating under all of the practical and legal constraints dictated by the exemption to the DMCA, we sought to assemble a corpus of 402 pre-selected theoretical works. We found that, despite the recent legal changes, and even with extensive support from a well-resourced library, it remains overly difficult to assemble a pre-selected corpus of scholarly works, even under ideal financial and institutional conditions. While scholars outside of the United Stats will face somewhat different legal restrictions on the collection of electronic texts than we did, we found that many of the obstacles we faced were practical, rather than regulatory, and in many cases, we found that scanning books was the easiest and most efficient route to digital versions of the texts we sought.

Introduction

Over the past fifteen years, subfields in the Digital Humanities (DH) that focus on computational text analysis have grown rapidly. But at the same time, an imbalance has emerged between the development of new computational and statistical methods for text analysis and the availability of material on which this analysis can be performed. While dataset publication and citation has increasingly become a standard in the sciences,[1] scholars in the digital humanities working on text, especially in-copyright text, face significant barriers to dataset publication and, subsequently, data sharing.[2] Whether under the rubric of Cultural Analytics, Computational Literary Studies, Stylometry, Distant Reading or Text Mining, the gains in these subfields have largely been methodological [Jänicke et al. 2015], [Underwood 2019], [Gius and Jacke 2022]. The large-scale resources that have enabled work in these fields, particularly for Anglophone texts, are the same that existed over a decade ago [Gale Cengage 2014], [Project Gutenberg n.d.].[3] These pre-assembled corpora, often sold and distributed by large publishers despite being mostly out of copyright, come with a host of challenges that complicate the early promises of statistical rigor and representative sampling made by DH practitioners [Jockers 2013] [Ramsay 2011]. The challenges of assembling relevant corpora for particular projects have been relegated to the bibliographical, archival, and, for in-copyright works, legal domains, which have long been subordinated to the interpretive and methodological concerns of the field [Bode 2020]. And while there exists a tacit assumption among practitioners that these challenges are slowly (if unevenly) being addressed, making large corpora of texts progressively more available, this availability continues to lag other developments in the field, and the gap between developments in methods versus in corpora may be widening. 1

This article focuses specifically on these issues, particularly in the legal domain, as they play out in the American context, where the 2021 introduction of the Text and Data Mining (TDM) exemption to the Digital Millennium Copyright Act (DMCA) represents a crucial step towards overcoming the legal dimension of these challenges.[4] While the implications of copyright for digital humanists are widely understood and discussed — if certainly not resolved — the work we present here focuses on the interplay between the licensing of ebooks and their protection via DRM. Unlike printed matter, whose use is restricted only by copyright, ebooks are subject to license agreements which restrict the ways in which they may be used, and often specifically disallow text-extraction. Even when such use is not prohibited, however, ebooks are generally protected from text-extraction by a variety of technical measures, generally referred to as “digital rights management” (DRM). The circumvention of DRM, in turn, is typically legally proscribed — in our case by the Digital Millenium Copyright Act in the United States, and in other nations by relevant local law (for example, in the European Union, the Directive on Copyright and Related Rights in the Digital Single Market requires EU member states to forbid the circumvention of DRM, stating in its preamble that “The protection of technological measures … remains essential to ensure the protection and the effective exercise of the rights granted to authors and to other rightholders under Union law”.[5]) 2

Unfortunately, compared to even the patchwork standardization of copyright (attempts at the harmonization of which date back to the 1886 Bern Convention), frameworks for licensing and DRM are significantly more varied, and many of the legal details of what we present here are specific to the US context. While the particular legal issues facing scholars will vary from jurisdiction to jurisdiction, however, we found significant practical and technical barriers to our corpus-compilation efforts that are, we believe, relevant and significant to scholars working under even considerably different legal frameworks. In our American case, the new TDM provision of the DMCA decriminalizes the act of extracting text protected by DRM software in the US for text and data mining research, and its explicit 3 goal is to allow scholars to build responsible corpora for research.[6] (Similar exemptions exist in other jurisdictions; for example article 3 of the Directive on Copyright in the Digital Single Market specifically protects “reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.” In addition to letting scholars assemble and share their own digital in-copyright corpora, the TDM exemption, and other similar frameworks, also create an opening for the discussion of the collection, use, and distribution of in-copyright texts.[7] As such, DH scholars working in the US have looked with excitement to these new provisions as a potential watershed in making possible the creation of more deliberate, representative, and accessible corpora of the kind previously only possible using out of copyright material [Bode 2018].

Yet the particulars of this exemption — in its current form — leave many of the same legal hurdles in place, particularly for scholars working independently or at smaller institutions without access to significant research budgets and well-resourced IT departments and legal teams. In order to work within the boundaries set by the exemption, scholars face new restraints around sourcing, processing, storing, and above all, sharing, these materials, many of which contradict the norms and aims of the field. Similarly, the easing of legal barriers to assembling corpora of in-copyright texts does not address the other kinds of challenges practitioners in DH face when sourcing texts. While the gains made possible by the passing of this exemption are a necessary step towards helping scholars research copyrighted materials, the exemption itself is anything but a sufficient solution. 4

Our goal in this paper is to document our attempt to take advantage of this new legal framework to build a bespoke corpus of largely in-copyright, born-digital texts for the purpose of text mining. Readers can access the corpus manifest. Starting not from a found corpus of pre-assembled works, but instead from a list of desired texts around a single subject, we sought to collect a complete corpus of 402 works of twentieth century literary theory, following all the rules set by the TDM exemption. In the end, we have discovered that even when supported by both a large, wealthy university library and a grant specifically funding the assembly of such a corpus, building a born-digital corpus remains difficult, if not impossible. Our struggles to do so implicate not just the legal mechanisms of copyright, licensing and DRM, but also the publishing and book selling industries, the collection and purchasing protocols of university libraries, and the general availability of texts themselves. We are particularly dispirited to report, then, that even armed with a considerable quantity of grant money, a supportive library, and a team of researchers working together on this project, we have found the acquisition of texts under these constraints to be slow, painful, and expensive. If efficiency had been our goal, it would have been unquestionably faster for us to cull our corpus of outrageously expensive titles, and then purchase, disassemble, and scan every book from a physical copy — not, alas, a ringing endorsement for the new legal regime. It is clear that the challenges of corpus construction extend far beyond the legal domain and into all facets of book publishing, selling, and archiving. 5

Project Overview

The corpus whose assembly we document here is composed of literary theory, collected in support of a research project examining the movement of ideas, concepts, and discourses between literary theory and literary criticism. Alongside the hand-curated corpus of literary-theoretical texts detailed here, this project will consider a collection of literary monographs based on recently recovered scans of books created by the Google Books initiative (see “Google Books Return”, below). To digital humanists working with textual corpora, this corpus will be a familiar example of the “found-corpus,” a pre-existing set of texts only partially aligned with our research goals, and marred by significant bibliographic and representational limitations. The principle of selection behind this set of critical texts is pragmatic in the extreme: it is simply the books that Stanford held which Google had not already digitized from another library’s collection. The scale of the collection is considerable, including 24,981 English-language titles under the Library of Congress “P” classification (“Language and Literature”) and its subclasses on literature and literary criticism. Our library’s records, moreover, distinguish between primary source and other texts, allowing us to identify, request, and access some 4,395 works that we were relatively confident represented chiefly works of literary criticism and analysis written in English. We cannot, and will not, claim that these texts are in any way representative of the field. In fact, based on our exploration of the overlap the between the texts we selected for our theoretical corpus and the works digitized in this Google Books data, we are nearly certain that this collection of texts is deeply idiosyncratic, missing exceptionally important works, and overdetermined by the particular strengths and weaknesses of the Stanford library’s collections. 6

We introduce this second corpus of literary criticism as a counterexample to the purpose-built in-copyright corpus that is our primary object of discussion. We will therefore not explore these works of criticism in this article, nor analyze the texts of literary theory that we have collected in any way. Instead, our focus here is methodological, foregrounding issues of corpus construction — how we found, obtained and tracked the modest number of texts we required, and the many obstacles we encountered in doing so. Many of these obstacles lie in the blurry interface between researchers and librarians, and we hope that by presenting them here, we can draw attention to the particular needs and expectations of each of these two groups and offer some insight into how best to negotiate this kind of collaboration. Of course, some of the issues we encountered were purely of our own devising, and many were, we would like to think, beyond our control, if sometimes retrospectively predictable. 7

Methods

Overview

Corpus Selection

For the research project underlying this corpus, we selected eight theoretical fields to focus on — feminism, Black studies, Marxism, psychoanalysis, post-structuralism, postcolonial studies, narratology and Russian formalism. These fields, we felt, had a variety of different relationships to literary criticism that would facilitate comparative analyses in our later research. Some felt distinctly dated, others seemed to us to have more lingering influence; some were older, dating to before the institutionalization of literary studies, others were more recent. This heterogeneity was an asset, too, insofar as we expected that the availability of texts from each of these schools would vary according to their dates, languages of composition, and other similar factors. While the specific selections we made are not the focus of the present work, we wish to highlight the necessarily incomplete nature of our case study. Comprehensiveness is not a reasonable standard for something as dynamic and contentious as twentieth-century theory, and so we do not want to even inadvertently give the impression that we believe that these fields and our resulting corpora represent “literary theory” writ large. 8

Within each of these fields, we carefully selected a few dozen works, mostly monographs, which we felt were important to that field and which we hoped would shed light on the questions we cared about.[8] We consulted tables of contents from anthologies, introductory texts, and readers, and relied in large measure on memory, expertise, and debates among ourselves and with outside scholars of these fields. We sought a corpus that, even if incomplete, represented a broad swath of texts that we felt were important to each field and which we hoped would facilitate meaningful, complex findings in our later research (see Table 1). In some — if not all — of our categories, we also had to make decisions as to whether a text was theory or literary criticism. How much of the monograph could be concerned with 9 interpreting primary material before it would be relegated to criticism rather than theory? What if an especially influential concept was introduced, but nestled within a carefully applied close reading? Importantly, however, we did not make our selections with any eye to the availability of digital (or other) versions of these texts: we wanted very specifically to explore the accessibility of literary-academic titles for text mining research, and we have been able to locate a text — albeit far from always a digital one — for nearly every text we initially listed.

Field	Number of Texts
Black Studies	76
Feminism	50
Marxism	51
Narratology	66
Post-Structuralism	36
Postcolonial	75
Psychoanalysis	41
Russian Formalism	24

Table 1. Table : number of texts per theoretical field. Some texts are included in multiple fields

10 A significant preliminary obstacle to our work was simply identifying books that corresponded to the texts that we had selected. Here, we ran into bibliographical problems that will be familiar to anyone who has assembled a corpus from a list of titles, as well as several new issues. In many cases, these problems stem from the particular features of the academic books in our corpus, which overrepresent translated works and often feature lengthy titles prone to abbreviation, sometimes with slightly amorphous punctuation dividing them (e.g., we initially included *Anti-Oedipus*, *A Thousand Plateaus*, and *Capitalism and Schizophrenia* in our corpus before resolving these into *Anti-Oedipus: Capitalism and Schizophrenia* and *A Thousand Plateaus: Capitalism and Schizophrenia*). As a preliminary step towards the compilation of our corpus, we sought to link each of our texts to a library catalog entry to find out if our library provides access to the text in question in an electronic form (or, in some cases, if our library has the text in *any* form).

11 Once we had checked for library catalog entries for each work, we consulted the corresponding MARC (Machine Readable Catalog) record, a standard for the representation and communication of bibliographic and related information in machine-readable form [Library of Congress n.d.] for each text, which our library makes available in its online catalog through what it terms “librarian view”. From this record, we then recorded information about whether a text had been included in Google’s digitization initiative, and, for texts where a library ebook was available, the details of that availability (this information was not available in the more public facing catalogue entry). Finally, we used this information to develop a preliminary plan for obtaining each of our texts.

Acquisition Process

12 We selected our acquisition methods with a careful eye to copyright law, the DMCA, and the license agreements that govern ebooks. While we are aware that a case can be made that fair use of copyright-protected texts produced via another party’s independent violation of the DMCA may not violate either copyright or the DMCA, this method represents the kind of “grey area” that we are avoiding.[9] The most direct route to such a legally unburdened corpus would have been to digitize each of the works ourselves, but we prioritized obtaining born-digital versions of our texts wherever it was possible to do so, both because — in principle — this offers a considerable saving of time and labor, and because we wanted to acquire the cleanest possible texts with the fewest errors introduced via either scanning or processing the texts (with optical character recognition [OCR]). For our research needs, our ebook processing and OCR-based pipelines were designed to produce plaintext UTF-8 encoded files, though other project might prefer to work with annotated data (e.g. XML or TEI), or even raw page images, and we anticipate that many of the issues we discuss here would apply in these cases as well.

13 Initially, we conceptualized our acquisition of texts quite simply (see Figure 1). We proposed to turn first to library-owned ebooks; then purchase ebooks as needed to fill in any remaining holes in the corpus; leaving, we naively assumed, a small residuum of texts to be scanned by the library or project members (e.g., long-out-of-print texts unavailable as ebooks).

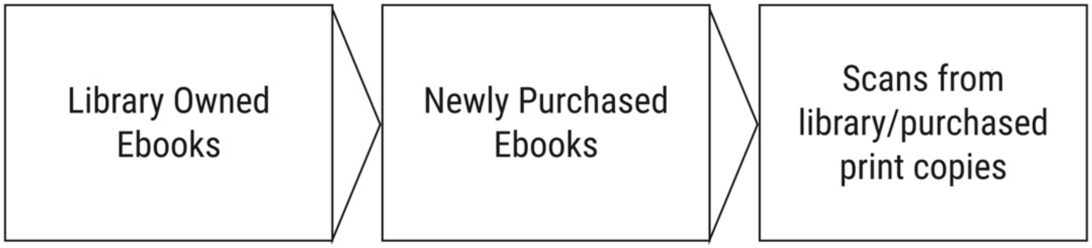

Figure 1. Initial process for acquisition of texts

14 Our workflow quickly grew vastly more complicated. We found that the library *owned* (as opposed to licensed) electronic copies of fewer texts than we expected, lacking ebook copies of widely read and cited theoretical texts. When purchasing ebooks, we were confronted by spotty availability, and we discovered later in the process that the availability of ebooks to the library changes relatively frequently. When we turned to digitization, we discovered that not all the texts which Stanford had provided to Google as part of the Google Books initiative had actually been scanned. Even with printed texts, there were unforeseen complexities, and as our scanning queue grew, we ultimately divided it into two different workflows. The result, as shown in Figure 2, was a far more tangled, burdensome, and slow process than we had anticipated, involving at least five and as many as eleven steps to go from identified text to file in the corpus.

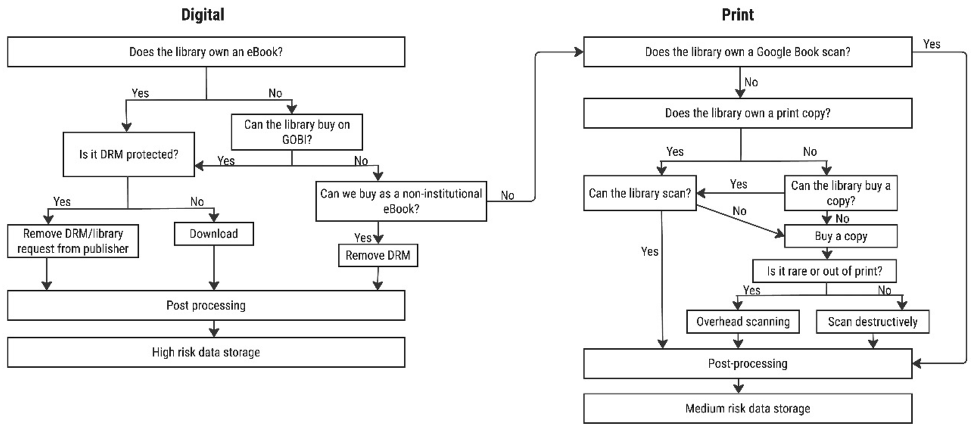
Digital
Print
Does the library own an eBook?
Yes
No
Can the library buy on GOBI?
Is it DRM protected?
Yes
No
Can we buy as a non-institutional eBook?
No
Yes
No
Remove DRM/library request from publisher
Download
Yes
Remove DRM
Post processing
High risk data storage
Does the library own a Google Book scan?
Yes
No
Does the library own a print copy?
Yes
No
Can the library scan?
Yes
Can the library buy a copy?
No
No
Buy a copy
Yes
Is it rare or out of print?
Yes
No
Overhead scanning
Scan destructively
Post-processing
Medium risk data storage

Figure 2. Final workflow for text acquisition

Types of acquisition

Electronic Books

Whenever possible, we prioritized the acquisition of texts via library ebooks. To remain in compliance with the TDM exemption to the DMCA, the text must be owned by Stanford University (or perpetually licensed), rather than available via subscription, as the exemption to the DMCA excludes ebooks acquired via subscription services. Of our 402 titles, 116 — slightly more than one quarter — were available as non-subscription ebooks via the library. Of these, 51 were available as full-text download (almost always as a PDF); while the remainder were protected by DRM or available only as chapter-by-chapter individual PDF downloads. Of these 65 DRM-protected or fragmented texts, 47 items were published by either University of Minnesota or Duke University Presses. In these cases, we took advantage of a previously unused term of our libraries' license agreements with the presses and/or distributors of the ebooks (entities such as ProQuest), and asked our library to submit requests for DRM-free complete copies of each of the texts for text and data mining purposes.[10] One publisher mentioned they had not received a similar request before, but after some back and forth to clarify what we were requesting, we eventually received delivery of the full-text PDFs. 15

In cases where the library did not already own or perpetually license an ebook, we requested that the library purchase the ebook if possible. The acquisition of ebooks by libraries is somewhat different from the processes by which individuals purchase the rights to such works. In Stanford's case, library ebook acquisitions are made via GOBI, a vendor system owned by EBSCO, the library services corporation likely more familiar to scholars for its databases. Ebooks can be purchased through GOBI (or similar competing systems) by libraries according to a variety of schemes determined by their publishers and/or platform (such as Project Muse) — for example, cheaper single-user vs more expensive unlimited-user licenses, DRM-free (for a higher price), etc. Regardless of the attendant terms, ebooks purchased in this way are considerably more expensive than those purchased by individual users, and, for libraries, also more expensive than print [Bailey et al. 2017]. While we cannot share exact costs for this project, the library purchased approximately 130 ebooks, with ebook prices for academic libraries ranging from USD $55-$200, averaging approximately $150 per book, for an approximate total cost of around $19,500. Originally, we had intended to finance these purchases via grant funding, but we were fortunate to discover that the relative importance of many of the texts we had selected meant the library was able to use our project to help justify a long-term shift in its collection. For many of the texts we had selected, the library has long had multiple print copies (especially for titles originally published before the digital era) but in many cases, in the ebook era, when it can afford to do so, our library would prefer to hold a single — perhaps less frequently consulted — print copy alongside a perpetually licensed ebook. 16

Google Books Return

When we could not obtain born-digital copies of the texts we needed, we turned to digitization, initially examining the set of texts that Stanford received from Google in exchange for participation in their Google Books digitization project. While these texts are older scans, and the provided OCR does not match modern standards, we deemed this an acceptable trade-off for the labor savings. (Of course, these digital versions were produced using extensive human labour, as works like Andrew Norman Wilson's *ScanOps* (2012) and the research of [Chalmers and Edwards 2017] make clear.) Unfortunately, the coverage of the Google Books collection is far from comprehensive, and even in the case of texts that Stanford provided to Google, there was no guarantee that they were ultimately digitized from Stanford's copy (the terms of the agreement between Stanford and Google give us access only to those works actually digitized from Stanford's own collection, not to all works provided for digitization).[11] For scholars fortunate enough to be working at institutions whose collections Google digitized more extensively, this may be a more fruitful source of digitized texts (if their libraries are equipped to provide access, that is). In the end, we were able to obtain 35 titles in this way, from a total of 53 provided to Google for potential digitization. 17

Scanned Books

A significant number of texts in our final corpus were, ultimately, derived from physical copies we purchased specifically for this project. In the end, we turned to self-digitization with greater frequency than we had initially anticipated, simply because not all texts we wished to include in our corpus were available in digital formats. In addition to relatively obscure items, including items out of print since the widespread adoption of ebook publication, we found that born-digital copies of some prominent works of theory simply did not exist (e.g., *S/Z* by Roland Barthes). Additionally, sometimes digital versions of texts *were* available, but not from a publication that offered ebook license terms allowing us to extract the text: the DMCA exemption does not in any way affect the contractual terms of ebook license agreements (notably, this is the case for many direct-to-consumer ebook distributors such as Amazon, although we were able to purchase a small number of non-institutionally licensed ebooks directly from some university presses). In these cases, we were also forced to fall back on scanning the texts. 18

We divided our scanning into several batches, according primarily to the expense and rarity of the physical book in question (see Figure 3). In this, we were once again significantly aided by the library, whose Digital Production Group agreed to scan a significant number of items (86) using their large-scale scanning system. In cases where a library copy of the item was available, this semi-automated scanning was our preferred option; however, the relatively slow pace of this process (which needed to contend with other, higher-priority digitization needs) meant that we sometimes opted to scan items ourselves. For items in print and available in relatively affordable editions, we opted for destructive scanning, in which the books are disassembled and scanned page-by-page using a feed scanner. For the majority (33 of 35 items) of our destructive scanning, we used a commercial scanning service. These companies offer a variety of services, typically priced per-item (e.g., the company we selected charges more for higher quality scans, OCR, or the option of receiving items directly from booksellers), and we found that the cost compared favorably to the cost of the books themselves. Additionally, however, we destructively scanned two items ourselves (both texts we had originally anticipated obtaining digitally). For these books, one of our project members used a bandsaw to remove the books' spines and we scanned the pages using a sheet-feed scanner. Since it requires some care and experience to produce cleanly cut pages that scan without issue in this way — not to mention access to power tools (or, ideally, a more specialized guillotine cutter such as would be used for bookbinding) — we think that for most projects, small lots of books are most practically scanned non-destructively, with larger lots scanned destructively by professional services. Finally, we scanned two out-of-print books manually using an overhead scanner. 19

Whatever the manner of scanning used, we converted our digital images to plain text using ABBYY FineReader, which, for our relatively modern texts in English, produced the best results of the software we tried.[12] Since FineReader is commercial software (and moderately expensive at that), we first tried a variety of other alternatives, including the OCR engine embedded in Adobe Acrobat (also commercial, and more expensive than FineReader, but something that some scholars may already pay for), as well as the open-source Tesseract, both of which performed noticeably worse than FineReader (we were particularly disappointed by the results from Tesseract). Depending on the nature of texts being digitized, as well as the downstream use cases (e.g., some users will prefer raw xml OCR output to plaintext), we anticipate that the relative tradeoffs of convenience, cost and performance will vary from project to project and we encourage anyone considering these questions to test a variety of options. 20

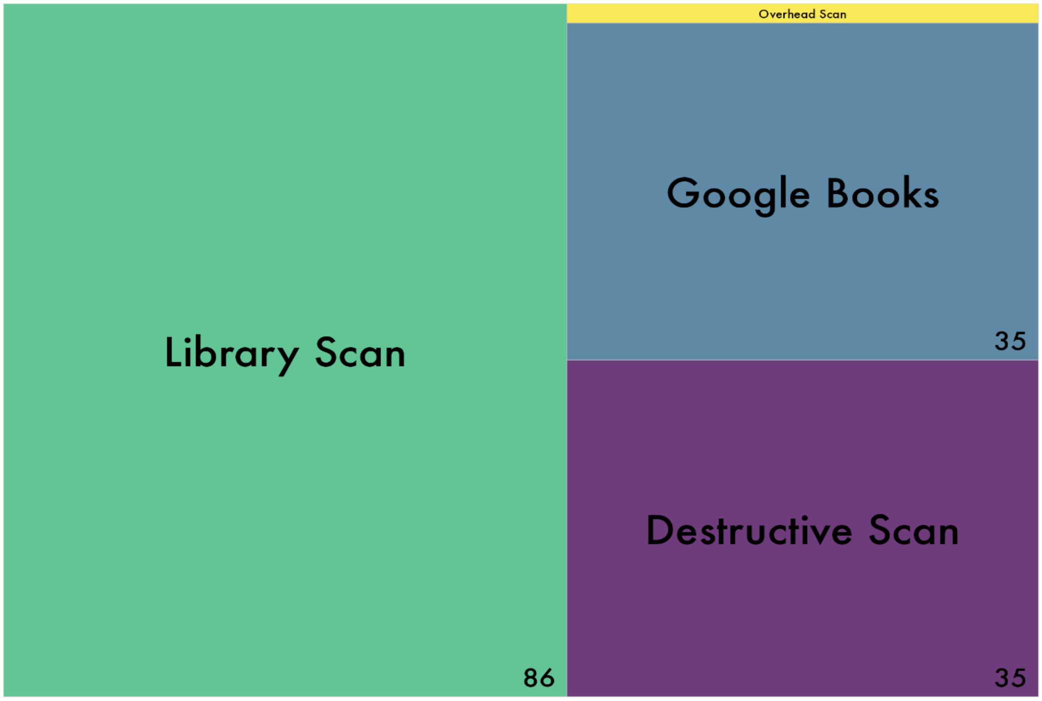

Figure 3. Sources of digitized books. Readers can download the code plots.

Articles

While our corpus of theory was assembled with a deliberate focus on the scholarly monograph, we included a number of articles and book chapters as well, in addition to a small number of miscellaneous shorter works (e.g. pamphlets, transcripts of speeches). Russian Formalism, for example, features significantly fewer monographs than other schools of literary theory: important works by Boris Eichenbaum, Yury Tynyanov, and Viktor Shklovsky translated into English are essays, not full-length books. While in a small number of cases we scanned older works from journals or collected editions, in most cases, we were able to obtain these texts in their entirety via standard scholarly databases,[13] confirming one of our project's premises, that computational metacritical and scientometric work on articles is considerably easier than similar projects focused on the monograph. 21

Open Access and Works in the Public Domain

A small but important part of our corpus consists of texts not protected by copyright, a group that includes both texts whose circumstances of publication mean that they are no longer protected in the United States (a category that will have a different parameters for scholars working in countries with different copyright terms), as well as texts that have been proactively released into the public domain or made available under other open-access (OA) terms.[14] Unfortunately, OA materials for human reading are not the same as OA materials for text-mining and may not be easy to obtain in a format compatible with computational study. Some OA texts were not available for complete download and required chapter-by-chapter navigation. (We encountered this same phenomenon with non-OA texts, where it seemed to be a text-protection mechanism, but many OA platforms appeared to do so simply to break longer texts into human-friendly units). While some OA 22 platforms were user-friendly for both text-mining and reading in a range of formats, this was far more often the case for public-domain works, which were usually already available for direct download in raw text form, albeit with rather uneven quality.

While OA and public domain texts are legally (and often practically) easier to work with than those protected by copyright, we found that identifying these texts as such was not always a straightforward process. OA monographs, in particular, are not always cataloged by libraries, meaning individual scholars must locate an OA copy and, more importantly, know that an OA copy might exist, as publishers have a financial interest in making OA versions of their texts as obscure as possible. We were surprised to find, moreover, that even the category of public domain texts could occasionally pose problems. In particular, while it is in principle relatively straightforward to identify texts published before the date of copyright as belonging to the public domain, translations (protected by copyright of their own) are not always as easy to date or identify. For texts like Vladimir Lenin's *Imperialism, the Highest Stage of Capitalism*, underlying copyright law regarding some translations produced in the earlier years of the USSR may have allowed us to use copies freely available online. But the complexities of international copyright law sometimes led us to purchase such texts anyway when we found the legal guidelines undecipherable, especially since these texts were cheap and frequently still in print. 23

Data management

Data Security Considerations

The DMCA exemption contains the stipulation that data obtained via the circumvention of DRM for the purposes of text and data mining be secured using "effective security measures", which are defined as those agreed to by the university and the rights-holder, or, absent such an agreement, those measures "that the institution uses to keep its own highly confidential information secure."[15] We anticipated from the outset of this project that this provision would prove significantly burdensome, and we found that this indeed was the case. Much of the burden has been interpretative: the text of the exemption does not clearly specify what kinds of measures should be taken by the institution to secure the data in question. Universities typically hold a wide variety of confidential information — employee records, medical records, banking details, and more — and it is not clear what standards should apply in the case of text mining as opposed to what [Borgman 2018] terms "grey data". Moreover, in developing workflows that complied with the exemption, we found not only that existing university guidance on data security and risk did not anticipate our particular situation, but also that the very tools set up by the university to help scholars assess their data risks were fundamentally unsuited for our particular situation. Our approach was ultimately determined by a series of in-person meetings with the university's general counsel and information security office, an additional expenditure of the project team and university's labor to comply with the exemption. These meetings led to the determination that data obtained via the DMCA exemption would be treated as "high risk", a data category at our institution that includes, among other things, personally identifiable health information, USA Social Security numbers, and credit card numbers.[16] 24

As digital humanists accustomed to working on our own devices and with varied technical skill sets, we have found the requirements for processing and collecting high-risk data challenging. While our university's data management policies allow personal computers to be configured for high-risk data and for such data to be stored locally, doing so requires a degree of oversight and monitoring that we felt was broadly unsuitable for personally-owned devices that are also used for non-university purposes. While it is our intention to perform much of the downstream analysis of our texts using our university's secure high-performance computing (HPC) environment, this was not a suitable solution for collection of our texts.[17] First, such an environment imposes a significant technical burden on a task that, otherwise, could be done without coding or command-line experience, familiarity with Amazon Web Services (used by our secure HPC environment), and similar skills. Second, and more importantly, the workflow we adopted to remove the DRM from protected ebooks was one that, to the best of our knowledge, was simply not compatible with such a computing environment. Instead — again attesting to the indirect expenses imposed by the TDM exemption — one of our project members furnished us with a spare computer which we configured for high-risk data. 25

Processing Ebooks and Removing DRM

Users who, pursuant to any of the previously granted exemptions of section 201 of the DMCA, have attempted to circumvent DRM for any of the allowed purposes have long reported technical difficulties.[18] In many cases, the lack of legitimate tools for such a task was a significant obstacle. The tools developed for removing the DRM from ebooks are, generally speaking, intended for piracy, not fair use, let alone use protected by an exemption to the DMCA. Because the development of such tools is intended to facilitate the violation of copyright, and is in violation of the DMCA (i.e., unlawful in the USA), their availability is subject to sudden takedown and cannot be guaranteed in the long term. Additionally, such tools are engaged in a kind of arms-race with the developers of DRM technology, and there is no guarantee that, at any given time, these tools will remain operational. 26

All the ebook DRM that we encountered in this project took the form of works distributed as Adobe Digital Editions. To extract the text from such works, we first downloaded them and used Adobe Digital Editions to extract a protected PDF file for each text. We were then able to remove the protection from this PDF using a plugin for the open-source ebook manager Calibre — it is this plugin specifically whose development is unlawful. Subsequently, Calibre's built-in ebook conversion tools allow the PDF to be converted to a text file. Technically, this process is relatively straightforward, but its reliance on Adobe Digital Editions poses two significant problems. First, there is no command-line interface for this piece of proprietary software, making it more challenging to use as part of a larger workflow. (Calibre does provide a command line interface.) Second, Adobe Digital Editions is limited to Windows and MacOS, precluding its use on most Linux-based HPC or remote computing environments. For our project, the limited availability of ebooks to process via the exemption meant that these limitations were principally of issue because of the aforementioned data management restrictions; for a larger project, however, they would necessitate a significant amount of manual labor. 27

Metadata and Database Management

In the early stages of our project, we considered whether or not to store the corpus metadata and the texts themselves in a database (as many DH projects are wont to do), and we made the wrong decision about databases (as many DH projects are wont to do). Initially, our thinking had been that with only about 400 texts, the total amount of data that we would need to track was relatively manageable, and its highly regular, tabular nature meant that a spreadsheet would be an adequate way to store this information.[19] Our approach was first vexed by book-historical research questions that led us to collect more thorough metadata of more kinds than we originally anticipated. We were interested in equipping ourselves to answer questions — or at least raise them — about, for example, the role of particular publishing houses in the development of literary theory, and we wanted to be able to attend closely to questions of translation, necessitating tracking translation dates as well as dates of first publication (in the original language as well as English), and so on. As columns and spreadsheets proliferated, recording consistent and complete metadata became onerous. 28

These choices resulted in endless manual reconciliation of our data against our library's records, often complicated by the fundamental differences in data models preferred by literary scholars and libraries. To us as critics, the basic unit of data is a *literary* (or critical) *work*, for example Franz Fanon's *The Wretched of the Earth*, 29 with associated metadata concerning, e.g., its first publication, its eventual translation into English, which version of the text is included in our corpus, the presence of an introduction by Jean Paul Sartre, and so on.[20] Our library, by contrast, tracks five individual versions of this text in English alone, in addition to two versions in French, one in Persian, one in Pashto, and a French edition of Fanon's *Œuvres.* For the library's purposes, texts come in *editions*, and while these texts are usefully cross-referenced by the inclusion in their catalog data of the original French title, they are fundamentally different books that exist in different numbers of copies — for example, two copies of the 1963-65 Grove Press edition translated by Constance Farrington, and one copy of Grove's 60th anniversary edition published in 2021 ("translated from the French by Richard Philcox; with commentary by Jean-Paul Sartre, Homi K. Bhabha, and Cornel West"; [Fanon 2021]. While in principle our work-oriented model could link each item to a particular instantiation available via the library (or an instantiation of that text purchased for this project), this proved exceptionally difficult to do, in large part because of the need to track library-collection data at the level of the work, not the edition. For example, we found ourselves in situations where we wanted, for one single work, to track that one edition had been sent to Google for potential inclusion in Google Books; to note that another ebook version was available only via subscription; and to record a reference to a new, purchased ebook of the same work. We found, moreover, that as we navigated the library ecosystem, we were continually returning to the same items' catalog entries to record data that we had not initially collected. A database that linked our 400-odd works directly to the (potentially multiple) MARC records associated with all the editions of each of the texts would have, ultimately, proven simpler.

Discussion

Analysis of sources

After sixteen months of work, we obtained, or, in many cases, created plain text files for 383 of 402 texts, as well as DRM-protected ebooks corresponding to the remaining 19 texts (which we are waiting to process so as not to create high risk data before we need to use it). Of these 402 texts, 206 — somewhat more than half — were born-digital files derived from ebooks, nearly exclusively "circulating" texts owned or perpetually licensed to Stanford University Libraries (see Figure 4). A small number of additional ebooks were purchased on non-institutional licenses. While born-digital sources, taken collectively, provide cleaner texts than digitized ones, they are encumbered by awkward workflows to remove DRM and extract text. Moreover, the need to robustly track license information for each text requires access to data that may not always be available to library users, and which is in any event likely unfamiliar to most scholars working outside of libraries. We found, then, that although ebooks presented a certain measure of (physical) labor savings and offer the highest text quality, they were correspondingly more complicated to work with than digitized sources, to the extent that, for a project of this size, digitization would have been the simpler and ultimately faster way to obtain our texts. Because many of these logistical costs are relatively fixed, larger projects may find greater benefit from working with library ebooks, and projects that already have the infrastructure to track licensing details, secure their data, and similar, may likewise find this balance tilted in favor of ebook sources. 30

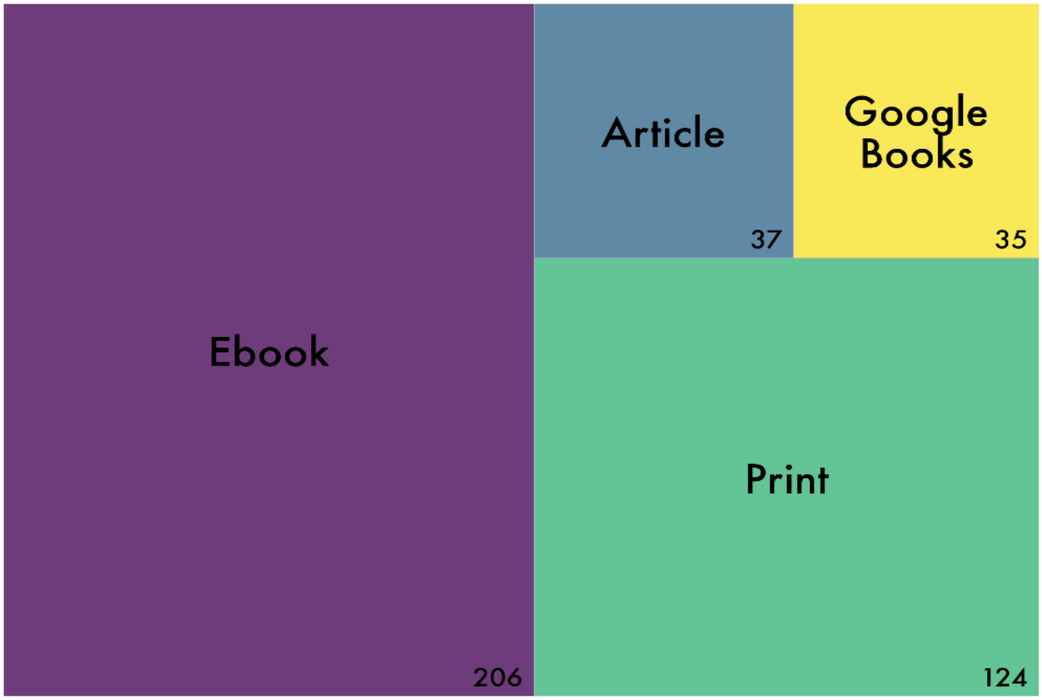

Figure 4. Overall corpus source types

One particular aspect of the library-scholar interface that we feel it is important to draw attention to is the complex interrelation of public facing library catalogues, internal library records, and scholarly metadata. For projects drawing upon library resources, and, *a fortiori*, when those projects are more bibliographically rigorous than ours, we think that aligning these three systems of bibliographic information will be invaluable. For scholars, this means acquiescing to the printed reality that texts come in multiple versions, which can be handled either by a many-to-one database relationship, or by simply committing to one particular version of a text, while for libraries, this means exposing data that, while not of general interest to most users, is necessary for scholars who might wish to cross-reference between, for example, call numbers, barcodes and book titles. Making this information available to scholars exposes them to a mode of thinking about books that is importantly different from the perspective adopted by most researchers. This allows researchers to expedite and clarify their communication with librarians, for example by providing barcode numbers as disambiguation in library requests, rather than more cumbersome edition information. 31

The single most important uses of MARC record information for this project has been to verify ebook ownership: like most academic library catalogues, Stanford's online system makes ebooks available to end-users in a way that does not expose the ebook's source, whereas the MARC records for each digital item include field 32

856, which contains “Electronic Location and Access” information, which our library uses to track whether a particular ebook was “subscribed” or “purchased”. To facilitate the use of library purchases for digital research and text mining, we suggest that information such as this be easily accessed by researchers, or even searchable as a filter option through the online catalogue. While the direct use of this information by researchers may be relatively esoteric, we think that a wider number of users concerned with access to (and the affordability of) library resources should be presented with information on which holdings their library owns, and which are merely licensed.

In particular, the impediments posed by commercial, individual (i.e., non-institutional) ebook licenses make the most straightforward approach to ebook acquisition significantly more burdensome. As the DMCA exemption provides no relief from restrictive license agreements, many texts are simply unavailable commercially under terms that permit the extraction of text for text-mining purposes. Even when texts are available without such restrictions, the texts in question often need to be sourced from smaller publishers and distributors (rather than large online resellers), increasing the logistical complexity of identifying, ordering, and tracking such works. 33

Texts scanned and digitized from originally printed sources (as opposed to born digital texts) offer the considerable advantage of simplicity, and with the increasing power and speed of modern scanning technology and OCR, provide the possibility of relatively clean texts, albeit less so when texts include tables or visual media.[21] Digitization, moreover, allows for the creation of output in a wide variety of formats, including markup-based OCR formats that preserve information about page layout and similar textual features, which can be difficult to extract from ebooks. The principal disadvantage of digitization is time: with overhead scanners, digitization is slow, especially if image quality is a priority, and mechanized solutions are imperfect and tremendously expensive. Feed-scanning is faster and produces acceptable results relatively quickly, but it precludes non-destructive scanning of books, necessitating higher costs for the end-users.[22] And, while disassembled texts can be stored for future re-scanning, most digital humanities groups are probably not equipped to store tens or hundreds of thousands of pages of unbound literature in anticipation of future developments in scanning technology. While the digital storage of scanned material is less of an issue than the physical, it is nevertheless worth calling attention to, since high-resolution images of thousands of pages of text lead rapidly to voluminous storage requirements; and though such images can be discarded, this essentially “freezes” the OCR output and prevents future reprocessing. 34

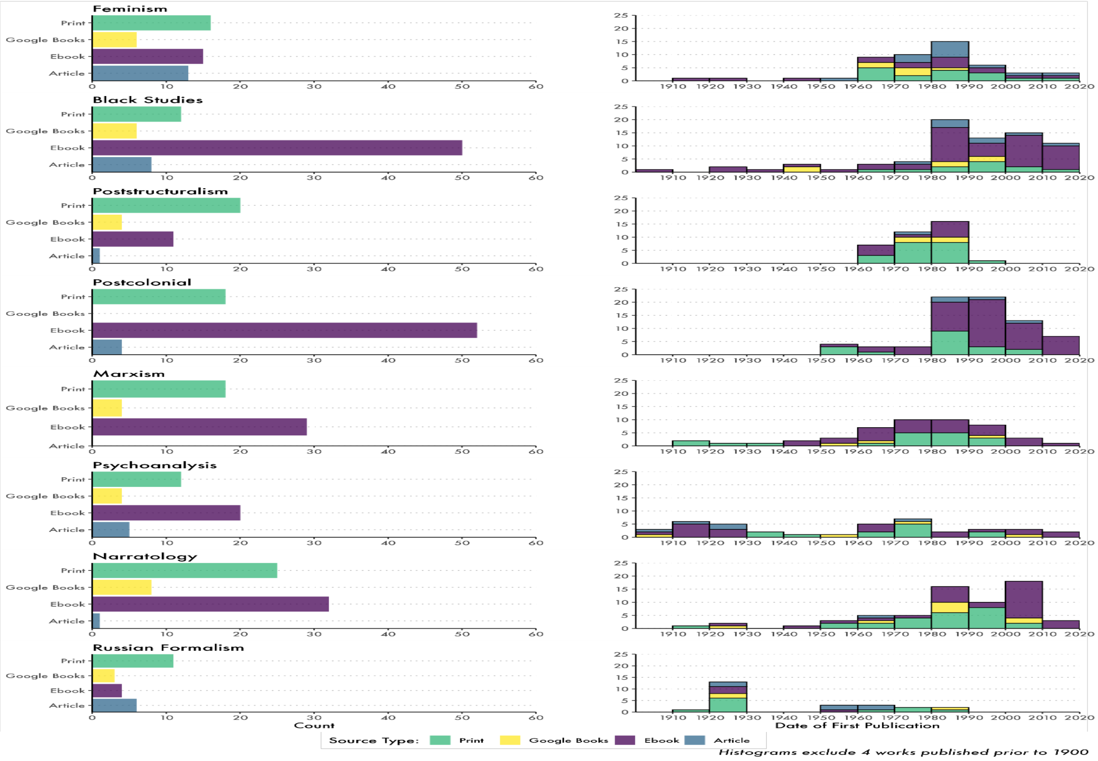

Figure 5. Distribution (by age and source type) of each theoretical field

The works included in our corpus span the range of the 20th and 21st centuries, with a small tail into the 19th (three works by Karl Marx and one by Sigmund Freud), and we had expected that the age of the texts in question would have some correlation with the medium in which it was possible to locate those texts. In practice, this effect was relatively minor: the median age of the works in each of our source type categories ranges from 1978 (works recovered from Google Books) to 1989 (ebooks) (see Figure 6). Of potentially greater interest, nearly all — 71 of 85 non-article works first published after 2000, including 29 of 32 from after 2010 — of the most recent works were available in born-digital format (see Figure 7). 35

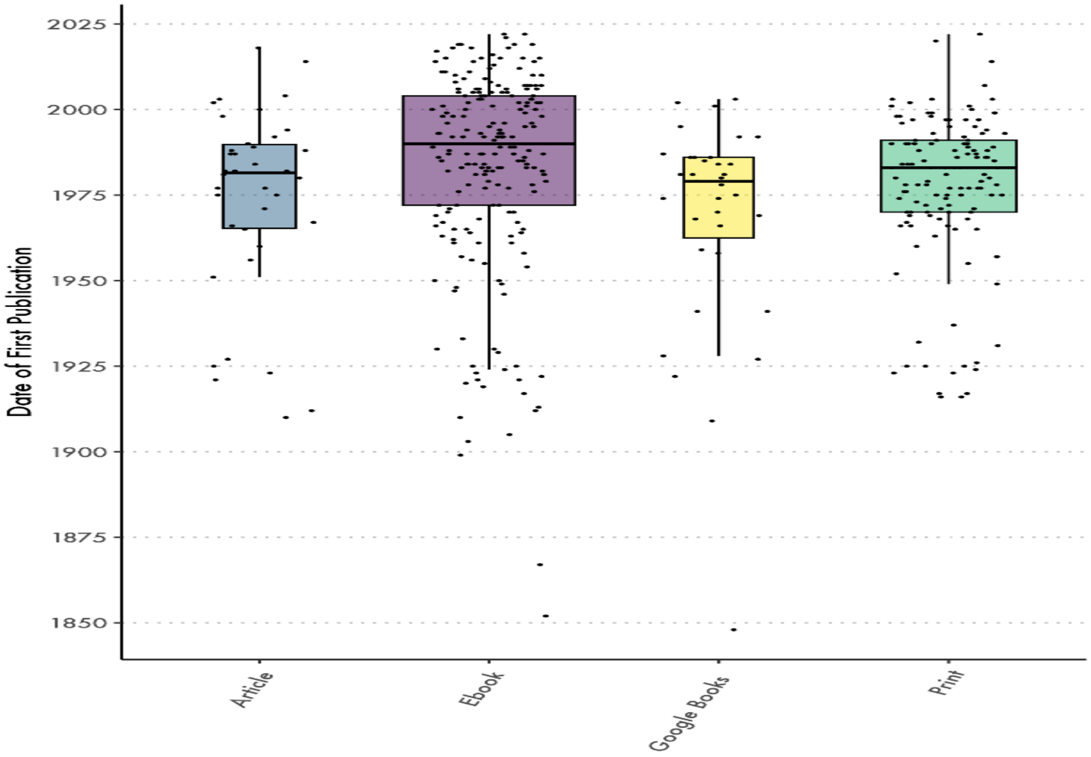

Figure 6. Distribution of texts' age by each source type category. Medians indicated by horizontal rule.

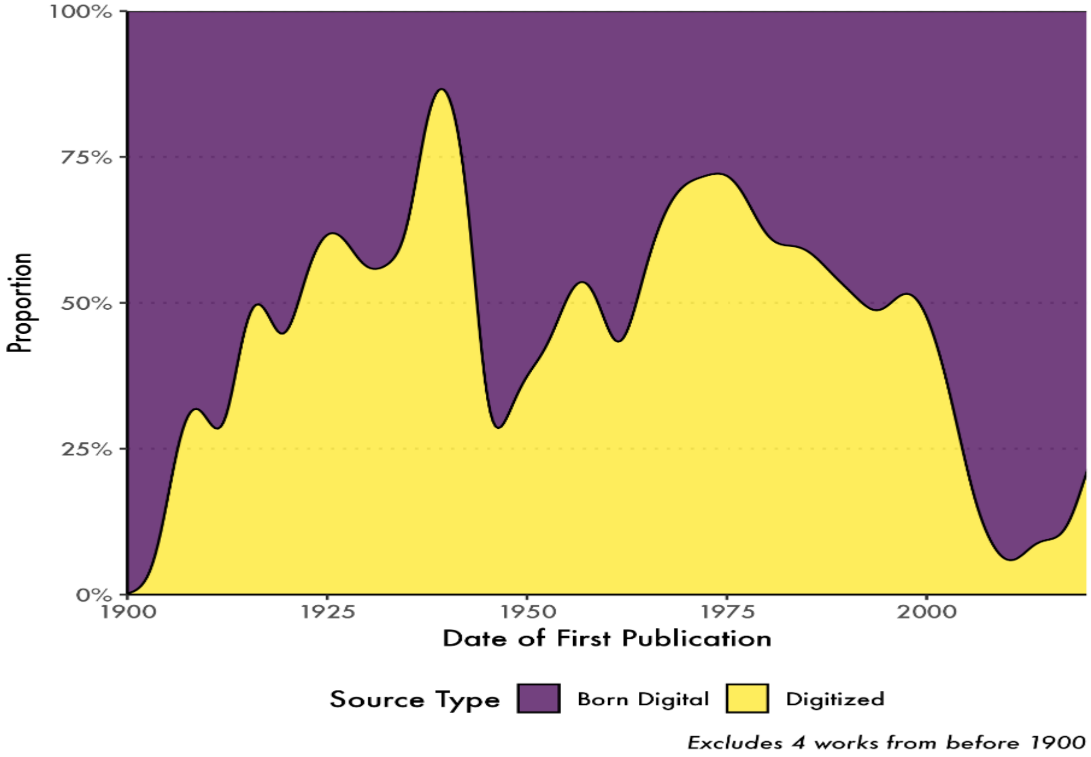

Figure 7. Proportion of corpus texts available in born-digital format.

Since a significant number of older texts are also available as ebooks, we hypothesize that ebook availability, while obviously correlated with recency, has much to do with the specifics of publication history, scholarly use, and other similar factors that we did not track (recent new editions, frequent course assignments, etc.). We hypothesize, for example, that for theoretical fields whose popularity peaked prior to 1990, such as Poststructuralism, libraries would own multiple print copies of books related to this field and thus be disinclined to repurchase ebooks. Theoretical fields with more recent popular engagement, such as Black studies and Postcolonial studies, had a higher proportion of ebooks even for texts written prior to 1990, as demand within libraries resulted in ebook purchases (see Figure 5). Since one of the main limitations of working with the DMCA exemption for this project has been the limited number of accessible ebooks, further research on the forces shaping ebook availability would be a great asset to digital humanists, since it would better equip researchers to understand whether the exemption might be valuable to them in assembling corpora. Since ebook availability varies from country to country, internationally collaborative work in this area would be especially valuable, since it would help to illuminate different patterns in ebook availability across different national and linguistic contexts. Similarly, we suspect that the trends in academic publishing that we observe may be quite different from those that obtain in the case of primary sources such as novels and other texts with significant commercial markets.

Articles and Monographs

Previous DH work on academic publication, especially that modelled after scientometric work conducted in the sciences and social sciences, has tended to focus on the scholarly article, at this point nearly the sole academic currency in most scientific fields [Goldstone and Underwood 2012], [Riddell 2014], [Feeney 2017], [Ambrosino et al. 2018], [Piper 2020]. Work in this area, moreover, has been greatly streamlined by requirements for open-access publication, as well as government-funded databases such as PubMed, which make the text of academic articles relatively accessible to scholarly research [Gusenbauer and Haddaway 2020], [Falagas et al. 2008]. In the humanities, by contrast, the state of open-access publication is significantly less advanced, and the monograph, typically produced as an expensive printed book by an academic publisher, remains an important element in the scholarly economy, albeit one increasingly supplemented by simultaneous publication as an expensive academic ebook. Monographs are well-established as central to humanistic citation [Thompson 2002] and tenure [Estabrook and Warner 2003]. In neither electronic nor codex form, however, are academic books particularly amenable to computational exploration, which we posit has been the reason for their exclusion even from studies of scholarship in the humanities, where their exclusion significantly reshapes the scholarly field. For this reason, the present project was deliberately designed to center the academic book and further work on monographs across a wide variety of humanistic fields is clearly needed. 37

Conclusion

Embarking on this project, we had hoped that, over the past decade, the increase in ebook publications and their subsequent heightened representation in library collections, together with the movement towards open access scholarship and, recently, the changes in the US legal regime surrounding in-copyright electronic media would make the task of assembling a corpus of bespoke texts easier or, at the very least, attainable. What we have found instead is that humanities scholars, facing the time and funding constraints endemic to our field, still face significant challenges when trying to source a bibliography of in-copyright texts for the purposes of text and data mining. Given the complexities (and expense) involved in assembling a corpus of e-books and legally extracting their texts, it is still the case that it may be preferrable to assemble a print corpus to scan, OCR, and hand correct. While intensive in terms of labor, it may still save a significant amount of time. 38

From a bibliographic standpoint, we found the necessity of consulting, comparing, and tracking the availability of texts across multiple digital venues and platforms (open-access, library ebooks, digitized texts from already extant collections, etc.) to bring with it significant administrative overhead. Such overhead requires the expertise and access permissions of a librarian, which may not be available to all research teams and is beyond the reach of individual researchers not affiliated with a library. While the recent changes to the DMCA have improved the legal situation for researchers wishing to work with copyrighted texts, the restrictions imposed by the DMCA exemption have made working under its terms particularly challenging. Projects considering this route would be well-advised to consider whether sufficient numbers of institutionally-owned, DRM-protected ebooks exist to justify the additional labor required to work with this category of text. We are more optimistic about the use of DRM-free library texts, which poses fewer legal obstacles and which might fruitfully be a supplemental source of texts for many projects. The expense involved in acquiring library ebooks, however, compares relatively unfavorably to digitization, and, particularly as the quality of OCR continues to improve, we are skeptical that this will play a significant role in the future acquisition of digital corpora. 39

Despite the difficulties that we have encountered, there are two particular domains of academic publishing and book acquisition in which relatively minor changes could offer substantial improvements to the process of corpus building and which we offer as recommendations for future work. First, the movement towards open-access publication for academic texts could be transformative for scholars looking to do cultural analytics work on contemporary scholarly sources. Although researchers have urged scholarly publication venues to embrace open-access publishing for much of the past decade, even those that do take up the significant financial and publishing challenges of doing so still undertake this work through a non-standardized and even scattershot approach [Eve 2014]. The variety of formats that open-access texts take, from PDF files to proprietary data formats on publisher websites, means that consistent use of these resources as textual data remains much more difficult than it should be. We believe that publishers who are willing to consider open-access publication should be encouraged to adopt data standards for open-access formats and to make text files and other machine-readable file formats of the open access materials available to researchers.[23] 40

Second, the importance of the library in our research pipeline is difficult to overstate, in terms of both resources and relationships: in particular, the willingness of library staff to work with team members to acquire these texts, and their financial ability to do so, has made our project possible. In particular, the inclusion of a hybrid library staff member among our project team members indicates the importance of the library/researcher relationship in successfully assembling a corpus of this nature, and we are fortunate to work at an institution whose infrastructure and resources allow the privileging of this relationship. Nevertheless, barriers remain. Even for forward-thinking libraries that have full-time staff dedicated to digital licensing and who include language in their licenses about obtaining full-text files of ebooks they have purchased, there are still significant barriers to actually obtaining these files. Delays by publishers and vendors as they work through their legal teams (despite license language); liaising between researchers and library staff to identify the correct vendor and contact information; clarifying that what is needed are the full text resources rather than access to a proprietary TDM platform: all combine to add to the difficulty of this work. A concerted effort by university libraries, particularly those with the resources of Stanford, to normalize the availability of collection material for TDM research from the publishers could help to create a sustainable set of practices for researchers at other, often less well-resourced institutions. Further, given their work with publishers, libraries themselves might be able to prioritize building corpora out of their collection materials that they could make available to scholars at their institutions. Many libraries are already engaged in building digital repositories of data and, to the extent that this could be extended to curating data that is available for TDM research, libraries seem excellently positioned to play a significant role in developing and disseminating these resources. Finally, in the US, the importance of institutional text ownership to the DMCA exemption provides an impetus for libraries to focus on purchasing texts rather than renting them through third-party content distributors. 41

In a global context, the exemption to the DMCA is not likely to be directly relevant to scholars not actively engaged in collaboration with US-based researchers, and the particular legal structures that we have had to navigate will be different from those encountered by researchers working in other countries. A large number of the issues we have encountered in assembling our corpora have *not* been legal, however, and we imagine that many of the practical ramifications of navigating ebook licenses, tracking subscription data, and similar will apply to such projects, even if, for example, the onerous data security provisions of the DMCA exemption prove to be singular. More fundamentally, we think our work offers a strong case for the benefits of digitization’s simplicity. Of course, large scale digitization projects already require extensive infrastructure; and for projects larger than ours, the benefits of adapting to a regime such as the DMCA, and the potential for increased savings of time and money through the use of ebooks will require careful consideration of the possibility of working with ebooks, whether obtained via licensing agreements allowing for text and data mining, or DRM-circumvention exceptions designed to facilitate research. 42

In the fall of 2024, the TDM exemption to the DMCA was renewed, subject to minor changes intended to better facilitate collaboration between institutions. The most substantive of these changes allows for the sharing of in-copyright digital corpora collected in line with the exemption between researchers at different institutions, even if they do not both own a copy of the text. (Previously, sharing corpora between institutions was only possible between direct collaborators working together on the same research project.) While this will prove particularly valuable for researchers working at less resourced institutions (since it will enable the construction of mutually beneficial and partially open corpora) the added administrative overhead of such collaborations will only add to the bookkeeping burdens faced by researchers that we describe above. Moreover, the work of building and maintaining the networks required for creating such shared resources is likely to fall disproportionately to less privileged scholars, and may not appeal to scholars who can more easily rely on their own institution's resources. We suspect, too, that developing corpora that can be shared across international borders will result in further complications due to the need to comply with two (or more) potentially contradictory regulatory frameworks. 43

Overall, then, while the bibliographic and material conditions have improved such that it is now possible to obtain a bespoke, hand-assembled corpus of selected Anglophone literary works in digitized format, *possible* does not indicate *easy*. Without the support of a well-resourced and cooperative library, we do not believe this project could have been completed at all. As it currently stands, the DMCA exemption does not significantly expand scholarly access to fair use text mining, and its data sharing restrictions may reify existing funding barriers for those not working within research-intensive institutions. Moreover, without significant shifts in the availability and cost of ebooks, and legislation to directly address the licensing thereof, frameworks such as the DMCA exemption seem unlikely to prove useful for scholars working at small to medium scales. While it is tempting to look at developments like the DMCA exemption, the recent push by universities to create open-access archives of scholarship, and the ever-increasing size of repositories such as HathiTrust, and speculate that we are on the cusp of a revolution in sourcing an exponentially larger number of texts for our work, we are all too aware that such claims echo the utopian predictions of a decade ago (e.g., [Poole 2013]). As much as things have changed (including the appetite of large language models for ever more terabytes of text), our research still takes place against a system that is structurally hostile to even the least responsible creation of bibliographically-based corpora. 44

Acknowledgements 45

This project is supported by a Public Knowledge grant from the Andrew W. Mellon Foundation. We are grateful to Julia Gershon and Sarah Sophie Schwarzhappel, undergraduate research assistants to the Stanford Literary Lab, for their aid in the collection and verification of some metadata, and to Casey Patterson, who consulted on the selection of Black studies texts. We are also tremendously grateful for the support of the Stanford University Library system and the help of all of its staff, without which it simply would not have been possible to do this project. 46

Author Contributions: Matt Warner managed the project, led the writing of the paper and produced the visualizations; Nichole Nomura and Carmen Thong (equal contributions) contributed to the selection of texts and the collection of data, and to the writing of the paper; Alix Keener managed the corpus and worked with the library on the acquisition and digitization of texts; Maciej Kurzynski and Alexander Sherman (equal contributions with Gabi Birch) contributed to the selection of texts; Gabi Birch contributed to the writing of the paper; Mark Algee-Hewitt (supervising author) contributed to the writing of the paper, managed the data and legal risk and wrote the original grant application. All authors contributed to the theorization and intellectual direction of the project. 47

Notes

[1] Borgman provides numerous examples of this in the sciences and social sciences, but is only able to point to the example of Archeology within the Humanities [Borgman 2018].

[2] Attempts to address the problems of data publication for professional metrics (tenure, funding and similar) have seen the recent development of hybrid models of dataset-article, such as *Cultural Analytics*'s "Data Set" articles, Post45's "data essays" or the *Journal of Open Humanities Data*'s "Data Papers".

[3] Large bibliographical efforts at the national scale have created important new resources for several, particularly European, languages [BNF n.d.], [SPK n.d.]. Yet, despite these efforts, given the under resourced nature of non-English archives, the challenges we describe in this paper are just as pressing, if slightly altered, for other languages.

[4] Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies. 37 CFR Part 201. For a research-oriented summary, see that of [Authors Alliance 2021]. The bulk of the research described here was conducted during the fall of 2022 and through 2023, under the first version of the TDM exemption, which has since been renewed, subject to minor modifications, particularly relating to collaboration, which we discuss in the conclusion below.

[5] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.) Preamble, 7.

[6] The pivot towards so-called "A.I." technologies rests largely upon companies' willingness to train large language models with questionably sourced in-copyright works. Given the for-profit nature of technologies like OpenAI's GPT4 or Google's Bard, the fair-use doctrine that underlies permissions for extracting in-copyright text for research purposes does not generally apply; however, the existence and use of such corpora has already been adopted by opponents of all text-mining work, research and industry alike.

[7] In addition to increasing the availability of recently published texts, the push to obtain born-digital texts lies in the field's need to overcome the systematic corpus errors introduced by OCRed text, which is endemic within the field.

[8] Because we made our selections for each field independently, we have several works that are associated with multiple fields (e.g., Sadiya Hartman's *Scenes of Subjugation*, which we included in both feminism and Black studies).

[9] The recent and increasing reliance on this argument being made by corporations engaged in the training of large language models (LLMs), moreover, makes us uncertain as to its future, and uncomfortable with its present-day ramifications. A growing pile of lawsuits have been filed on behalf of authors against companies using copyrighted texts as training data for their LLMs [Saveri and Butterick 2023], [Reisner 2023]. The legality of using copyrighted texts, or even copyrighted art images, as training data is in the process of debate [Knibbs 2023], [Helms and Krieser 2023], [Sekhon et al. 2023], [Pike 2022], [Jung 2020]. Most academic work on this issue is still in its nascent stages, with promising new work across fields beginning to come up as preprints (such as [Chu et al. 2023]).

[10] We selected these presses as they were the only ones from whom we needed a substantial number of texts. We later learned that when the ebooks are distributed by a third party (such as Project Muse or ProQuest), these requests are handled by the distributor, not the publisher, which would have made a broader request more practical (though it is not clear if every distributor still checks each request with every publisher).

[11] In addition to whether Stanford specifically could access the digitized text because of its place in Google's scanning order, the coverage of our recovered corpus is also fundamentally influenced by Google's digitization pipeline and material workflow. Google's process excluded fragile, large, small, tightly-bound, brittle, and uncatalogued books from its scanning queue, taking an "opportunistic rather than systematic approach to digitization [that] may amplify existing selection biases in physical print collections" [Chalmers and Edwards 2017, 13]. For the purposes of our project, Google's priorities — far removed from those of DH and literary researchers and of the libraries Google pulled books from — resulted in a corpus that is not only biased towards the canonical and available, but also excludes some of our highest priority texts.

[12] Note that the MacOS and Windows versions of FineReader are not equivalent; our testing showed the full-featured Windows version to noticeably outperform the Mac version.

[13] We are fortunate that we had access to these databases and recognize that this access is subject to our individual institution's financial decisions. Scholarly database access is notoriously costly to access at the level of the institution, as well as for independent researchers [Björk 2021], [Bosch et al. 2019].

[14] Note that OA is not the same as open-source (OS) — which, while predominantly used as a framework for understanding software, has been applied to other domains (see, for example, the essays in [Hawkins 2021]).

[15] Specifically: The term "effective security measures" is defined as: "(1) Security measures that have been agreed to by all interested copyright owners of literary works and institutions of higher education; or (2) Security measures that the institution uses to keep its own highly confidential information secure." (b)(5)(ii)(B)

[16] We use the term "high-risk" below in reference not just to our own institution's data risk classification scheme, but also, more broadly, to gesture to data deemed particularly sensitive by other analogous institutional schemes for classifying and handling data that poses privacy, financial, ethical, legal or other risks.

[17] As our university also maintains a separate HPC environment for non-high-risk data which allows non-paying users to access the computational resources of paid users when they would otherwise be idle, switching to the per-hourly secure HPC environment will also entail additional financial costs. This cost-differential reflects in significant part the primary use case for high-risk compliant computational resources, biomedical research funded on an entirely different scale from all other forms of academic research: in the US, for example, the total federal appropriations for the National Institute of Health is $49B compared to $9.45B for the National Science Foundation, and $207M for the National Endowment for the Humanities [NIH 2023], [NSF 2023], [NEH 2023].

[18] This is true for both deliberate DRM/TPM [Fisher 2020] and the lack of interoperability found in older, proprietary software [McDonough et al. 2010].

[19] For convenience and to combat the proliferation of columns, we ultimately split this spreadsheet into two unique-identifier linked sheets, one containing data associated primarily with the acquisition of the texts (library records, price information, availability for purchase, etc.) and one with more traditionally bibliographic metadata.

[20] Of course, we are not ignorant of the multiplicity of such texts and the differences between editions — but we are accustomed to viewing these items as, fundamentally, editions *of a text*, a unifying structure that in most cases connects them. Even so, earlier projects from the Literary Lab have confronted the process of selecting between multiple editions of, for example, *Robinson Crusoe*, whose text changes drastically, on a sentence-by-sentence level, between editions from the early and the late eighteenth-century.

[21] OCR has improved dramatically over the last decade (notably, since our Google Books files were digitized), although benchmarking and measuring that improvement for individual corpora remains difficult [Hegghammer 2021], [Neudecker et al. 2021]. As the quality of OCR has known downstream effects in text-mining that vary by method (e.g. topic modeling or natural language processing), the benefits of born-digital texts over digitized ones will vary by project [Hegghammer 2021].

[22] Google Books scanned their books using manual labor at a cost of US$10 each, a non-destructive approach dependent on “an army of human page-turners,” a sizable amount of computational power, and an agreement with libraries that did not require them to purchase the books in question [Leetaru 2008].

[23] Given that the concerns around open access availability lie in the possibility of consumptive use on the part of researchers (that is, that they could read the open access version instead of buying the book), making the text available in a format amendable to TDM (for example a text file) should not increase this risk, as most readers do not want to consume text in its raw form.

Works Cited

Ambrosino et al. 2018 Ambrosino, A. et al. (2018) “What topic modeling could reveal about the evolution of economics”, *Journal of Economic Methodology*, 24(4), pp. 329–348.

Authors Alliance 2021 Authors Alliance. (2021) “Update: Librarian of Congress Grants 1201 Exemption to Enable Text Data Mining Research”, *Authors Alliance*, 27 October. Available at: https://www.authorsalliance.org/2021/10/27/update-librarian-of-congress-grants-1201-exemption-to-enable-text-data-mining-research/ (Accessed: 19 October 2023).

BNF n.d. Bibliothèque nationale de France. (n.d.) “Gallica: The BNF Digital Library”. Available at: https://www.bnf.fr/en/gallica-bnf-digital-library (Accessed: 20 October 2023).

Bailey et al. 2017 Bailey, T.P., Scott, A.L. and Best, R.D. (2017) “Cost Differentials between E-Books and Print in Academic Libraries”. Available at: https://doi.org/10.5860/crl.76.1.6.

Björk 2021 Björk, B.-C. (2021) “Why Is Access to the Scholarly Journal Literature So Expensive?”, *portal: Libraries and the Academy*, 21(2), pp. 177–192. Available at: https://doi.org/10.1353/pla.2021.0010.

Bode 2018 Bode, K. (2018) *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press. Available at: https://doi.org/10.3998/mpub.8784777.

Bode 2020 Bode, K. (2020) “Why You Can't Model Away Bias”, *Modern Language Quarterly*, 81(1), pp. 95–124.

Borgman 2018 Borgman, C.L. (2018) “Open Data, Grey Data, and Stewardship: Universities at the Privacy Frontier”, *Berkeley Technology Law Journal*, 33(2), pp. 365–412. Available at: https://doi.org/10.15779/Z38B56D489.

Bosch et al. 2019 Bosch, S., Albee, B. and Romaine, S. (2019) *Deal or No Deal | Periodicals Price Survey 2019*, *Library Journal*. Available at: https://www.libraryjournal.com/story/Deal-or-No-Deal-Periodicals-Price-Survey-2019 (Accessed: 19 October 2023).

Chalmers and Edwards 2017 Chalmers, M.K. and Edwards, P.N. (2017) “Producing "one vast index": Google Book Search as an algorithmic system”, *Big Data & Society*, 4(2), p. 205395171771695. Available at: https://doi.org/10.1177/2053951717716950.

Chu et al. 2023 Chu, T., Song, Z. and Yang, C. (2023) “How to Protect Copyright Data in Optimization of Large Language Models?” Available at: https://doi.org/10.48550/ARXIV.2308.12247.

Estabrook and Warner 2003 Estabrook, L. and Warner, B. (2003) *The Book as the Gold Standard for Tenure and Promotion in the Humanistic Disciplines*. Urbana-Champaign: University of Illinois.

Eve 2014 Eve, M.P. (2014) *Open Access and the Humanities: Contexts, Controversies and the Future*. Cambridge: Cambridge UP.

Falagas et al. 2008 Falagas, M.E. et al. (2008) “Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses”, *The FASEB Journal*, 22(2), pp. 338–342. Available at: https://doi.org/10.1096/fj.07-9492LSF.

Fanon 2021 Fanon, F. (2021) *The Wretched of the Earth*. 60th anniversary edition. Translated by R. Philcox. New York: Grove Press.

Feeney 2017 Feeney, M. (2017) “What Can Text Mining Reveal about the Use of Newspapers in Research?”, in *Collecting, Preserving, and Transforming the News - for Research and the Public. IFLA International News Media Conference*, Reykjavik, Iceland.

Fisher 2020 Fisher, K. (2020) “Copyright and Preservation of Born-digital Materials: Persistent Challenges and Selected Strategies”, *The American Archivist*, 83(2), pp. 238–267. Available at: https://doi.org/10.17723/0360-9081-83.2.238.

Gale Cengage 2014 Gale Cengage. (2014) “Eighteenth Century Collections Online”.

Gius and Jacke 2022 Gius, E. and Jacke, J. (2022) “Are Computational Literary Studies Structuralist?”, *Journal of Cultural Analytics*, 7(4). Available at: https://doi.org/10.22148/001c.46662.

Goldstone and Underwood 2012 Goldstone, A. and Underwood, T. (2012) “What can Topic Models of PMLA Teach Us About the History of Literary Scholarship?”, *Journal of Digital Humanities*, 2(1).

Gusenbauer and Haddaway 2020 Gusenbauer, M. and Haddaway, N.R. (2020) “Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources”, *Research Synthesis Methods*, 11(2), pp. 181–217. Available at: https://doi.org/10.1002/jrsm.1378.

Hawkins 2021 Hawkins, S. (ed.) (2021) *Access and control in digital humanities*. Abingdon, Oxon ; New York, NY: Routledge.

Hegghammer 2021 Hegghammer, T. (2021) “OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment”, *Journal of Computational Social Science*, 5(1), pp. 861–882. Available at: https://doi.org/10.1007/s42001-021-00149-1.

Helms and Krieser 2023 Helms, S. and Krieser, J. (2023) “Copyrights, Professional Perspective - Copyright Chaos: Legal Implications of Generative AI”, *Bloomberg Law*. Available at: https://www.bloomberglaw.com/external/document/XDDQ1PNK000000/copyrights-professional-perspective-copyright-chaos-legal-implic (Accessed: 15 October 2023).

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

Jung 2020 Jung, G. (2020) “Do Androids Dream of Copyright?: Examining AI Copyright Ownership”, *Berkeley Technology Law Journal*, 35(4), pp. 1151–1178.

Jänicke et al. 2015 Jänicke, S. et al. (2015) “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges”, in *STAR-State of The Art Report*. Eurographics Conference on Visualization (EuroVis), Ed. R. Borgo, F. Ganovelli, and I. Viola.

Knibbs 2023 Knibbs, K. (2023) “The Battle Over Books3 Could Change AI Forever”, *Wired*, 4 September. Available at: https://www.wired.com/story/battle-over-books3/ (Accessed: 15 October 2023).

Leetaru 2008 Leetaru, K. (2008). “Mass Book Digitization: The Deeper Story of Google Books and the Open Content Alliance.” *First Monday*, 13. Available at: https://doi.org/10.5210/fm.v13i10.2101.

Library of Congress n.d. Library of Congress (no date) “What is a MARC Record and Why is it Important?” Available at: https://www.loc.gov/marc/umb/um01to06.html (Accessed: 20 October 2023).

McDonough et al. 2010 McDonough, J. et al. (2010) “Preserving Virtual Worlds Final Report”.

NEH 2023 National Endowment for the Humanities. (n.d) “Appropriations of the National Endowment for the Humanities”. Available at: https://www.neh.gov/neh-appropriations-history.

NIH 2023 National Institute of Health. (n.d) “Appropriations of the NIH”, *Appropriations of the NIH*. Available at: https://www.nih.gov/about-nih/what-we-do/nih-almanac/appropriations-section-1.

NSF 2023 National Science Foundation (n.d) “Appropriations of the NSF”, *Appropriations of the NSF*. Available at: https://new.nsf.gov/about/budget/fy2023/appropriations#:~:text=The%20%22Consolidated%20Appropriations%20Act%20of,above%20the%20FY%202022%20appropriation.

Neudecker et al. 2021 Neudecker, C. et al. (2021) “A survey of OCR evaluation tools and metrics”, in *The 6th International Workshop on Historical Document Imaging and Processing*. HIP '21: The 6th International Workshop on Historical Document Imaging and Processing, Lausanne Switzerland: ACM, pp. 13–18. Available at: https://doi.org/10.1145/3476887.3476888.

Pike 2022 Pike, G. (2022) “Copyright and AI”, *Information Today*, 39(9), pp. 26–27.

Piper 2020 Piper, A. (2020) *Can We Be Wrong: The Problem of Textual Evidence in a Time of Data*. Cambridge: Cambridge University Press (Elements in Digital Literary Studies).

Poole 2013 Poole, A.H. (2013) “Now is the Future Now? The Urgency of Digital Curation in the Digital Humanities”, *Digital Humanities Quarterly*, 7(2).

Project Gutenberg n.d. “Project Gutenberg” (n.d.). Urbana, Illinois.

Ramsay 2011 Ramsay, S. (2011) *Reading machines: Toward an Algorithmic Criticism*. Urbana-Champaign: University of Illinois Press (Topics in the Digital Humanities).

Reisner 2023 Reisner, A. (2023) “Revealed: The Authors Whose Pirated Books Are Powering Generative AI”, *The Atlantic*, 19 August. Available at: https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/ (Accessed: 13 October 2023).

Riddell 2014 Riddell, A.B. (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: Boydell & Brewer, pp. 91–114.

SPK n.d. Stiftung Preußischer Kulturbesitz (n.d.) “Deutsche Digitale Bibliothek (German Digital Library)”. Available at: https://www.preussischer-kulturbesitz.de/en/about-us/tasks-of-national-interest/deutsche-digitale-bibliothek.html (Accessed: 20 October 2023).

Saveri and Butterick 2023 Saveri, J. and Butterick, M. (2023) “LLM litigation” · Joseph Saveri Law Firm & Matthew Butterick, *LLM Litigation*. Available at: https://llmlitigation.com/ (Accessed: 13 October 2023).

Sekhon et al. 2023 Sekhon, J., Ozcan, O. and Ozcan, S. (2023) “ChatGPT: what the law says about who owns the copyright of AI-generated content”, *The Conversation*. Available at: http://theconversation.com/chatgpt-what-the-law-says-about-who-owns-the-copyright-of-ai-generated-content-200597 (Accessed: 15 October 2023).

Thompson 2002 Thompson, J.W. (2002) “The Death of the Scholarly Monograph in the Humanities? Citation Patterns in Literary Scholarship”, *Libri*, 52(3). Available at: https://doi.org/10.1515/LIBR.2002.121.

Underwood 2019 Underwood, T. (2019) *Distant Horizons: Digital Evidence and Literary Change*. Chicago: Chicago UP.

Recommendations

DHQ is testing out three new article recommendation methods! Please explore the links below to find articles that are related in different ways to the one you just read. We are interested in how these methods work for readers—if you would like to share feedback with us, please complete our short evaluation survey. You can also visit our documentation for these recommendation methods to learn more.

SPECTER Recommendations

Below are article recommendations generated by the SPECTER model:

1. The Push and Pull of Digital Humanities: Topic Modeling the What is digital humanities? Genre, 2020, Elizabeth Callaway, University of Utah; Jeffrey Turner, University of Utah; Heather Stone, TETON Sports; Adam Halstrom, University of Utah
2. Recovering the London Stage Information Bank: Lessons from an Early Humanities Computing Project, 2017, Mattie Burkert, Utah State University
3. A Review of Bridget Whearty's Digital Codicology: Medieval Books and Modern Labor (2022), 2025, Loren Lee, University of Virginia
4. Covers and Corpus wanted! Some Digital Humanities Fragments, 2016, Claire Clivaz, Swiss Institute of Bioinformatics
5. Working on and with Categories for Text Analysis: Challenges and Findings from and for Digital Humanities Practices, 2023, Dominik Gerstorfer, Technical University of Darmstadt; Evelyn Gius, Technical University of Darmstadt; Janina Jacke, Kiel University

DHQ Keyword Recommendations

Below are article recommendations generated by DHQ Keywords:

1. Classics in the Million Book Library, 2009, Gregory Crane, Tufts University; Alison Babeu, Tufts University; David Bamman, Tufts University; Thomas Breuel, Technical University of Kaiserslautern; Lisa Cerrato, Tufts University; Daniel Deckers, Hamburg University; Anke Lüdeling, Humboldt-University, Berlin; David Mimno, University of Massachusetts, Amherst; Rashmi Singhal, Tufts University; David A. Smith, University of Massachusetts, Amherst; Amir Zeldes, Humboldt-University, Berlin
2. Conclusion: Cyberinfrastructure, the Scaife Digital Library and Classics in a Digital age, 2009, Christopher Blackwell, Furman University; Gregory Crane, Tufts

University

3. Cyberinfrastructure for Classical Philology, 2009, Gregory Crane, Tufts University; Brent Seales, University of Kentucky; Melissa Terras, University College London
4. Digital Criticism: Editorial Standards for the Homer Multitext, 2009, Casey Dué, University of Houston, Texas; Mary Ebbott, College of the Holy Cross
5. Using word vector models to trace conceptual change over time and space in historical newspapers, 1840–1914, 2022, Jaap Verheul, Utrecht University; Hannu Salmi, University of Turku; Martin Riedl, University of Stuttgart; Asko Nivala, University of Turku; Lorella Viola, University of Luxembourg; Jana Keck, German Historical Institute Washington; Emily Bell, University of Leeds

TF-IDF Recommendations

Below are article recommendations generated by the TF-IDF Model:

1. The Ebook Imagination, 2022, Simon Peter Rowberry, Department of Information Studies, University College London
2. Pertinent Discussions Toward Modeling the Social Edition: Annotated Bibliographies, 2012, Ray Siemens, University of Victoria; Meagan Timney, University of Victoria; Cara Leitch, University of Victoria; Corina Koolen, University of Victoria; Alex Garnett, University of Victoria
3. Ooligan Press: Building and Sustaining a Feminist Digital Humanities Lab at a R-2, 2020, Kathi Inman Berens, Portland State University; Abbey Gaterud, Chemeketa Community College; Rachel Noorda, Portland State University
4. Lessons from the Library: Extreme Minimalist Scaling at Pirate Ebook Platforms, 2022, Martin Paul Eve, Birkbeck College, University of London
5. The New Place of Reading: Locative Media and the Future of Narrative, 2011, Brian Greenspan, Carleton University