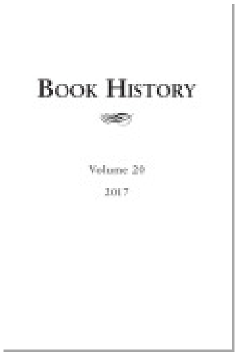
"Q i-jtb the Raven": Taking Dirty OCR Seriously

Ryan Cordell



Book History

Volume 20

2017

Book History, Volume 20, 2017, pp. 188-225 (Article)

Published by Johns Hopkins University Press
DOI: https://doi.org/10.1353/bh.2017.0006

“Q i-jtb the Raven”

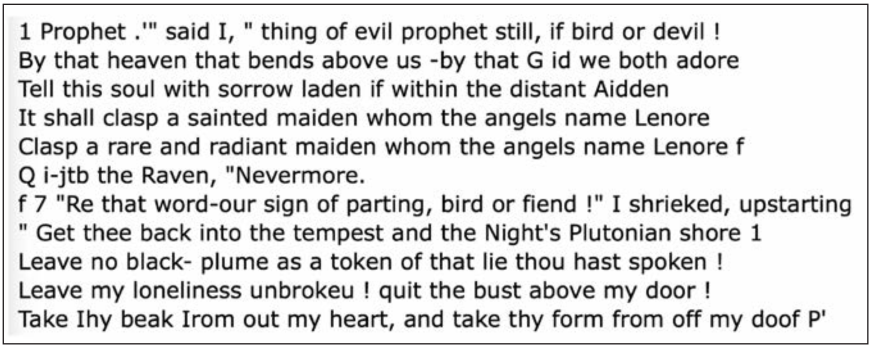
Taking Dirty OCR Seriously

Ryan Cordell

I. Introduction

On November 28, 1849, the *Lewisburg Chronicle, and the West Branch Farmer* published one of the most popular poems of the nineteenth century, Edgar Allan Poe’s “The Raven.” This reprinting falls somewhere in the middle of an enumerative bibliography of Poe’s poem, which was widely printed, reprinted, and parodied in period newspapers. The *Lewisburg Chronicle*’s “Raven” is one version among many produced after Poe’s death in 1849—“By Edgar A. Poe, dec’d”—interesting as a small signal of the poem’s circulation and reception. It is just such reprinting that we are tracing in the Viral Texts project, in which we use computational methods to automatically surface patterns of reprinting across nineteenth-century newspaper archives.[1] While it might seem an obvious point to make, methods such as text mining rely not on nineteenth-century newspapers themselves, but on those newspapers as remediated by mass digitization, a phrase that shorthands elaborate systems of scholarship, preservation, bureaucracy, human labor, machine processes, and economics. Seen in this light, the *Lewisburg Chronicle* version of “The Raven” also becomes interesting as a digitized object in the twenty-first century, in which at least one iteration of the poem’s famous refrain has been rendered by optical character recognition (OCR) as, “Q i-jtb the Raven, ‘Nevermore.’”[2]

Though conversations about large-scale digital archives typically revolve around page images, it is OCR-derived text that underlies our searches and more complex forms of digital text analysis, and it is OCR that has received the least sustained attention from bibliographers and book historians. What then is *this* text? Where did it come from, and how did it come to be the way it is? How should we understand the relationship between the OCR underlying a digital archive and images through which we typically experience those archives? How then should we understand the relationships among those digital components and the analog technologies of microfilm



1 Prophet .'" said I, " thing of evil prophet still, if bird or devil !
By that heaven that bends above us -by that G id we both adore
Tell this soul with sorrow laden if within the distant Aidden
It shall clasp a sainted maiden whom the angels name Lenore
Clasp a rare and radiant maiden whom the angels name Lenore f
Q i-jtb the Raven, "Nevermore.
f 7 "Re that word-our sign of parting, bird or fiend !" I shrieked, upstarting
" Get thee back into the tempest and the Night's Plutonian shore 1
Leave no black- plume as a token of that lie thou hast spoken !
Leave my loneliness unbrokeu ! quit the bust above my door !
Take Ihy beak Irom out my heart, and take thy form from off my doof P'

Figure 1. A sample of the OCR-derived text for "The Raven" in the CA Lewisburg Chronicle, including the line that gives this essay its title.

and print from which they are derived? Why is the *Lewisburg Chronicle* from Lewisburg, Pennsylvania, available through the Library of Congress' open-access Chronicling America archive, while Philadelphia's *Pennsylvania Freeman*, which would print "The Raven" five months later, is available through Readex's commercial archive, America's Historical Newspapers? How were these decisions made and by whom?

Our primary perspective on the digitized text thus far has been that of the textual critic entirely "concerned with . . . the reconstruction of the author's original text." As W.W. Greg contended in 1932, however, "criticism may just as rightly be applied to any other point in the transmission of the text." For Greg, the bibliographer's concern must be "whole history of the text" in which "the author's original is but one step"—albeit likely an important step—"in the transmission." Greg describes "the text" not as a single individual, but instead as a lineage: "We have in fact to recognize that a text is . . . a living organism which in its descent through the ages, while it departs more and more from the form impressed upon it by its original author, exerts, through its imperfections as much as through its perfections, its own influence upon its surroundings."[3] In the eight decades since its publication, Greg's notion of the living text has significantly influenced work in bibliography, book history, and critical editing. Theories of the variorum text, the fluid text, and the social text have refined a vocabulary for discussing transmission, circulation, and difference as essential features of any literary work. Scholars have experimented with ways to represent fungible texts both in print, such as John Bryant's "fluid text" edition of *Moby Dick* (Longman 2009), and using digital tools, such as NINES' Juxta Commons collation platform.[4]

While scholars revel in revealing the fluidity of texts from the hand- and machine-press eras, however, we rarely note—except, perhaps, in dismissal—the variora emerging online. Just as cheap, pirated, and errorful American editions of nineteenth-century British novels teach scholars much about economics, print technology, and literary culture in that period, dirty OCR illuminates the priorities, infrastructure, and economics of the academy in the late 20th and early 21st centuries. Literary scholars know to distinguish when they build an argument about "The Raven" from its 1845 printing in *Graham's Magazine* or from a twenty-first-century critical edition of Poe's poetry; we understand that both can be appropriate sources, depending on the nature of our claims and the evidence demanded by those claims. We do not require all arguments be constructed from first sources, which would unduly strain much work, but we do require that scholars appropriately account for the sources they use. Similarly, we must reckon with mass digitized historical texts as new and discrete bibliographic objects, which is to say as objects worthy of and available for source criticism.

In the following pages I sketch a critical bibliography of a single digitized newspaper issue, the above-mentioned *Lewisburg Chronicle, and the West Branch Farmer* of November 28, 1849.[5] I frame both the material and social histories that led to its digitization and online publication, drawing from the metadata presented on the Chronicling America site; the metadata embedded in the digitized objects available for download there; and my own research into the funding programs, grant proposals, organizational structures, and project workflows of the National Newspaper Digitization Project and its subsidiary, the Pennsylvania Digital Newspaper Project. I argue that we must understand mass digitized texts as assemblages of new editions, subsidiary editions, and impressions of their historical sources, and that these various parts require sustained bibliographic analysis and description. Such media-specific theorization helps disentangle discussions of large scale digitization from myths of surrogacy—or worse, replacement—which hinder both their creation and their use.

II. Digital Bibliography

In their recent overview of digital scholarship for *Book History*, Matthew Kirschenbaum and Sarah Werner call for the "perspectives of book historians . . . on the large-scale digitization efforts underway at such places such as Google Books, the Internet Archive, HathiTrust, Gallica, and other institutions that are actively aiming to make print resources available as digital objects."[6] Book historical perspectives are particularly wanted, they argue, given the challenges of uneven metadata quality, single-copy digitization, and flat digital representation of the materials in these archives. Such issues become acute as researchers increasingly move beyond search to address digitized archives through computational text and image analyses, as do many of the projects described by Kirschenbaum and Werner. To complement such analyses we require more robust methods for describing digital artifacts bibliographically: accounting for the sources, technologies, and social realities of their creation in ways that make their affordances and limitations more readily visible and available for critique.

As Alan Galey argues of e-books in "The Enkindling Reciter: E-books in the Bibliographical Imagination," digitized newspapers are also "human artifacts, and bear the traces of their making no less for being digital," though they "bear those traces in ways bibliographers have yet to explain thoroughly."[7] Galey's article is a masterful example from a growing body of literature around the bibliographic description of "born digital" materials: artifacts which originated as files on a computer. Galey's analysis of *The Sentimentalists* deftly negotiates between the distinct *objects* (file formats through which an e-book is distributed), *states* (examinations of a given e-book file through different software platforms), and *instances* (renderings of a given e-book file on different hardware) of a single electronic "book." The e-book, Galey shows, is "a complex transaction of electronic signals, material inscriptions, and cultural codes" which we interpret through "layers and layers of digital tools and interfaces."[8] These layers extend beyond the hardware and software, including too the "cultural, technical, and even political forces" that drive the production and use of e-books.[9]

For born-digital texts, the challenge for bibliographers is apprehending the necessary layers for description in a medium still rapidly evolving. In *Mechanisms* and related work, Matthew Kirschenbaum describes a "forensics" that can provide an "account of electronic texts as artifacts—mechanisms—subject to material and historical forms of understanding."[10] Through a doubled focus on "the twin textual and technological bases of inscription (storage) and transmission (or multiplication)" Kirschenbaum offers careful bibliographic accounts of seemingly ineffable electronic texts, such as William Gibson's electronic poem *Agrippa*, which was initially distributed in 1992 on 3.5" floppy disks that encrypted after a single use, but which was almost immediately replicated through video tape recording and transcription and distributed via electronic bulletin boards. Kirschenbaum and Werner trace the bibliographic impulse across fields such as software studies, critical code studies, and platform studies, the latter of which can be exemplified by Nathan Altice's *I Am Error*, a rigorous account of the intertwined material and social meanings of the Nintendo gaming platforms of the early 1980s.[11] Likewise we might look to the "Preserving Virtual Worlds" team's attempts to apply a FRBR (functional requirements for bibliographic records) model to classic computer games such as *Mystery House*, *ADVENTURE*, and *Spacewar!* The PVW team notes, "even the simplest electronic 'text' is in fact a composite of many different symbolic layers" and the same is true of the digitized historical text.[12]

Digitized historical objects are curiously *less apprehensible* than born-digital objects because they invite familiar print contextualization and interpretations. As N. Katherine Hayles notes, "media constantly engage in a RECURSIVE dynamic of imitating each other, incorporating aspects of competing media into themselves" while our tools for theorizing texts remain "shot through with unrecognized assumptions specific to print."[13] When looking at a digitized issue of the *Lewisburg Chronicle*, in a very real sense we do not *see* the digital object, but instead a facsimile or worse, a surrogate, of the print object from which it is derived. Scholars often use evidence discovered in large-scale digital archives as if they were discovered through analog means—a newspaper page is cited, but not the archive from which it was drawn—eliding or at least downplaying the web interfaces, human labor, and algorithmic search technologies that shaped what is discovered (and what is not).[14] As John W. Maxwell contends, "hardware-centric thinking is so pervasive" that it blinds us to the most essential workings of computers for reading: "We still tend to think of computers as physical machines rather than the software that makes them go . . . we think of the iPad, but not the Safari browser that comes on every iPad. We think of e-readers, but not of the HTML and CSS software that makes them work."[15] In the case of digitized historical texts, we perhaps think of the screens through which we access them, but we do not necessarily think of the digitized materials themselves as software. The computer is treated as a window to the physical archive, rather than as an integrated system for remediation of the archive.

The first challenge for a serious bibliography of digitized materials, then, is one of apprehending: of *seeing* the digital object as such, as an artifact with a distinct materiality and sociology. This is the task Whitney Anne Trettien sets out in her study of *English Reprints Jhon Milton Areopagitica*, a print-on-demand book that, while an "artifact . . . of *print*" is, nonetheless, “a thoroughly *digital* object, produced from electronic information gathered by software searching enormous databases.” While Trettien notes that book historians often share “a general distrust of reprints . . . since reprints are (ostensibly) just that, reprints, unmediated by the intellectual labor of editing,” she insists print-on-demand reprints might offer “better perspective from which to understand our own historically-constructed assumptions about plain text and facsimile image, printed book and electronic file.”[16] Likewise, large scale, digitized historical archives offer an opportunity for scholars to thicken our understanding of the media they represent and our continually evolving relationships of reading and remediation toward the analog and digital archive.

When we treat the digitized object primarily as a surrogate for its analog original, we jettison the most compelling qualities of both media. The unique use of the digital medium, broadly considered, is the capacity to computationally trace patterns across corpora of various sizes, to “draw these materials into computable synthetic relations at macro as well as micro levels.”[17] Though corpus level data analysis might seem the province of a small subset of researchers, Ted Underwood reminds us that “[a]lgorithmic mining of large electronic databases has been quietly central to the humanities for two decades. We call this practice ‘search,’ but ‘search’ is a deceptively modest name for a complex technology that has come to play an evidentiary role in scholarship.”[18] In other words, the digital medium has already transformed humanistic research, though we rarely acknowledge or reflect on this reality. Both the predominant public interfaces of large-scale archives (focused on page images) and common modes of representing those materials in scholarship (a citation to the historical newspaper itself) encourage a fundamental misrecognition of the machine reading in which we are all engaged. To put it directly: our machine-read research in digitized archives is only occasionally predicated on their images. To adequately theorize *any* research conducted in large-scale text archives—including research that includes primary or secondary sources discovered through keyword search—we must avoid the myth of surrogacy proffered by page images and instead consider directly the text files they overlay.

III. OCR as Compositor

Discussions of algorithms in the humanities often focus on those that analyze already-digital texts. As Hayles outlines, the varieties of machine reading "range . . . from algorithms for word-frequency counts to more sophisticated programs that find and compare phrases, identify topic clusters, and are capable of learning," all methods at work, to one degree or another, in recent books such as Stephen Ramsay's *Reading Machines*, Matthew L. Jocker's *Macroanalysis*, or Franco Moretti's *Distant Reading*. Hayles argues, "Given the scope, pervasiveness, and sophistication of contemporary programs used to parse texts, it seems to me quite reasonable to say that machines can read."[19] However, a more pervasive, complex, yet largely uncommented variety of "reading machine" is OCR (optical character recognition), a type of software that bridges image and text analysis to mimic the identificatory functions of the human eye and brain. In the following section I ask whether it is also reasonable to say that machines can *compose*—in the bibliographic rather than authorial sense—and if so, how recognizing such composition might shift our accounts of digitized historical texts.

When scholars search in mass online archives, they do not search directly the words in the archives' source texts. Instead, they search an underlying text file, often hidden from the interface and likely encoded, though typically not at a fine level of detail, in a markup language such as XML (eXtensible Markup Language).[20] In most digital newspaper archives, these text files are created through OCR software, which attempts to recognize alphabetic characters on a page image and create a machine-readable text file from them. OCR data underlies most large-scale digital newspaper archives as well as larger book repositories, such as Google Books or the Internet Archive.

On one hand, the results of mass OCR processing of book, magazine, or newspaper pages are remarkable. Where human transcription would be prohibitively expensive and slow, through OCR words printed on thousands or millions of physical texts become, almost immediately to scholarly timelines, machine readable data that can be identified and computationally analyzed. However, OCR engines are also infamously unreliable, particularly for historical texts. Depending on the type, age, and conditions of a given set of historical documents, as well as on the procedures, hardware, and software of their digitization, OCR quality ranges widely. Most newspaper digitization efforts—and many for books—rely on scans of microfilm, adding the limitations of that earlier mass-preservation technology to the limitations of our current scanners. The OCR derived from such images is “errorful,” to borrow a term from computer science, carrying traces of remediations over decades of scholarly activity.

Errorful OCR influences our research in ways by now well expounded by scholars, inhibiting, for instance, comprehensive search. Were I to search “Quoth the Raven” in the Chronicling America database, its search engine would not find the line that gives me my title.[21] As Andrew Stauffer cautions,

> Algorithmic searching and text mining can guide us towards new patterns and connections that are only visible through the power of digital processing. Yet it must be remembered that this mode of research focuses almost exclusively on the verbal content of idealized models of nineteenth-century printed materials, models that are themselves vitiated by numerous localized errors in character recognition.[22]

Critiques that remind scholars about the uncomprehensiveness of search within digital archives are necessary, particularly while many researchers remain unaware of the underlying data structures upon which they rely. Moreover, such critiques have spurred notable attempts to correct errorful OCR in large scale digitization projects. Bonnie Mak notes that since 1999 the Text Creation Partnership (TCP) has mobilized “legions of outsourced ‘vendors’ who have keyboarded and tagged over 40,000 early English texts” in EBBO, as well as another 8,000 for Gale Cengage’s Eighteenth-Century Collections Online (ECC) and Readex’s Evans Early American Imprints (Evants-TCP).[23]

Such hand correction is less common for newspapers, largely because they are too voluminous. The Australian Newspaper Digitization Program, the results of which are available through the National Library of Australia’s Trove portal, invested significant money in “manually correcting the titles, subtitles, and first four lines of article text” over 21 million newspaper pages (as of July 2016), mostly through low-cost, offshore editing services.[24] In addition, Trove’s interface allows users to manually correct the archive’s text data, so that by July 2013, “more than 100 million lines of text” had been corrected through “crowd-sourced effort.”[25] Even so, the majority of Trove’s text data remains uncorrected OCR, as 100 million hand-corrected lines constitute only a small percentage of the total lines across 21 million newspaper pages.[26] Like Trove, some commercial newspaper archives hand correct titles and headlines, but most have deemed correction of articles too time consuming and costly. Thus the bulk of keyword search within large scale newspaper archives, public and commercial, depends on the output of OCR.

Given these realities, critiques that both begin and end with the imperfections of OCR foreshorten the bibliographic imagination. The mass digitized book, newspaper, or magazine is never simply a transparent surrogate for a corresponding physical object. It is instead a complex assemblage of new impressions and editions—in the full bibliographic senses of those words—which, while it "departs more and more from the form impressed upon it by its original author," nonetheless "exerts, through its imperfections as much as through its perfections, its own influence upon its surroundings." For books from the hand-press period, the bibliographic definition of edition is quite clear: "all the copies of a book printed at any time (or times) from substantially the same setting of type," including "all the various impressions, issues, and states which may have derived from that setting."[27] G. Thomas Tanselle expands slightly the conception of edition "in order to include modern methods of book production which do not involve actual type setting," arguing that "an edition should be defined as all copies resulting from a *single job of typographical composition* [my emphasis] . . . whether printed from type (set by hand or by machine), or plates, or by means of a photographic or electronic process." Tanselle argues, "all copies that derive from the same initial act of assembling the letterforms belong to the same edition."[28] In 1975 Tanselle is not yet thinking about digital processes, but his enlarging of the edition proves useful for naming the text files underlying large scale digital archives. While the *image* of a particular digitized text may reflect precisely the setting of type in the edition from which it was scanned, the *computer-readable text data* was reset in a new "job of typographical composition" by OCR software. Thus the OCR derived from any historical text constitutes a new edition of that text.[29]

If OCR produces a new edition of the text, we might think of OCR as a species of compositor: prone to transcription errors, certainly, but nonetheless resetting the type of its proof texts into .txt or .xml files rather than a galley. More specifically, we might think of OCR as a compositor setting text in a language it does not comprehend—as we know compositors sometimes did in the printing house—copying letters and words by their form rather than their sense. In making this argument, I want to pressure the distinction between OCR as an "automatic" process and composing type as a "human" process. To maintain such a dichotomy, we must ignore a long and fascinating interplay between technology and human labor in textual production. Both movable type and optical character recognition attempt to automate laborious aspects of textual production, and we can only speak of editions as such, whether printed or digital, within an industrialized framework.

Through the hand press period, Gaskell insists both the compositor and correctors "worked more or less automatically, and did not necessarily take in the general sense of what [they were] reading."[30] Gaskell points to Charles Manby's Smith's account of his life as a journeyman printer, and his story of "a reader employed for years together on an evening paper, every line of which he read and corrected professionally in the course of the day, who yet called for the same paper and read it regularly over his pipe and glass of grog in the evening, with the design of making himself acquainted with the news." As Smith notes earlier in the same section from which Gaskell quotes, readers working in print shops had "knowledge . . . sometimes sufficiently various" but "generally anything but profound," as they "in the course of fifteen or twenty years' practice . . . may have read detached and fragmentary portions of ten thousand volumes" but likely had "never read a dozen through from beginning to end."[31] In other words, while human compositors and correctors typically did recognize the words they composed, their processes could be more automatic than we might countenance.

A brief survey of nineteenth-century printers' manuals indicates that compositors at least occasionally composed in languages they did not themselves read, as when a volume mostly in English included quotations from continental European or classical texts. Printer's manuals routinely explained the alphabets, diacritical marks, and basic grammar of other languages, clearly expecting some compositors to require this help as they set type they could not read, or read fluently. Horace Hart's famous *Rules for Compositors and Readers at the University Press, Oxford* provides appendices with tips for setting French, German, Latin, and Greek words, noting in one case that "if the MS. is in well-written German script, and the compositor is acquainted with the German characters, he will find little difficulty in setting this up in German type," clearly implying through its "if . . . and" that some compositors would not be acquainted with German characters (thus requiring the help offered in the appendix) and that messy manuscripts would also force compositors to compose by forms rather than sense.[32]

Sylvere Monod demonstrates the latter conundrum at play even when compositors shared a language with authors, as when the compositors who hurriedly set Charles Dickens's *Bleak House* "were obviously unfamiliar even with the sometimes odd names and eccentric speech of several characters." Monod asks, "When faced with a manuscript which they were not always able to decipher, what could such men do?" answering, "They did just what ordinary human beings could be expected to do . . . they made guesses, with varying luck, they tended to substitute the expected for the unexpected, i.e., attempted to normalise Dickens's English; or, occasionally, they just gave up in despair, i.e., omitted the words they could not read at all or set up pure nonsense." While Dickens corrected over 700 compositor's errors at the proof stage, Monod shows that he missed at least 159, and that "the part played by the compositors in the evolution of the text is not negligible."[33] Though compositors generally understood the sense of the lines they composed, evidence from the printing house indicates that regular moments of more primitive kinds of text recognition punctuated compositors' working lives: moments which we might compare with the workings of optical character recognition.

As Rose Holley describes, OCR "attempts to replicate the combined functions of the human eye and brain" to identify alphabetic characters in a scanned image file, "which is why it is referred to as artificial intelligence software." Holley's is one of the best plain language descriptions of OCR processing for newspapers I have read, and so I will quote her at some length:

> The first step of the OCR software is to analyse the structure of the newspaper page. It divides the page into elements such as blocks of texts (columns), tables, images, etc. The lines are divided into words and then into characters. Once the characters have been singled out, the program compares them with a set of pattern images stored in its database. It analyzes the stroke edge, the line of discontinuity between the text characters, and the background. Allowing for irregularities of printed ink on paper, each algorithm averages the light and dark along the side of a stroke, and advances numerous hypotheses about what this character is. Finally, the software makes a best guess decision on the character. This character is given a confidence rating. The encoding of this confidence is dependent on the software or schema used to represent the OCR results. Therefore, a confidence rating encoded according to the ALTO standard for newspapers is an integer within the range of 0-9, 9 being very confident. A secondary level analysis may then take place at word level (since now a word is formed). The built-in English dictionaries and possibly dictionaries of other languages are checked to see if the word matches. If it does, the confidence rating of the characters may be increased. The built-in dictionaries have

> a complicated relationship with the algorithms and the hypotheses, and how they integrate together is usually kept confidential by the software companies.

Not unlike Dickens's compositors, OCR engines "ma[k]e guesses, with varying luck" by comparing the marks on a given image with "a set of pattern images stored" in memory, and they err toward the expected over the unexpected. As Holley notes, "Some OCR software has the capacity for 'training'" to recognize "old fonts or material distorted" in a regular way, though such training "is incredibly time consuming and has therefore not been used for large scale text digitisation projects."[34] Such training and other refinements characterize much of the research literature on OCR, as computer scientists, often in collaboration with humanists, seek to improve OCR of complex code switching in multilingual 15th–17th century documents;[35] apply methods of handwriting recognition to better identify the close or touching characters, ligatures, and elaborate styling of German Fraktur fonts;[36] or correct OCR errors by aligning the output from multiple OCR engines ("the more the better"), using differences among them to correct errors specific to particular engines;[37] to name just a few notable examples.

By labeling OCR an "automatic" process, we elide this substantial field of research by colleagues in Computer Science. While OCR certainly automates certain acts of transcription, it does so following constantly developing rules and processes devised by human scholars.[38] OCR cannot be said to understand the text it transcribes in the same way as compositors would have understood text in their native languages, but we should recognize with Hayles that "the line between (human) interpretation and (machine) pattern recognition is a porous boundary, with each interacting with the other."[39] By attempting "to replicate the combined functions of the human eye and brain" to recognize and reliably copy alphabetic characters from copy texts, OCR researchers create a "compositor function" that operates across sets of page images. We might think of OCR researchers as analogous to the writers of compositors' rulebooks, outlining templates and guidelines that help the software evaluate images and make decisions about the proper characters to transcribe. OCR performs its work tirelessly, in some sense "automatically," but it is yet a coalescence of human intentions.[40] A text created using OCR must be considered a new edition of its historical source, conversant with but not identical to the images that dominate scholars' experiences of digitized archives. Essentially, the OCR edition of any text can only be understood through attention to digital processes of creation, publication, and access. OCR thus serves as a dramatic illustration of my larger call to take digitized historical texts seriously through the specificities of their medium.

IV. Bibliography for Mass Digitized Editions

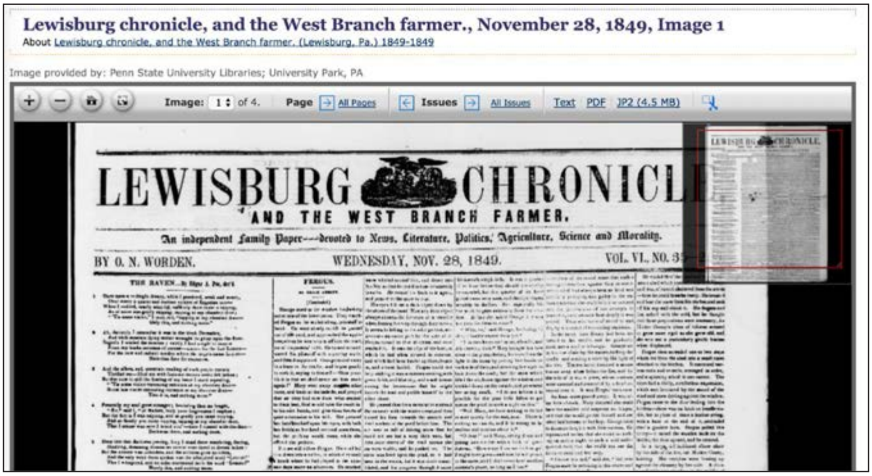
My focus on OCR thus far also pressures myths of surrogacy, in large part because OCR-derived text so rarely *can* stand in for the texts from which it derives. In large-scale historical archives, OCR is rarely presented—or, we might extrapolate, intended—as a reading text for human beings. It exists to facilitate keyword search and other kinds of computational text analysis, and is thus a specific kind of edition created for machine readers. The page images typically presented alongside or overtop OCR are aimed at human readers and imitate the bibliographic codes of their print forebears. A full accounting of any digitized historical text must understand all its constituent parts and their relationships to each other, and to their historical source.

To account for the relationships among the parts that constitute a digitization—multiple images, OCR, interface—we might take as a starting point Fredson Bowers's mid-twentieth-century suggestions for describing proliferating outputs in the machine printing era: "It is necessary to conceive not only of impressions with their issues but also of a family of subsidiary editions stemming from the parent edition type-setting, some having a direct line of descent, others a collateral." Bowers insists to his contemporaries that they must "broaden the old terms to take account of the complexities attending machine-printing," and we must likewise broaden the terms of bibliographic inquiry to take account of mass digitization and hybrid modes of online publication.[41] As Gaskell notes, the industrial printing processes of the machine-press period complicate definitions of edition, though bibliographers largely agree that printings taken from exactly the same typesetting, such as those from stereotype plates, "must . . . be regarded as part of the original edition," and that even "a long pause in a book's printing history . . . does not in itself make a new edition."[42] Following from these principles, we would consider the images provided by large scale online archives as new impressions, belonging to the same editions as the historical originals from which they are derived.

Nevertheless these images too carry important traces of 20th and 21st century processes, like a book reprinted from stereotype plates, but on paper watermarked with a date far distant from the period of its type's setting. A digital page image is offered to readers through an interface, so the image itself is only one element among the "layers and layers of digital tools and interfaces" that comprise a digitization.[43] As a new impression, such an image clearly "operates in reference to, and intermittently transmits something of, the various circumstances associated with the object that is being represented" while simultaneously, through its role as both source text and graphical overlay for the archive's new OCR-derived editions, linking its historical antecedent to its twentieth- and twenty-first-century descendants.[44] Mak describes the layers of a digitization's image, text data, metadata, and interface, "as palimpsests" altering each other, "a particular synthesis of traditional and emergent technologies" that are "challenging to locate for scholarly analysis."[45] These challenges are exemplified (and sometimes exacerbated) by the interfaces of large-scale digital archives, which struggle to reconcile a rhetoric of surrogacy with technical systems that demand media-specific representation.

In this section I outline a bibliographic approach to the mass digitized archive that attempts to account for the complex, palimpsestic, editional families generated through digitization. I organize this discussion around a series of questions in order to suggest a procedure scholars might follow when seeking similar information about other digitized collections. While the precise organization, technologies, and file formats will vary among archives, a technically informed bibliographic accounting can help us move beyond myths of surrogacy and more precisely locate the objects, both historical and contemporary, of our analyses. I do not here propose a system for replacing metadata standards or information models used by libraries and other archival institutions, though many could better represent relationships among historical originals and the digital remediations thereof. I will not have space to delve into library and information science literature around archival models like FRBR (Functional Requirements for Bibliographic Records),[46] OAIS (Open Archival Information System),[47] or EAD (Encoded Archival Description), or metadata standards like the Dublin Core.[48] My aim is more practical. Scholars' research often leads them through multiple digital archives—and digital catalogs of analog materials—developed to diverse standards, some in line with best practices and some not. The procedures described below can help scholars better decipher the information provided through archives' metadata and information models, and to bring that information into conversation with insights gleaned from analysis of archives' digital artifacts themselves. Ultimately, this piece exhorts scholars to investigate and thus better understand the composition (both technical and social) of the digitized archives they use and to integrate such source criticism into any scholarship that makes claims from the digitized archive.

1. A bibliographic investigation of digitized historical materials should first ask, *What metadata does the archive's interface provide about the historical originals of its materials, its digitized edition(s), or both?* The answer to these questions is likely to be a complex version of the latter possibility. Chronicling America's interface, for instance, foregrounds information about the historical newspapers themselves, but also includes clues about their digitization that attentive readers can use to spur further investigation.



Lewisburg chronicle, and the West Branch farmer., November 28, 1849, Image 1

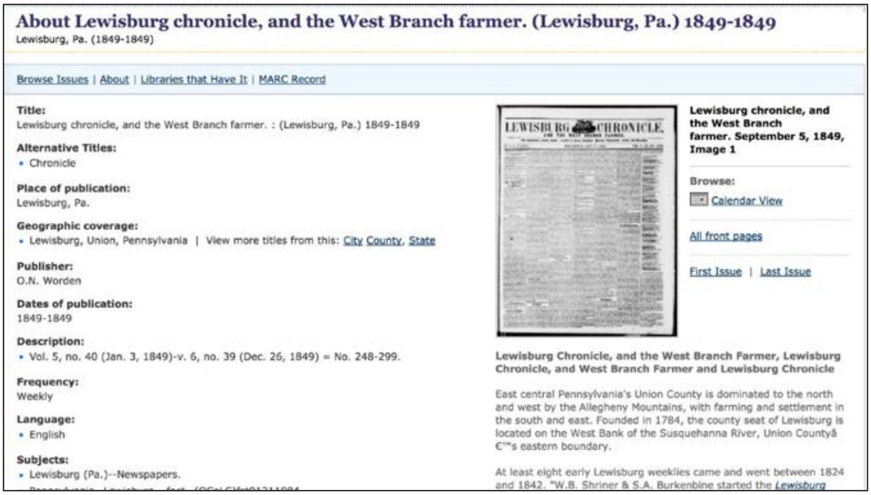
About Lewisburg chronicle, and the West Branch farmer. (Lewisburg, Pa.) 1849-1849

Image provided by: Penn State University Libraries; University Park, PA

Image: 1 of 4.
Page
All Pages
Issues
All Issues
Text
PDF
JP2 (4.5 MB)
LEWISBURG CHRONICLE
AND THE WEST BRANCH FARMER.
An independent Family Paper---devoted to News, Literature, Politics, Agriculture, Science and Morality.
BY O. N. WORDEN.
WEDNESDAY, NOV. 28, 1849.
VOL. VI, NO. 3
THE RAVEN

Figure 2. The presentation of a page from for the Lewisburg Chronicle, and the West Branch Farmer in the Chronicling America interface.

To glean some details about a particular CA digitization, users can click the link at the bottom right of the interface's page viewer (where it reads in Figure 2, "Provided by Penn State . . . ") to see the newspapers digitized by a given NDNP awardee. From the awardee's page for the Penn State University Libraries, one can browse the batches they have contributed to CA and learn that the November 28, 1849, issue of the *Lewisburg Chronicle* was uploaded in batch_pst_fenske_ver02 on July 9, 2013 at 8:07pm.[49] Tracing this fact requires significant dedication to clicking through enigmatically titled links, however. The most prominent metadata provided in CA's interface for individual newspaper issues is found in a headline atop the page viewer (see Figure 2), and lists the historical newspaper's name, issue date, and the image number of the displayed page. This headline mixes description of its historical original—in the newspaper name and date—and one layer of its digitization—in the image number.



About Lewisburg chronicle, and the West Branch farmer. (Lewisburg, Pa.) 1849-1849
Lewisburg, Pa. (1849-1849)

Browse Issues | About | Libraries that Have It | MARC Record

Title:
Lewisburg chronicle, and the West Branch farmer. : (Lewisburg, Pa.) 1849-1849

Alternative Titles:
- Chronicle

Place of publication:
Lewisburg, Pa.

Geographic coverage:
- Lewisburg, Union, Pennsylvania | View more titles from this: City County, State

Publisher:
O.N. Worden

Dates of publication:
1849-1849

Description:
- Vol. 5, no. 40 (Jan. 3, 1849)-v. 6, no. 39 (Dec. 26, 1849) = No. 248-299.

Frequency:
Weekly

Language:
- English

Subjects:
- Lewisburg (Pa.)--Newspapers.

Lewisburg chronicle, and the West Branch farmer. September 5, 1849, Image 1

Browse:
Calendar View

All front pages

First Issue | Last Issue

Lewisburg Chronicle, and the West Branch Farmer, Lewisburg Chronicle, and West Branch Farmer and Lewisburg Chronicle

East central Pennsylvania's Union County is dominated to the north and west by the Allegheny Mountains, with farming and settlement in the south and east. Founded in 1784, the county seat of Lewisburg is located on the West Bank of the Susquehanna River, Union Countyâ €™s eastern boundary.

At least eight early Lewisburg weeklies came and went between 1824 and 1842. "W.B. Shriner & S.A. Burkenbine started the *Lewisburg*

Figure 3. A section of the catalog record for the *Lewisburg Chronicle, and the West Branch Farmer* in the Chronicling America interface.

Clicking on the name of the newspaper just below the title—where, in this case, it reads "About Lewisburg Chronicle . . . "—brings users to a longer library catalogue and narrative description of the newspaper. These prose narratives were written for each CA newspaper by the awardees who digitized them (more on awardees below) and offer insight into the aims, audience, and affiliations of the historical publications, for example describing various instantiations of the *Lewisburg Chronicle* as having "featured infrequent but often extensive ruminations on the emerging great issue of the age—slavery in America, and its extension to new states joining the union—although it would be the mid-1850s before the paper formally identified itself as following a Republican banner."[50]

Perhaps most interesting in this catalog record, however, is its complex representation of the newspaper as an archival object. There are links (e.g. "Browse Issues," "All front pages") that testify, though indirectly, to the particular digitization of the *Lewisburg Chronicle* presented in CA, as well as IDs and other records (e.g. LCCN, OCLC, and MARC) that help link this newspaper to larger systems of information management. The majority of these fields, however, refer to the newspaper as an abstraction rather than a specific material text. In the link to "Libraries that Have It," "It" refers to the *Lewisburg Chronicle, and the West Branch Farmer*, which various libraries have in physical copies or microform. The "It" does not specifically refer to the CA digitization of the *Lewisburg Chronicle*, which is perhaps implied to be had by all through CA itself. As Stauffer argues, such representations of metadata "conflate multiple" and distinct "copies into a single WorldCat entry" and foster "the growth of bibliographic monocultures."[51] Stauffer shows the damage of such models to print collections, which are deaccessioned based on digital accessibility to a single scanned copy, thus losing the evidences of print variation and readerly use across distinct physical copies.

Such models also flatten our understanding of digitizations, encouraging misrecognitions that stymie both computational and bibliographic thinking. When the "It" of an interface refers only to a print edition, it reinforces codexical thinking in the midst of a digital edition rich with machine reading possibilities. By foregrounding the material and social circumstances of digitized archives, however, we both better apprehend our object and defend against the de-accessioning of physical books. If the digitized version of a text is a new, clearly-distinct edition, then it cannot stand as "one representative copy" for all the physical copies of its print original.[52] A "'surrogate first' policy that restricts access to print copy" makes no sense when the digital edition cannot be considered a surrogate.[53]

2. As Mak, Galey, Kirschenbaum and others make clear, an archive's interface is only one layer of a digitization: and, importantly, the layer least useful to most machine reading tasks, which rely on underlying strata of data and metadata. Thus a bibliographic investigation of digitized historical materials should next ask, *What information about this archive's historical originals and their digitization can be gleaned from metadata encoded across the image and text files it provides?* By downloading the files an archive provides, for example, scholars can then use a range of tools to read their Exif (Exchangeable Image File Format) data.[54] This Exif data is created with the images during digitization or format conversion, and can carry information about both the files themselves—such as the size, resolution, or color profile of a given JPG, TIFF, or PDF image—as well as about the digitization process—such as the date of scanning, the source of the scanned image, or the software package used for scanning. Indeed, there is a bibliographic irony here: while scanned page images visually seem the closest simulacrum of their print sources—new impressions of their historical editions—they often carry embedded metadata that more clearly testify to the digitization process than the interfaces through which we access them. However, because these files were created at different moments in the digitization process, quite possibly by people or organizations with distinct pri-

orities, the bibliographic details borne by each file differ. A complete bibliographic account should investigate each available file's metadata and collate the evidence they provide.

In Appendix 1, I have transcribed the full Exif data for the JPG and PDF files of my central *Lewisburg Chronicle* issue, provided through the CA interface, as well as its archival TIFF file. This latter file I learned of through the NDNP's technical guidelines (more on these below), which require each CA issue to include an archive-quality TIFF.[55] That latter file is not served to users as part of CA's interface, likely because they are quite large and it would be server-intensive for too many users to download them frequently.[56] To obtain the archive-quality TIFF of this *Lewisburg Chronicle* issue, I wrote directly to the Library of Congress; after a few days a librarian sent me a link through which I could download the TIFF.

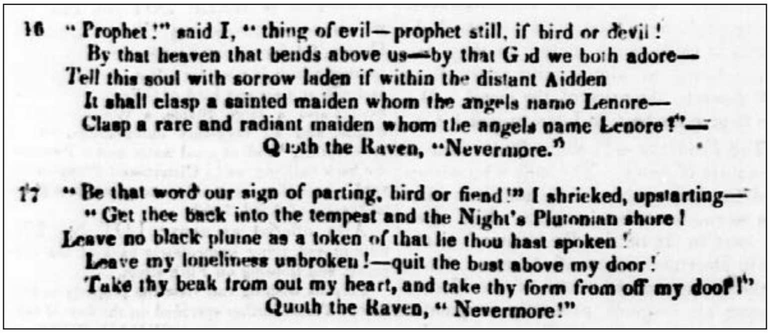
Here I highlight a few details of bibliographic interest gleanable from CA's Exif data. First, metadata describing the image files themselves can be found in the Exif for all three CA image files. A few lines from the JPG Exif data, for instance, read so:

```
Compatible Brands: jp2
Image Height: 6997
Image Width: 5412
Number Of Components: 1
Bits Per Component: 8 Bits, Unsigned
Compression: JPEG 2000
```

These might seem mundane details, but I suggest we compare them to details of page size, format, signing, and collation in descriptive bibliographies of printed books. Those features offer insight into the printing process and allow scholars to distinguish editions, impressions, and states. Similarly, metadata about images' pixel widths or color profiles could help distinguish scanned impressions and establish the order in which the constituent parts of a digitization were created.

The PDF Exif metadata for the *Lewisburg Chronicle* offers more historiographic details, including precisely when the file was originally created: "Create Date: 2012:01:18 12:56:08 " or January 18, 2012 at 12:56 and 8 seconds. In addition, the PDF file includes the following field: "Identifier: Reel number 00280776 35A. Sequence number 50 " that tells us that this newspaper was scanned from microfilm, and which reel was scanned. A scholar might happen to know that the majority of CA's newspapers were scanned from microfilm, but if not (or if working in a less well documented archive) such a detail in Exif metadata would be revealing. Indeed, in CA this detail usefully correlates with what we can glean about the NDNP's digitization priorities from its guidelines (discussed below).

We can triangulate these two data points with a detail only available in the TIFF file metadata, which lists the camera (which can mean a literal camera or, as in this case, a scanner) used to create it: "Camera Model Name: Eclipse 300D, SN# sn632129 ." The Eclipse 300 is a rollfilm scanner produced by the company nextScan which can, according to its promotional materials, "scan 350 pages per minute."[57] That this particular device was used to create our image—particularly when combined with the reel number given in the PDF's Exif metadata—allows us to say with certainty that the *Lewisburg Chronicle, and the West Branch Farmer* was digitized not from a paper original, but from microfilm, and so to know that the OCR text derived from this scan is at least three remediations removed from the historical newspaper. Knowing the scanner—the hardware used—to create these files is just as important to a full understanding of this digital edition as knowing the printing press used is to a full understanding of a printed book or newspaper.

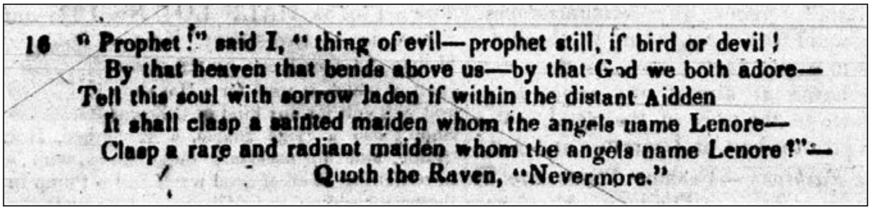


16 "Prophet!" said I, "thing of evil—prophet still, if bird or devil!
By that heaven that bends above us—by that God we both adore—
Tell this soul with sorrow laden if within the distant Aidden
It shall clasp a sainted maiden whom the angels name Lenore—
Clasp a rare and radiant maiden whom the angels name Lenore?"—
Quoth the Raven, "Nevermore."

17 "Be that word our sign of parting, bird or fiend!" I shrieked, upstarting—
"Get thee back into the tempest and the Night's Plutonian shore!
Leave no black plume as a token of that lie thou hast spoken!
Leave my loneliness unbroken!—quit the bust above my door!
Take thy beak from out my heart, and take thy form from off my door!"
Quoth the Raven, "Nevermore!"

Figure 4. Detail from Chronicling America's archival TIFF file for the Lewisburg Chronicle, and the West Branch Farmer's reprinting of "The Raven," showing how high-contrast scanning from microfilm exacerbates inking errors from the original and microfilm.

We see in the image above, for instance, a detail from the Chronicling America page image of the *Lewisburg Chronicle*'s printing of "The Raven," where the left side of the "u" is lightly inked, while the "o" has a diagonal white strip through it. These issues cause the OCR to read uppercase "Q", space, lowercase "i," and lowercase "j" in place of "Quo." The ink is rather



16 "Prophet!" said I, "thing of evil—prophet still, if bird or devil!
By that heaven that bends above us—by that God we both adore—
Tell this soul with sorrow laden if within the distant Aidden
It shall clasp a sainted maiden whom the angels name Lenore—
Clasp a rare and radiant maiden whom the angels name Lenore?"—
Quoth the Raven, "Nevermore."

Figure 5. A high-quality scan of the microfilm for "The Raven" in the Lewisburg Chronicle, including the line that gives this essay its title. Microfilm scanned by the Bucknell University Library. The inking errors from the digital image are less pronounced (though still present) in this earlier remediation.

too heavy in the case of the final "h," leading the OCR to read it as a closed lowercase "b." OCR reads "Q i-jtb the Raven" rather than "Quoth the Raven."

These inking errors are present, though less dramatic, in the microfilm scanned for Chronicling America, which is shown in Figure 5. Scans from microfilm increase contrast in order to highlight letterforms and remove as much visual noise—e.g. dirt specks, stains—as possible. This is done to increase the *overall* effectiveness of the OCR process, but that increased contrast can also exacerbate relatively minor issues such as light inking or damaged type, as shown in the microfilm (which likely exacerbates the same issues, even less dramatic, in the physical copy from which the microfilm was created). In other words, knowing the scanner helps us understand the decisions made during the digitization process which led to the errorful OCR on which our searches and computational analyses rest. These decisions are *understandable*: with limited funding, time, and manpower, scanning from microfilm allowed dramatically more newspapers to be digitized than could have been from paper. But those decisions should also be *understood* by scholars using these resources, as they provide essential context for any work done in mass digitized archives.

Finally, the PDF's Exif data hints at the software package that created it: "Producer: itext-paulo-138 (itextpdf.sf.net-lowagie.com)." iText is a software package for managing large-scale PDF creation, and uses the ABBYY Finereader OCR software.[58] Indeed, we can correlate this snippet with details in the header of another component of this digitized newspaper issue family, the XML file.[59] The *Lewisburg Chronicle*'s XML file, a selection of which is available in Appendix 1, includes important additional metadata about the OCR processing that generated the new machine-readable edition of this newspaper. These XML fields show, for instance, that the OCR was performed using the Abbyy Finereader 9 software, with a "Predicted Word Accuracy" of 98.2 percent.[60] The XML also lists "iArchives OCR" as "processing software." This reference to iArchives points us away from CA's interface and toward its institutional history. iArchives is a company in Lindon, Utah, which many of the first grantees under the NDNP contracted to perform their scanning. This detail in the XML file, then, indicates Penn State University Library did not digitize this newspaper issue in house.[61]

3. This bit of XML metadata reminds us that a full bibliographic account of a digitized text must also concern itself with the institutional, financial, social, and governmental structures that lead one historical object to be digitized, while another is not. The next question for a bibliography of a digitization should be, *What can be learned about the material and sociological processes of this digitization through paratexts such as grant applications, digitization guidelines, or project reports?* In other words, how have the groups and institutions involved in a given digitization shaped the content, structure, and technical capabilities of the archive in which scholars find it? In Ian Milligan's study of newspapers cited in Canadian dissertations, he demonstrates quantitatively that overall citations of newspapers have increased in "the post-database period," but also that those citations draw ever more disproportionately from those papers which have been digitized over those which have not: "Before digitization, a newspaper like the *Ottawa Citizen* was roughly equivalent in historical usage to the *Toronto Star*, as one might expect, given their relative prominence in Canadian history. After the *Star* was digitized and made available, however, it became far more prominent" in dissertations.[62] In other words, decisions about what to digitize ripple throughout the scholarly record from then on, a phenomenon we should mark in scholarship drawn from digitized texts.

In the case of the *Lewisburg Chronicle*, understanding the decisions that led to its digitization requires delving into a range of paratexts related to the United States' National Digital Newspaper Project (NDNP) and its grantees. CA is not a single digitization project run by the Library of Congress, but the portal to data generated by the NDNP, which awards grants to groups in individual states seeking to digitize their historical newspapers. According to a 2014 impact study, NDNP "is a joint partnership between the Library of Congress and the NEH to create a searchable database of culturally significant newspapers from every U.S. state and territory published between 1836 and 1922 . . . . Formally launched in 2004, NDNP grew out of the NEH's United States Newspaper Program, which sponsored the preservation microfilming and description of millions of pages of historic newspapers in every state." The study provides numbers as well, claiming, "NDNP awards enable each state partner to digitize approximately 100,000 pages of historically significant newspapers over a two-year period. States can also apply for supplemental funding, and a number have already completed the digitization of more than 300,000 pages. At present, approximately 8 million pages of historic newspapers from 32 states and the District of Columbia have been digitized."[63] The newspapers digitized through NDNP grants are often collected in state-level archives as well as being aggregated in CA. Thus one can find webpages for the California Digital Newspaper Archive, the North Carolina Digital Newspaper Archive, and so forth, which may include some new content not yet aggregated in CA.

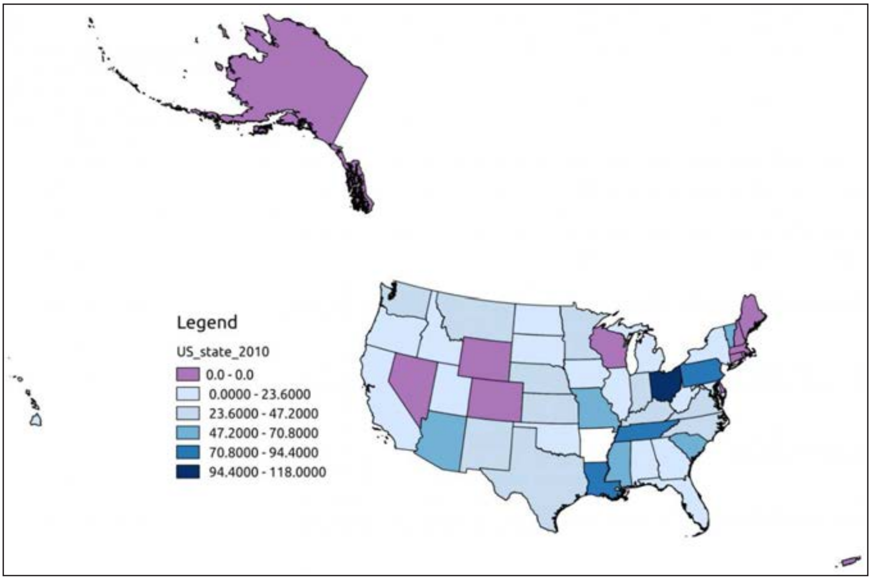


Figure 6. A map of state contributions to Chronicling America as of 2015, prepared by Viral Text Project research assistant Abby Mullen. States are darker based on how many historical newspapers they have contributed to CA.

Such state-level granting, however, means that some states are well represented, others less so, while many are not represented at all. For instance, my home state of Massachusetts has not yet participated in the NDNP, meaning that CA includes no papers from Boston or other Massachusetts towns. In addition, NDNP's state-level granting results in disparate selection criteria from state to state. In their applications for NDNP funding, groups must articulate a rationale for choosing "historically significant newspapers" from their state. While these rationales share many features, they are not identical.[64] Penn State University Libraries was one of the first NDNP grantees and they have received two rounds of supplemental funding. As a result, they are the third largest CA contributor, having added 364,350 pages to the repository as of March 5, 2015, putting them behind only the Library of Congress itself (1,019,504 pages) and the Library of Virginia in Richmond (381,900 pages).[65] The three phases of the Pennsylvania Digital Newspaper Project (PaDNP) are described on a Penn State University Library webpage, while Phases I and II are also documented through blogs hosted at PSU.[66] These blogs, taken together with Penn State's grant proposals, press releases, and NDNP program digitization guidelines, provide unique insight into the processes through the *Lewisburg Chronicle*, a rural paper from Union County, came to be collected in CA while nineteenth-century Philadelphia is represented by a single newspaper, the *Evening Telegraph*. The Phase I blog, for instance, includes posts in which participants propose candidates for digitization, discuss the availability of those titles through various libraries and the feasibility of scanning them, and winnow the candidates for both intellectual and practical reasons.

In short, Penn State's plans for digitization emphasized certain kinds of geographic and demographic representativeness over others. Under their Phase 1 grant of $393,65 (2008–10), for instance, the group digitized Pennsylvania newspapers published between 1880 and 1922. The newspapers chosen were "currently on microfilm" and were chosen by Penn State in consultation with the State Library of Pennsylvania and the Free Library of Philadelphia. These stakeholders used census data to identify the 10 cities with the largest populations between 1880–1922, and from these cities, "48 publications were reviewed for initial consideration, with the final selection made by an advisory board of researchers, scholars, librarians and historians." PaDNP ultimately digitized four titles in Phase I, the *Lancaster Daily*, the *Pittsburg Dispatch*, the *Scranton Tribune*, and the *Evening Public Ledger*.[67]

Penn State's Phase II grant of $393,489 (2010–12) expanded PaDNP's time frame to 1836–1922 while "title selection efforts focused on 17 Pennsylvania counties without any known digitized newspapers." In Phase II PaDNP digitized 45 new titles, including 151,968 pages, from the State Library of Pennsylvania, the Free Library of Philadelphia, Bloomsburg University Library, and the Pennsylvania Historical and Museum Commission. Finally, Penn State's Phase III grant of $321,526 (2012–14) allowed the group to digitize 109,025 pages from 39 newspapers published between 1836–1922. The foci for this round included:

- Titles from four counties with very little or no digitization
- Tiles that represent the Commonwealth's German and Italian ethnic heritage
- Titles that cover World War I and the Spanish influenza epidemic.[68]

Over their three rounds of funding, then, Penn State sought to digitize newspapers from as many counties as possible, meaning they prioritized breadth of geographic coverage over trying to ascertain the most influential newspapers in the state or other measures of diversity. Phase II's blog includes a graph and map summarizing the situation of Pennsylvania newspaper digitization as of 2010. Union County, home of the *Lewisburg Chronicle*, is notably blank in the map, which is why this county was one of 17 chosen for Phase II of the PaNDP.[69] The *Lewisburg Chronicle, and the West Branch Farmer* was scanned in Phase II of the PaNDP because its county had been ignored in previous digitization efforts, both public and commercial. Reading the narrative about the *Lewisburg Chronicle*, prepared for the PaNDP and now provided on CA, we understand that the paper was chosen among those published in Union County because of its stability, as "At least eight early Lewisburg weeklies came and went between 1824 and 1842" before the *Chronicle* brought a regular publication to the county.[70]

I cite these details neither to defend nor deride the choices made by PaDNP. Given finite time and resources, any mass digitization effort must privilege certain features from all possible corpora over others. Had the PaDNP chosen papers based on their historical influence, they would have produced a corpus skewed in another way: toward Philadelphia over more rural areas in the state. Such geographic diversity, however, mutes other kinds of diversity a newspaper archive might strive toward. As Benjamin Fagan points out, "While digital copies of 215 white newspapers published before 1865 have been made publicly available through the Library of Congress's Chronicling America project, that archive contains no black newspapers printed during the same period." Fagan notes that 46 black newspapers from after 1865 have been digitized for the project, but this is from an overall archive of 1,799 newspapers at the time of his article's publication, meaning that black newspapers make up less than 3% of the overall titles in CA.[71] Researchers building arguments, computational or otherwise, from CA or its subsidiary archives must understand such choices and qualify their claims in light of the socio-technical processes that shape mass digitization, including analog archival histories that proscribe the boundaries of possibility for contemporary digitization.

4. A historical object in a large scale digital archive evidences not only the time period of its print original and its online present. It witnesses also to cultural and scholarly modes of transmission that bridge the period of the historical object's creation and our experience of its digitized editions. The final question for a thorough bibliography of a digitization might then be: *What can be learned about a given digitization through paratexts about previous remediations?* Clues throughout the files in CA and related records show how the digitized edition of the November 28, 1849, *Lewisburg Chronicle, and the West Branch Farmer* mediates not only its paper original, but also intermediary states. The vast majority of newspapers digitized for the NDNP are scanned from microfilm cataloged or created between 1982–2011 under the auspices of the US Newspaper Program (USNP), "a cooperative national effort among the states and federal government to locate, catalog, and preserve on microfilm newspapers published in the United States from the eighteenth century to the present."[72] As they were for the later NDNP, Pennsylvania State University was one of the earliest participants in the USNP: "As an early applicant of the USNP in 1985, Pennsylvania held the distinction of being the largest state geographically to participate in the USNP . . . . Moreover, the site at Penn State was the first to conduct field work due to the lack of a large repository in central Pennsylvania." The PaNP received $1,903,196 in NEH grant funding. A record of the Pennsylvania Newspaper Project's activities between 1985–88 can be found on a PSU website, which includes biographical details of the PaNP's staff, yearly reports from the project's field team, and technical details of the project's work.[73]

The PaNP reports are, frankly, remarkable historical and bibliographic records in their own right, attesting to the herculean efforts of field librarian Becky Wilson and field catalog librarian Sue Kellerman. These scholars "journeyed throughout 30 rural counties in central Pennsylvania seeking out collections of newspapers, old and new," visiting "public libraries, historical societies, newspaper publishing offices, and in collections held by private citizens" in order to "catalog all newspapers ever published in Pennsylvania." These reports are far from dry records, and instead transcribe from the field team's notebooks many narrative (and often quite amusing) anecdotes of their travels across the state; interactions with newspaper collectors, local historians, and librarians; and the difficult detective work of the project. To quote only one short example from Union County (home of the *Lewisburg Chronicle*), the field team reported in 1985, “In Union County we ran into our first non-cooperative collector, Harry Feltman, reputed to have the finest collection of papers in Union County, and who will not return our calls or answer our letters.”[74] Wilson and Kellerman’s reports for the PaNP deserve more sustained treatment than I have room for in this article, but I want to briefly focus on how they help further historicize the digitization of the *Lewisburg Chronicle* in CA. Their 1985 report is the only to mention visits to Lewisburg, while the March entry in the 1986 report lists Union among “14 Counties Completed as of end of March 1986.”[75] In particular, the 1985 report notes that “in Union County we visited Bucknell University, Union County Historical Society, Packwood Museum, Herr Memorial Library, two newspaper offices, New Berlin Heritage Association, four private individuals, and examined the contents of boxes from an estate (left to Bucknell).” Their visit to Bucknell is perhaps most important to this account, as the Bertrand Library at Bucknell is one of the only that owns the 1849 *Lewisburg Chronicle, and the West Branch Farmer* on microfilm. Indeed, the Library of Congress’ *Catalog: Newspapers in Microform, United States, 1948–1983* lists only four institutions holding this paper: microfilm masters were cataloged at MICOR (The Micrographics Corporation) in Cornell Heights, Pa., and the Microfilm Corporation of Pennsylvania in Pittsburgh, while microfilm service copies were cataloged at Pennsylvania State Library in Harrisburg and Bucknell University in Lewisburg.[76] While I cannot yet say with certainty which microfilm was used for the *Lewisburg Chronicle* digitization, it is certain that Wilson and Kellerman’s research contributed directly to the NDNP’s decisions 25–30 years later about which papers could and should be digitized, as well as to the newspaper descriptions now found on CA.

A digitized historical text bears traces of its original, the processes of its digitization, and a series of decisions over decades or centuries about documentation, collection, access, and preservation. If taken seriously by scholars, moments of archival remediation can uniquely illuminate the processes through which our scholarly sources come to be, which in turn can instruct our efforts to build more democratic and representative historical collections. More immediately, however, an understanding of a digital corpus’s sociological outlines and the technical composition of its materials allows us to qualify the claims we make from it while benefiting from the possibilities of access, comparison, or analytical scale enabled by digitization. Beginning from the specific example of the *Lewisburg Chronicle*, we can see that the constitution and provenance of digitized archives are, to some extent at least, knowable and describable. Just as details of type, ink, or paper, or paratext such as printer's records can help us establish the histories under which a printed book was created, details of format, interface, and even grant proposals can help us establish the histories of corpora created under conditions of mass digitization.

V. Conclusion

The digital November 28, 1849, *Lewisburg Chronicle, and the West Branch Farmer* is a family of new editions and impressions comprising at least six parts: an archival TIFF, a JPG, a PDF, an OCR-derived text file, an XML file, and a web interface. Bibliographic clues about the technical and social processes behind its digitization are scattered among its interface, its constituent files—not all of which are available through CA's public interface—and paratexts of the NDNP and related programs. While the precise parts of a digitization may differ among mass digitized archives, all digitizations comprise layers of interface, image, and text that can offer unique bibliographic clues. Scholars should come to the digitized archive primed to analyze the interactions and tensions among a single digitization's editions and impressions. One of the most compelling reasons to take bibliography seriously for digitized historical texts is that doing so forefronts their createdness: the chain of human labor that led to their present existence. When we apprehend the November 28, 1849, *Lewisburg Chronicle* primarily as a surrogate, we elide not only the scanning and OCR processing this article has primarily discussed, but also a series of human acts by many people over 150 years, including preservation, curation, collection, cataloging, and description. As Jerome McGann insists—channeling D.F. McKenzie—"No book"—or no newspaper—"is one *thing*, it is many *things*, fashioned and refashioned repeatedly under different circumstances. Its meaning . . . is in its use."[77] Digitization does not remove a historical artifact from material culture, but adds another stratum of computational materiality to its social text.

Given this fact, what then is the meaning of the mass digitized text when used through forms of machine reading, whether keyword search or topic modeling? First, its meaning cannot *only* be surrogacy. Even when a scholar finds an article through keyword search and then reads it *as though it were* its paper original, the fact of its finding necessitates that it be contextualized as a digital artifact within a particular editional family. Second, its meaning must be relational, enmeshed in the hypertextual and database structures through which we identify, compare, and discuss digitized historical texts. Printed materials also inhere within networks of relationship, of course, but the unique affordances of the digital medium—pattern detection across vast textual fields—foreground connections among texts over textual uniqueness. The digital archive flattens the material texts it represents in ways that foster machine reading and foreclose some varieties of close reading. I offer this point as neither praise nor condemnation, but as another insistence that we must grapple with the digital on its own terms rather than through the lens of surrogacy.

In closing, I turn to the practical questions these reflections raise for scholars working with digitized materials, and in particular for those involved in computational analysis at the corpus scale. I see no reason to single out other scholars for oversights I have shared in past publications, so instead I will illustrate the following points through personal reference. On the one hand, for projects drawing on many thousands of digitized texts (as we do with newspapers in the Viral Texts Project) it would be unreasonable to expect description of them all at the level of detail I demonstrate in this article. In an earlier piece, for instance, we described our data thus: "In the first iteration of the project, we focused on pre-Civil War newspapers in the Library of Congress's Chronicling America online newspaper archive, in large part because its text data is openly available for computational use. The pre-1861 holdings comprise 1.6 billion words from 41,829 issues of 132 newspapers." Creating a full critical bibliography for even the 132 separate digitized newspapers—putting aside individual issues—would require far more time, effort, and money than the project's entire scope or budget would allow.[78]

However, it is likewise evident that the broad strokes with which we outline the archival context of our research in the quote above are insufficient. Descriptions of source data in computational text analysis often resemble ours: they are accounts of the size of the data in works, in words, or in gigabytes, but offer little account of provenance. A welcome exception to this tendency can be found in Matthew Wilkens's "The Geographic Imagination of Civil War-Era American Fiction," in which he summarizes the principles of selection for the Wright American Fiction Collection:

> The literary corpus used in the present study . . . is based on the volumes cataloged by Lyle Wright in his *American Fiction, 1851–1875: A Contribution toward a Bibliography* (1958). Wright's bibliogra-

> phy attempts to list "the fiction . . . written for adults by Americans and printed in the US" between the dates of his title; he specifically excludes reprints, religious tracts, children's literature, genres other than narrative fiction, serials, and books by non-American writers published in the US. Wright consulted both physical copies held in a range of libraries and lists of newly published titles from contemporary sources in the compilation of his bibliography.[79]

Wilkens's brief bibliographic account conveys the core of his corpus's historical composition. Paired with robust technical description of sampled works from a given corpus, similar accounts would more clearly mark the bounds within which we make arguments in the database age. Indeed, we must develop protocols for data sampling that would allow scholars to use specific examples to estimate bibliographic characteristics across mass corpora such as CA.

Energetic work is needed from bibliographers and book historians in constructing bibliographies of mass digitized archives, including those that resist scholarly description, such as Google Books. It is quite likely that Google would not eagerly share all the data and metadata for their holdings in the same way that CA does, but investigative work along the lines I propose here is nonetheless possible, for instance by consultation with the academic libraries who partnered to digitize many of the materials in Google Books. Robust, critical bibliographies of mass digitized archives would federate information available through interfaces, metadata, and paratextual materials in order to outline the technical and social composition of such resources.

Acknowledging digitized historical texts as new editions is an important first step toward developing media-specific approaches to the digital that more effectively exploit its affordances; more responsibly represent the material, social, and economic circumstances of its production; and more carefully delineate its limitations. Massive, errorful OCR archives necessitate close bibliographic and book-historical attention that both leverages their powers while historicizing their creation and use. Importantly, such historicizing does not foreclose the possibility of new OCR editions that benefit from our colleagues' research. Instead, it recognizes the realities of our editions at a given moment—and the limits those realities place on research—while giving us a vocabulary for naming the iterative realities of digitized text. We must ask what kinds of questions mass digitized editions might make more tractable, and how those questions can more consciously dovetail with archival, bibliographic, and book historical research priorities. In order to learn how to ask pressing humanities questions best answered through computational means, we must take the digitized text seriously within its own medium. That, I would argue, is the primary challenge facing bibliographers and book historians in our moment: not a technical challenge, but a challenge of imagination.

Appendix 1: Chronicling America File Format Data

Below is the metadata investigated for the November 28, 1849, issue of the *Lewisburg Chronicle, and the West Branch Farmer* in four file formats, as provided through the Chronicling American interface or, in the case of the TIFF file, directly by email from the Library of Congress. I read this Exif data using the the free command line application Exiftool (http://www.sno.phy.queensu.ca/~phil/exiftool/). I have reproduced the entire Exif data for the JPG, TIFF, and PDF files, and a portion of the TEI/XML header. This metadata is discussed in Section IV, "Bibliography for Mass Digitized Editions," above.

1. *JPG Exif metadata*

```
ExifTool Version Number: 10.01
File Name: seq-1.jp2
Directory: .
File Size: 4.5 MB
File Modification Date/Time: 2015:12:10 15:55:47+01:00
File Access Date/Time: 2015:12:11 13:37:35+01:00
File Inode Change Date/Time: 2015:12:10 15:55:48+01:00
File Permissions: rw-r-----
File Type: JP2
File Type Extension: jp2
MIME Type: image/jp2
Major Brand: JPEG 2000 Image (.JP2)
Minor Version: 0.0.0
Compatible Brands: jp2
Image Height: 6997
Image Width: 5412
Number Of Components: 1
```

```
Bits Per Component: 8 Bits, Unsigned
Compression: JPEG 2000
Color Spec Method: Enumerated
Color Spec Precedence: 0
Color Spec Approximation: Not Specified
Color Space: Grayscale
Warning: Can’t currently handle huge JPEG 2000 boxes
Image Size: 5412x6997
Megapixels: 37.9
```

2. *PDF Exif Metadata*

```
ExifTool Version Number: 10.01
File Name: seq-1.pdf
Directory: .
File Size: 826 kB
File Modification Date/Time: 2015:12:10 15:55:50+01:00
File Access Date/Time: 2015:12:10 16:10:22+01:00
File Inode Change Date/Time: 2015:12:10 15:55:55+01:00
File Permissions: rw-r-----
File Type: PDF
File Type Extension: pdf
MIME Type: application/pdf
PDF Version: 1.4
Linearized: No
Page Count: 1
Page Mode: UseNone
Format: application/pdf
Title (en): Lewisburg Chronicle, and the West Branch Farmer..(Lewisburg, Pa.) 1849-11-28 [p ].
Description (en): Page from Lewisburg Chronicle, and the West Branch Farmer. (newspaper). [See LCCN: sn85055199 for catalog record.]. Prepared on behalf of Penn State University Libraries;     University Park, PA.
Date: 1849:11:28
Type: text, newspaper
Identifier: Reel number 0028077635A. Sequence number 50
Create Date: 2012:01:18 12:56:08-07:00
Producer: itext-paulo-138 (itextpdf.sf.net-lowagie.com)
Modify Date: 2012:01:18 12:56:08-07:00
```

3. *TIFF Exif Metadata*

```
ExifTool Version Number: 10.01
File Name: 0050.tif
Directory: .
File Size: 36 MB
File Modification Date/Time: 2015:12:10 20:34:01+01:00
File Access Date/Time: 2015:12:11 14:25:24+01:00
File Inode Change Date/Time: 2015:12:10 20:34:01+01:00
File Permissions: rw-r-----
File Type: TIFF
File Type Extension: tif
MIME Type: image/tiff
Exif Byte Order: Little-endian (Intel, II)
Subfile Type: Single page of multi-page image
Image Width: 5409
Image Height: 6997
Bits Per Sample: 8
Compression: Uncompressed
Photometric Interpretation: BlackIsZero
Fill Order: Normal
Document Name: 0028077635A
Make: Eclipse
Camera Model Name: Eclipse 300D,SN# sn632129
Strip Offsets: 493
Orientation: Horizontal (normal)
Samples Per Pixel: 1
Rows Per Strip: 6997
Strip Byte Counts: 37846773
X Resolution: 300
Y Resolution: 300
Planar Configuration: Chunky
Resolution Unit: inches
Page Number: 0 1
Software: iArchives, Inc., 3.240
Modify Date: 2012:01:18 12:55:32
Artist: Penn State University Libraries; University Park,
PA; iArchives
File Source: Unknown (microfilm)
```

```
Image Unique ID: 50
Image Size: 5409x6997
Megapixels: 37.8
```

4. *XML Header*

```
<sourceImageInformation>
    <fileName>
/mnt/192.168.101.196/data01/jobq/root/projects/production/
newspaper/Penn_State/NDNP_2010/batch_fenske/LCF_18490905-
18530325/ocr/0050.tif
    </fileName>
</sourceImageInformation>
<OCRProcessing ID=”OCR.0”>
    <ocrProcessingStep>
        <processingStepSettings>
Conf:0.982abbyy9.option.analyze-manual-zones:falseL
ang:engInverted:falseConf.SPead:0.982source-image:/
mnt/192.168.101.196/data01/jobq/root/projects/production/
newspaper/Penn_State/NDNP_2010/batch_fenske/LCF_18490905-
18530325/ocr/0050.tifCharacter     Count:23869Abbyy9.cache-
check.base-page-CRC:d254ba62a1df36afbda05ea115359d19Predic
ted  Word  Accuracy:98.2%Abbyy9.cache-check.wrapper-version:
1.3Node   Count:5572abbyy9.option.ocr-auto-pictures:trueSus
picious  Character  Count:2357Abbyy9.cache-check.image-CRC:8
0460689eebc65a511ad6a5614064cbbDictionary Words:4824abbyy9.
option.hyphenation:trueabbyy9.version:9.0.0.7394-3Engine:
Abbyy9abbyy9.option.analyze-zones:trueOption Count:5572Word
Count:5572width:5409height:6997xdpi:300ydpi:300
        </processingStepSettings>
        <processingSoftware>
            <softwareCreator>iArchives</softwareCreator>
            <softwareName>iArchives OCR Framework</software-
            Name>
            <softwareVersion>Multiple</softwareVersion>
        </processingSoftware>
    </ocrProcessingStep>
</OCRProcessing>
```

Notes

1. For more on the Viral Texts Project at Northeastern University, see http://viraltexts.org.

2. OCR is a term for computer programs that identify machine-readable words from a scanned page image, and is the source for most of the searchable data in large-scale digital archives.

3. W.W. Greg, "Bibliography—an Apologia," in *Collected Papers*, ed. J.C. Maxwell (Oxford: Clarendon Press, 1966), 257, 259. Originally printed in *The Library* 8 (Sep. 1932).

4. Juxta `, http://juxtacommons.org/.

5. http://chroniclingamerica.loc.gov/lccn/sn85055199/1849-11-28/ed-1/seq-1/.

6. Matthew Kirschenbaum and Sarah Werner, "Digital Scholarship and Digital Studies: The State of the Discipline," *Book History* 17 (2014): 417.

7. Alan Galey "The Enkindling Reciter: E-Books in the Bibliographical Imagination," *Book History* 15 (2012): 214.

8. Galey, "The Enkindling Reciter," 211, 213, 240.

9. Galey, "The Enkindling Reciter," 211, 214.

10. Matthew Kirschenbaum, "Text Messaging," in *Mechanisms: New Media and the Forensic Imagination* (Cambridge, Mass.: MIT Press, 2008), 17.

11. Nathan Altice, *I Am Error: The Nintendo Family Computer / Entertainment System Platform* (Cambridge, Mass.: MIT Press, 2015).

12. Jerome McDonough, Matthew Kirschenbaum, Doug Reside, Neil Fraistat, and Dennis Jerz, "Twisty Little Passages Almost All Alike: Applying the FRBR Model to a Classic Computer Game," *Digital Humanities Quarterly* 4, no. 2 (2010), http://www.digitalhumanities.org/dhq/vol/4/2/000089/000089.html.

13. N. Katherine Hayles, "How We Read: Close, Hyper, Machine," *ADE Bulletin* 150 (2010): 72.

14. For a useful investigation of citation practices for digital resources in American literary scholarship, see Lisa Spiro and Jane Segal, "Scholars' Usage of Digital Archives in American Literature," in *The American Literature Scholar in the Digital Age*, ed. Amy E. Earhart and Andrew Jewell (Ann Arbor: University of Michigan Press, 2011), http://dx.doi.org/10.3998/etlc.9362034.0001.001.

15. John W. Maxwell, "E-Book Logic: We Can Do Better," *Papers of the Bibliographical Society of Canada*, 51, no. 1 (2013): 43

16. Whitney Anne Trettien, "A Deep History of Electronic Textuality: The Case of *English Reprints Jhon Milton Areopagitica*," *Digital Humanities Quarterly* 7, no. 1 (2013), http://digitalhumanities.org:8081/dhq/vol/7/1/000150/000150.html.

17. Jerome McGann, "From Text to Work: Digital Tools and the Emergence of the Social Text," *Romanticism on the Net* 40–41 (February-May 2006), http://www.erudit.org/revue/ron/2006/v/n41-42/013153ar.html, paragraph 33.

18. Ted Underwood, "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago," *Representations* 127, no. 1 (Summer 2014): 64.

19. Hayles, "How We Read," 72.

20. The Text Encoding Initiative (TEI), for instance, is a consortium of scholars, librarians, and technologists "which collectively develops and maintains a standard for the representation of texts in digital form . . . chiefly in the humanities, social sciences and linguistics." The TEI is a specific variety of XML, a more broad-based data encoding scheme, that focuses on encoding scholarly materials. A good many digital archival resources for the humanities will be encoded in TEI, though the level of detail in that encoding varies widely based on the archives' sizes, funding, staffing, and so forth. In many cases, tags drawn from TEI standards will encode only basic metadata about largely uncorrected body text, while in other (often smaller) proj-

ects, the TEI is used for intricate encoding of bibliographic details to the level of the word. The latter level of detail is extremely uncommon in mass digitization projects, however, for reasons I discuss later in this section. For more on the TEI see http://www.tei-c.org/index.xml.

21. Though given the repetition of that phrase in this particular poem, it would find the *Lewisburg Chronicle* witness of the poem. The problem with errorful OCR is more dire in cases of a search term not so frequently repeated within the pertinent document.

22. Andrew Stauffer, "The Nineteenth-Century Archive in the Digital Age," *European Romantic Review* 23, no. 3 (2012).

23. See for instance Bonnie Mak, "Archaeology of a Digitization," *Journal of the Association for Information Science and Technology* 65, no. 8 (August 2014): 1519, and the website of the Text Creation Partnership http://www.textcreationpartnership.org/home/. The outsourcing of hand correction itself raises important ethical issues about which scholars should be engaged.

24. Katherine Bode, "Thousands of Titles Without Authors: Digitized Newspapers, Serial Fiction, and the Challenges of Anonymity," *Book History* 19 (2016): 284–316, doi: 10.1353/bh.2016.0008.

25. Marie-Louis Ayres, "'Singing for Their Supper': Trove, Australian Newspapers, and the Crowd," paper presented at the 79th IFLA World Library and Information Congress, Singapore 2013, accessed 11 July 2016, http://library.ifla.org/245/1/153-ayres-en.pdf.

26. To put this in perspective, I made a rough line count of a nineteenth-century newspaper in my personal collection, the *Salem Gazette* of May 28, 1811. Keeping in mind variations in layout due to advertisements, mastheads, headlines, and so forth, I count approximately 100–120 lines per column, with 4 columns per page. Even taking the lower figure to calculate estimate, this would mean Trove includes at least 8.4 billion lines across its 21 million pages. Thus 100 million corrected lines make up approximately 1.2% of the archive's text. The 4 lines per article hand-corrected offshore would increase this percentage somewhat, but I suspect still would not translate to more than a single-digit percentage of the total searchable text in the archive.

27. Philip Gaskell, *A New Introduction to Bibliography*, St. Paul's Bibliographies (New Castle, Delaware: Oak Knoll Press, 1995), 313.

28. G. Thomas Tanselle, "The Bibliographic Concepts of 'Issue' and 'State,'" *The Papers of the Bibliographic Society of America* 69, no. 1 (First Quarter, 1975): 18.

29. My thinking on these distinctions owes much to David L. Gants, who in a talk on digital bibliography at the Rare Book School in Charlottesville, Virginia (July 2016), explored a similar line of reasoning and who in conversation afterwards helped me refine my own definitions.

30. Gaskell, *New Introduction to Bibliography*, 113.

31. Charles Manby Smith, *The Working-Man's Way in the World; Being the Autobiography of a Journeyman Printer* (New York: Redfield, 1854), 305, 304. Accessed via the Internet Archive, 30 July 2016, https://archive.org/stream/workingmanswayin00smit#page/304/mode/2up/search/paper.

32. Horace Hart, *Rules for Compositors and Readers at the University Press, Oxford*, Nineteenth Edition (London: Henry Frowde, 1905).

33. Sylvere Monod, "'When the Battle's Lost and Won . . . ': Dickens *v.* the Compositors of Bleak House," *The Dickensian* 69 (1973): 5, 12.

34. Rose Holley, "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs," *D-Lib Magazine* 15, no. 3–4 (March/April 2009), http://www.dlib.org/dlib/march09/holley/03holley.html.

35. Dan Garrette et al., "Unsupervised Code-Switching for Multilingual Historical Document Transcription," NAACL 2015, http://www.aclweb.org/anthology/N/N15/N15-1109.pdf.

36. Thomas M. Breuel et al., "High-Performance OCR for printed English and Fraktur using LSTM networks," ICDAR 2013, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.433.4006&rep=rep1&type=pdf.

37. William B. Lund et al., "How Well Does Multiple OCR Error Correction Generalize?," DRR 2014, http://www.researchgate.net/publication/260084914_How_Well_Does_Multiple_OCR_Error_Correction_Generalize.

38. The body of OCR research is too broad to succinctly cite here, but the main conference in the field is ICDAR (International Conference on Document Analysis and Recognition), http://2015.icdar.org/, with relevant work often presented—and collected in the conference proceedings of—CVPR (Computer Vision and Pattern Recognition), EMNLP (Empirical Methods in Natural Language Processing), SIGIR (Special Interest Group in Information Retrieval), and NIPS (Neural Information Processing Systems). It should be noted for *Book History* readers unfamiliar with the conventions of computer science research that presentations at top flight conferences (almost all with published proceedings) are generally preferred to journal publication, as the prior are more timely and widely read.

39. N. Katherine Hayles, "How We Read," 72–73.

40. For the lovely phrase "coalescence of human intentions," I am indebted to Michael F. Suarez in several courses at the Rare Book School in Charlottesville, Virginia.

41. Bowers, *Principles of Bibliographic Description:* 382–83.

42. Gaskell, *New Introduction to Bibliography*, 313–14.

43. Galey, "The Enkindling Reciter," 240.

44. Mak, "Archaeology of a Digitization," 1516.

45. Mak, "Archaeology of a Digitization," 1515.

46 See http://www.ifla.org/publications/functional-requirements-for-bibliographic-records.

47 See http://public.ccsds.org/publications/archive/650x0m2.pdf.

48 See http://dublincore.org/.

49. http://chroniclingamerica.loc.gov/awardees/pst/ and http://chroniclingamerica.loc.gov/batches/batch_pst_fenske_ver02/.

50. http://chroniclingamerica.loc.gov/lccn/sn85055199/.

51. Stauffer, "My *Old Sweethearts*: On Digitization and the Future of the Print Record," *Debates in the Digital Humanities 2016*, ed. Matthew K. Gold and Lauren F. Klein (Minneapolis: University of Minnesota Press), 2016, 220.

52. Stauffer, "Nineteenth-Century Archive," 340.

53. Laurel Brake, "London Letter: Researching the Historical Press, Now and Here," *Victorian Periodicals Review* 48, no. 2 (Summer 2015): 251.

54. I read the Exif data of the JPG, PDF, and TIFF files for the November 28, 1849, *Lewisburg Chronicle, and the West Branch Farmer* using the the free command line application Exiftool http://www.sno.phy.queensu.ca/~phil/exiftool/.

55. CA's technical guidelines have shifted over time, but scholars can access each version, depending on the date their issue of interest was digitized: • Technical Guidelines for 2007 and 2008 Awards, http://www.loc.gov/ndnp/guidelines/archive/techspecs0708.html • Technical Guidelines for 2009 and 2010 Awards, http://www.loc.gov/ndnp/guidelines/archive/techspecs09.html • Technical Guidelines for 2012 Awards, http://www.loc.gov/ndnp/guidelines/archive/guidelines1213.html

56. The TIFF file for just the first page of the November 28, 1849, *Lewisburg Chronicle*, for instance, is nearly 40MB, which means the four pages of this issue would approach 160MB. One can extrapolate from here to understand the servers that would be required to host hundreds of thousands of pages for ready download, when only a fraction of users likely require those files for the bibliographic work I outline here.

57. The Eclipse 300 Microfilm Scanner was released by nextScan in 2009, http://www.nextscan.com/nextscan-introduces-the-eclipse-plus-high-production-rollfilm-scanner-with-lumintec-light-line-illumination-system/#.VRF8DVxCNFI.

58. http://itextpdf.com/.

59. This location of this file is not obvious, though one can find it through the CA website. In short, one must click the "Text" link to the left of the PDF and JPG links for the issue. This brings up a simplified interface that reproduces a small version of the page image and the OCR text derived from it. There is no apparent metadata here, but at the very bottom of this page one can find a link to an XML file that encodes the raw text. XML is a markup language, like the more familiar HTML (HyperText markup Language) that underlies most webpages. XML's primary use is to storing data in a format that is both human- and machine-readable. The most familiar version XML in the humanities is the TEI, or Text Encoding Initiative, which is a group "whose mission is to develop and maintain guidelines for the digital encoding of literary and linguistic texts." "About the TEI," http://www.tei-c.org/About/.

60. Because Abbyy Finereader is a commercial product, the software that predicts its accuracy is not freely available for inspection. As such, we should not make too much of the figure presented here, which certainly does not align with a human reader's assessment of the page's overall similarity to the words on the page images.

61. In October of 2010 iArchives was acquired by Ancestry.com (http://Ancestry.com), and their technologies essentially became the for-profit Newspapers.com (http://Newspapers.com) that feeds many of the results you get when searching family trees in Ancestry.

62. Ian Milligan, "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010," *Canadian Historical Review* 94, no. 4 (December 2013): 550, 567.

63. Jamie Mears, *National Digital Newspaper Program Impact Study 2004–2014*, National Endowment for the Humanities (September 2014), https://www.loc.gov/ndnp/guidelines/docs/ndnp_report_2014_0.pdf.

64. Mears, *National Digital Newspaper Program Impact Study 2004–2014*.

65. http://chroniclingamerica.loc.gov/awardees/pst/ and http://chroniclingamerica.loc.gov/batches/batch_pst_fenske_ver02/.

66. An overview of the PaDNP's three phases can be found at https://www.libraries.psu.edu/psul/digipres/panp/padnp.html, while production blogs for Phase I and Phase II can be found at http://www.personal.psu.edu/kkm111/blogs/pa_digital_newspaper_project/ and http://www.personal.psu.edu/kkm111/blogs/padnp2/, respectively. As a side note, I would express some hope that Penn State University Libraries will take care to preserve these resources, which are valuable bibliographic assets for researchers using the CA corpus.

67. http://www.personal.psu.edu/kkm111/blogs/pa_digital_newspaper_project/2009/03/four-titles-to-be-digitized-by-padnp.html.

68. All figures taken from https://www.libraries.psu.edu/psul/digipres/panp/padnp.html.

69. These graphs can be found at http://www.personal.psu.edu/kkm111/blogs/padnp2/digitized-titles.html.

70. This narrative can be found at http://chroniclingamerica.loc.gov/lccn/sn85055199/.

71. Benjamin Fagan, "Chronicling White America," *American Periodicals* 26, no. 1 (2016).

72. http://www.neh.gov/us-newspaper-program.

73. The reports on the "Our Story" website (https://sites.psu.edu/ourstorycentralpausnewspaperproject/) deserve much fuller explication than I can offer here. Its descriptions of archival and bibliographic resolve through constant travel, poor weather, obstinate private dealers, and numerous other obstacles provide a compelling and entertaining account of how historical and other scholarly resources come to be. I hope to treat these accounts in more detail in a future article.

74. "May 1985" report, https://sites.psu.edu/ourstorycentralpausnewspaperproject/reports-1985/.

75. “March 1986” report, https://sites.psu.edu/ourstorycentralpausnewspaperproject/reports-1986/.

76. Library of Congress Catalogs, *Newspapers in Microfilm, United States, 1948–1983*, Volume 2 (Washington: Library of Congress, 1984), 24. For a guide to the abbreviations in these records, see Volume 1 of this same catalog.

77. McGann, “From Text to Work,” paragraph 36.

78. David Smith, Ryan Cordell, and Abby Mullen, “Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers,” *American Literary History* 27, no. 3 (August 2015).

79. Matthew Wilkens, “The Geographic Imagination of Civil War-Era American Fiction,” *American Literary History* 25, no. 4 (2013): 830.