

# EEBO and Us

*Michael Gavin*

## ABSTRACT

*This essay responds to criticisms leveled against “How to Think About EEBO” (Gavin [2017] 2019). While my history of Early English Books Online was faulted for paying insufficient attention to the experiences of scholars using the collection, I argue that using digital tools for literary and bibliographical research can create a false sense of mastery that leaves scholars blind to the features of digital textuality that are most important historically and theoretically.*

---

PETER C. HERMAN’S ESSAY, “EEBO AND ME: AN AUTOBIOGRAPHICAL Response to Michael Gavin”, offers an engaging and welcome rejoinder to my earlier piece, “How to Think About EEBO”.<sup>1</sup> In my article, I describe how the *Early English Books Online* (EEBO) corpus was published. My essay traces this history across a series of remediations, briefly recounting how the surviving record of early print was catalogued, photographed, scanned, and eventually transcribed into machine-readable form. Through these transformations, I argue, EEBO introduced new textual structures to early modern studies that are particularly well suited for quantitative study. My essay paid scant attention to the experiences of scholars who use the collection through its online interface, an exclusion that Herman aims to correct. He argues that I focus on the wrong things and that my emphasis on EEBO’s formal, essentially quantitative structure leads to a mistaken conclusion. What really matters, according to Herman, is not how EEBO works under the hood but how it’s experienced by scholars like himself, for whom the ability to perform quick searches has proven transformative. Whereas my essay emphasized how the EEBO corpus synthesizes textual data with bibliographical metadata in a vast network of collocation, Herman prefers instead to emphasize how scholars are now able to traverse that network through keyword searches to find particular books more easily. In his response, Herman misrepresents and seems to misunderstand some of

---

1. See HERMAN 2020 and GAVIN 2017 [2019].

the finer technical points in my article, but the substance of our disagreement rests on this difference of emphasis. So while I'm grateful for the opportunity to correct misunderstandings, more importantly I'd like to use this occasion to offer some additional context for my piece by reflecting on what I consider the bigger picture.

"How to Think About EEBO" was conceived and written as part of my larger effort to understand the growing role of computation in textual studies, a line of inquiry I have pursued elsewhere through articles that provide theoretical overviews and technical demonstrations of quantitative methods. (I'm one of *those* scholars. . .) However, quantification remains mired in controversy, making progress slow and difficult, and so I hoped, in my study of EEBO, to draw readers' attention to fundamental questions about digital textuality while side-stepping anything that tends to provoke knee-jerk hostility. I also decided to suppress all prior commentary from book historians and others about new media and cybertext, choosing instead to go back to Charles Goldfarb, the inventor of Standard Generalized Markup Language (SGML), whose innovative reconceptualization of textuality continues to astonish for its originality and impact. My hope was that readers would bring fresh eyes to theoretical problems that our predecessors in the 1990s and early 2000s failed to solve.

Indeed, I alluded only briefly to how other scholars interpreted new hypertext formats like SGML. In the 1990s, discussion often centered around the advent of electronic books. In 1994, Sven Birkert's *Gutenberg Elegies: The Fate of Reading in an Electronic Age* eloquently lamented the end of the print era. Many commentators worried that print culture had collapsed and that "the book" faced extinction. Others were motivated to look for historical analogues to the sweeping changes brought by digital technology. Phrases like "from the codex to the computer" and "from Gutenberg to Google" became common tropes as scholars sought to fold new technologies into longer histories of textuality and reading. To understand digital texts and to learn from them was cast as a dialectical process of mutually informing comparison: we can better understand computers by acknowledging continuities in text technologies and reading practices, and we can better understand the histories of those technologies and those practices through comparison to computing. Thus, scholars could study topics like "information overload", "discontinuous reading", and "new media" as subjects in the history of print. Computers themselves could be studied like books as material objects; media archaeology developed a niche following; and Matthew Kirschenbaum's award-winning *Mechanisms: New Media and the Forensic Imagination* (2008) literally put a hard drive under

the microscope to photograph the binary code inscribed on its surface, just to prove that digital texts were as “material” as any other.

For most literary scholars, such forensic work wasn’t necessary for answering the questions raised by cybertext. The best evidence — and, sometimes, the only evidence — was their testimony of their own experience with computers. You didn’t need to know anything about computing to reflect on how reading from a screen felt new and strange. Scholars could assert the privileges of the lay user when confronted by technology while still claiming the authority of expertise. In *Shakespeare and the Book* (2001), David Kastan’s final chapter was actually titled “From Codex to Computer”, but he dismissed and discounted all attention to computing as such. The “actual technology remains unfathomable to most users” (2001, 111), he wrote. However, that very opacity had for Kastan a beneficial side-effect: “the novelty and, for most of us, the daunting unintelligibility of the medium [. . .] conspicuously foregrounds the material make-up of the text” (2001, 116). He continued: “Once we seriously pose the questions the electronic medium has foregrounded for us, we find ourselves at the heart of the issues that animate contemporary textual theory” (2001, 117). Such questions include: “What is the relation between the text’s formal and its material principles of causality? Do texts exist independently of the medium in which they appear?” (2001, 117). To discuss computing “seriously”, according to Kastan, was to proffer answers to questions like these, for which our training and our inherited theoretical tradition were presumptively adequate. By disregarding everything he doesn’t already know and treating the object in front of him as an unfathomable and unintelligible novelty, Kastan could return to more familiar topics about which he felt comfortable reflecting.

A few years later, in his foreword to the collection *Electronic Textual Editing* (2006), G. Thomas Tanselle quibbled with some of Kastan’s conclusions but concurred with the general thrust: “The philosophical conundrum of where a text resides is the same as it has always been”, he assured readers (2006, 6). Computer-mediated text is just the latest installment in a long and continuous history of textuality; our inherited theoretical frameworks are comprehensively sufficient to the task. “What the computer offers [. . .] is a new way of producing and displaying *visible* text” (TANSELLE 2006, 3, emphasis added). Nothing to see here but the things you see here. “Procedures and routines will be different”, Tanselle declared, but “concepts and issues will not” (2006, 6). Our theories have it covered. There’s nothing new we need to learn.

It's a strange notion: *Anything we don't understand about computers can be dismissed as unfathomable, because the only vocabulary we need to speak authoritatively about computing comes from textual theory, and so therefore to know about computers we need only to infer from our experience of using them to read and edit texts.* This presumption has three parts that are closely related but worth keeping distinct. First, that all technical operations invisible to the eye can be disregarded. Second, that textual theory is adequate for identifying what is true and evaluating what is important about computer-mediated discourse. Third, that the experience of reading, interpreting, and editing computer-mediated texts provides an adequate empirical basis for testing any theory.

When stated outright, these assumptions might seem shaky, but in retrospect it is easy to see why scholars like Kastan and Tanselle would find them intuitive and appealing. "The computer is a tool", Tanselle wrote, "and tools are facilitators" (2006, 3). And what would scholars do with these tools? They would use digital collections to compare textual variants, trace references, and search by keyword. Tanselle called this loose cluster of practices "radial reading", and he correctly pointed out that scholars have been reading radially for a very long time. "When people say that the computer makes possible certain kinds of textual research", he writes, "they are using the word possible inexactly to mean 'practically feasible.' But what is feasible is a relative matter" (2006, 4). To confuse what is feasible with what is possible, according to Tanselle, is "symptomatic of the exaggerated claims one often hears about computers" (2006, 4). The hubbub surrounding hypertext was interesting only insofar as it highlighted the features of print technology that always gave the codex its expressive power.

All of this brings us back to Herman. His autobiographical response is welcome because he provides an unusually vivid and detailed testimony of how *Early English Books*, even when still in microfilm, facilitated precisely the kind of radial reading Tanselle describes. Herman describes using the collection for his dissertation, which required searching the short-title catalogues for references, retrieving boxes, threading the film, and scrolling to the needed sources. The process felt laborious but the results were serendipitous: "I got a sense of the wide range of books published in the sixteenth and seventeenth centuries. I saw sermons, poems, plays, cookbooks, more sermons, biblical commentaries, and government announcements. In addition, while looking for attacks on poetry, I was schooled in the extraordinary range of the early modern book trade, including the extraordinary range of publishing styles and fonts" (2020, 209). This research method

was still hindered, though, by the difficulty of working with the machines. Herman highlights “four major advantages that EEBO has over the microfilms”:

First, the database is now searchable (e.g., title, author, subject, printer, year). Second, we can access the files within seconds, as opposed to waiting for someone to deliver the microfilm box, or pawing through a giant file cabinet to find it yourself, and then scrolling through until you finally reach the book you want. Third, the files are all downloadable, so now, we can develop our own library of primary sources. Finally, *you can do all of this from your desktop at home! You don't have to be in a library!*

(2020, 210, emphasis original)

Whereas Tanselle stood pat on a theoretical nicety about the conditions of possibility — a nicety true on its own terms but irrelevant on any other — Herman adopts a far more sensible outlook. By changing what is practically feasible, he argues, EEBO changed how early modern scholarship was practiced: “By making access to early modern books much easier, EEBO allows for greater and greater historical grounding in our scholarship. So much so that EEBO changed the protocols of peer review: people are now *expected* to use EEBO. Relying on contemporary editions, let alone snippets of quotations from other works, is no longer sufficient” (2020, 211–12, emphasis in the original).

On all of these points, Herman and I agree. In fact, I had a very similar experience with the microfilm and with EEBO as a student (as did, I'm sure, many other people), and in writing its history my goal was to describe the information architecture that made our shared experiences possible. Although Herman writes as if we disagree on these points,<sup>2</sup> I believe our

---

2. A key point of misunderstanding seems to hinge on my use of the word “abstraction”. Herman argues that “while Gavin is surely right when he claims that the short-title catalogues offered ‘a compilation of metadata already powerfully abstracted from the paper, cardboard and leather on the shelves’, my experience is the opposite [. . .]. [R]ather than abstraction, reading the STC microfilms gave me a vastly more concrete sense of my topic” (2020, 209). Of course, I did not mean to suggest that EEBO's data structures direct attention away from the source texts, nor that they somehow obviate the particularity of individual books. Nothing about the TCP files invalidates one's subjective experience of textual concreteness. Nor does my reference to Charles Goldfarb's phrase, the “death of the document”, imply as much. Quite the contrary. The abstractions of the metadata are precisely what make possible the experience of each book's

views have much in common, and at times his narrative echoes mine quite closely.<sup>3</sup> Where Herman claims “EEBO and Me”, I see EEBO and us.

However, we do disagree on two points. One is a relatively small matter of definition. Herman writes, “Just to be clear, EEBO stands for ‘Early English Books Online’, and the database consists of pdf files of the STC microfilms” (2020, 210). Yet, as I tried to explain, there’s a lot more to EEBO than the files that scholars download from the website. Herman mistakes EEBO’s most visible parts for its actual whole. By defining the resource as a collection of PDFs that just happen to be “searchable”, Herman dismisses

---

historical specificity, per se. Abstraction is not contrary to particularity. Indeed, metadata facilitate the *thisness* Herman refers to as “*this* phrase from *this* newspaper published on *this* date” (2020, 214, emphases in the original).

3. For example, I write that EEBO provides a “surrogate” for printed books; he calls it a “substitute” (2020, 209). He says that *Early Modern Books* gave him a “sense of the wide range of books published in the sixteenth and seventeenth centuries”; I argue that the catalogue and collection were most valuable for “giving scholars a sense of ‘what there was’” in the archive. Herman argues that “EEBO has altered scholarship by allowing easy and nearly immediate access to early modern books” (2020, 210); I argue that it “provided a record system that made library holdings visible and therefore accessible” ([2017] 2019, 87). I must confess I find Herman’s essay quite baffling on this score, and I struggle to resist the temptation to believe that he simply didn’t read my piece with much care. For example, Herman writes, “In his article, Gavin poo-poops exactly why EEBO has made such a difference in scholarship: ‘If information technology just winds up in your hands as a printed book — if we have merely ‘gone full circle’ to where we started — something hasn’t gone right.’ But in my experience, something has gone exactly right when this happens. Information technology has put into our hands, and on our desktop or laptop screens, the collected holdings of the Bodleian, the Folger, the Huntington, and the Newberry libraries (to choose but a few). That is not small, and it doesn’t deserve to be denigrated” (2020, 212). In the passage Herman quotes, I am describing the print-on-demand service imagined by Eugene Power (1958) and now offered by ProQuest. When Herman says that EEBO has “put into our hands [ . . . ] the collected holdings of the Bodleian, the Folger, the Huntington, and the Newberry”, he lapses into metaphor. EEBO couldn’t do such a thing, nor would anyone want it to. Except, perhaps, for Power. Power actually thought his greatest achievement would be realized when scholars adopted the habit of buying print-on-demand paper copies of microfilmed books. Power held this idea in such esteem that he actually titled his autobiography *Edition of One*. Herman agrees with me that this scheme misunderstands EEBO’s true value, but in his eagerness to take offense he fails to notice our agreement and accuses me of denigrating features of EEBO we both prize.

as mere functionality the elements of EEBO that I tried to call attention to: the metadata and the TCP transcriptions. It's true that good metadata and clean transcriptions make bibliographical and keyword search possible, but my goal was to draw attention to them on something like their own terms. The metadata and the transcriptions bind the collection together as a coherent whole. Within the EEBO-TCP corpus, no single document exists on its own. Each is part of a larger body. Bibliographical metadata (like date of publication, author, etc.) organize the corpus into sets of documents. Lexical metadata (including both TEI tags and, more simply, word forms) identify paradigmatic relations across those sets. Every element contained in every EEBO-TCP document exists within this self-contained system of relations. That's what makes keyword search possible. But it's also what makes statistical analysis possible, and so anyone hoping to study the corpus quantitatively must carefully attend to this structure.

Here, then, is the second point of disagreement. Herman dismisses this view as hogwash: "With more than a touch of techno-utopianism, Gavin proposes that with EEBO-TCP, we have finally reached the promised land where 'everything [. . .] is connected to everything else'" (2020, 208). But there is nothing techno-utopian about my observation; it's a straightforward description of the EEBO-TCP corpus, or any corpus for that matter. However, because scholars like Herman (and like Kastan before him) know digital texts only through their experience as lay users, and because they think of computing (like Tanselle) as a mere tool for displaying and searching visible text, direct attention to the corpus seems counter to (rather than simply different from) the experience of reading. And because they take for granted that their experience is adequate to a proper and full understanding, such scholars dismiss as mere function precisely those features of digital textuality that are most important historically and theoretically. This dismissiveness enables another: claims about the historical importance of textual computing can be dismissed as exaggerations, and the people making those claims can be dismissed as naïve or utopian.

The ability of the computer to abstract the attributes and contents of files — what users like Herman experience as the ability to *search* — is not just a convenience, nor is it some abstruse magic. It reflects the fundamental condition of discourse when recorded for computation. As I have argued elsewhere, "keyword searching" and "quantitative textual analysis" are much more closely related than most scholars realize.<sup>4</sup> Their histories

---

4. For a brief overview of the intellectual history, see GAVIN 2018; for a more detailed introduction to the mathematical concepts and their relation to questions of textuality, see GAVIN 2020.

are intertwined in the information sciences, and they share a common theoretical foundation grounded in mathematics. As book historians and literary scholars, nothing in our bibliographical training prepares us to study these topics in an intellectually responsible way; nothing in a typical theory seminar introduces students to the concepts they would need to do so.

When confronted by an object like EEBO, it is tempting to focus on familiar problems that raise familiar questions. What's the difference between a book and an image? What aspects of the material text *can't* be digitized? What is it like to read on a screen? Is the collection representative of the surviving print record? What has been preserved and what has been lost at each stage from publication to digitization? What about reprints and variants? What about manuscript sources, material culture, and oral transmission? Haven't they been left out? Such questions are valid and important, but we have been asking them for a long time. They aren't enough. They miss a more important point.

In the twentieth century, a profound discovery was made in the information sciences. Namely, that documents could be identified and classified using simple statistical measures of the words they contain. This might seem like a technical accomplishment, a feat of engineering. And so it was. But it did something more. It revealed a fundamental commensurability between the mathematical structures of computationally recorded discourse and the meaningful structures of language as used and intended by persons. On the basis of this discovery, textuality has been totally transformed. Computation has infiltrated everything from how governments track dissidents to how people have sex. Information scientists were able to do this because they know things about language we don't know. I find that exciting, because it means there is a lot for us to learn. I wrote that history of EEBO to share what the collection looks like through a different set of eyes. I see it, not only as a remediation of rare books, but also as an opportunity to investigate these newly discovered principles of language and textuality.

Another commenter has said that my history "presumes a bibliographical knowledge that many humanities students do not possess" (GREGG 2021, 36). But, as should be clear, that is not my argument at all. Instead, I ask: *What if bibliographical knowledge is a trap?* When I look back on my training in book history, I believe that Kastan's, Tanselle's, and Herman's assumptions were everywhere operative. Those assumptions came wrapped in a story with a clear moral. I was trained to take for granted that rigorous, historically aware scholarship provided a needed corrective to the exaggerated claims of techno-enthusiasts, determinists, utopians, futurists, and neoliberal hacks. Unlike them, we're careful. We historicize. We appreciate

complexity and nuance. We know not to fall for their hocus-pocus act, for their TED talks. We're the good guys.

It was a great story, told with vigor, and true in many respects. It motivated a huge outpouring of excellent scholarship, much of which was foundational to my graduate training in the early 2000s. I remain indebted and grateful to its tellers, some of whom I cherished as teachers. But what if that story's moral framing also distracted us from other things? What if bibliographical knowledge has left our discipline stuck always asking the same kinds of questions no matter our object of study? How would we know?

*University of South Carolina*

## Works Cited

- GAVIN, Michael. 2017 [2019]. "How to Think about EEBO". *Textual Cultures* 11.1–2: 70–105.
- . 2018. "Vector Semantics, William Empson, and the Study of Ambiguity", *Critical Inquiry* 44. 4: 641–73.
- . 2020. "Is There a Text in My Data? (Part 1): On Counting Words". *Cultural Analytics* @ <https://doi.org/10.22148/001c.11830>.
- GREGG, Stephen H. 2021. *Old Books and Digital Publishing: Eighteenth-Century Collections Online*. Cambridge: Cambridge University Press.
- HERMAN, Peter C. 2020. "EEBO and Me: An Autobiographical Response to Michael Gavin, 'How to Think About EEBO'". *Textual Cultures* 13.1: 207–16.
- KASTAN, David Scott. 2001. *Shakespeare and the Book*. Cambridge: Cambridge University Press.
- POWER, Eugene. 1958. "O-P Books, A Library Breakthrough". *American Documentation* 9.4: 273–76.
- , with Robert ANDERSON. 1990. *Edition of One: The Autobiography of Eugene B. Power*. Ann Arbor: University Microfilms International.
- TANSELLE, G. Thomas. 2006. "Foreword". In *Electronic Textual Editing*, edited by Lou BURNARD, Katherine O'BRIEN O'KEEFFE, and John UNSWORTH. New York: Modern Language Association.