Early English Books Online: Mass Digitization and the Archive

Early English Books Online: Mass Digitization and the Archive

Conner Wilson

Abstract

This review examines the originations and contemporary usage of the online archive Early English Books Online (EEBO). Highlighting the recent advancements in digital historiography, alongside considerations of inherent archival bias, this article demonstrates a variety of circumstances in which the scholar is encouraged to look beyond the digital archive itself. EEBO here is proposed as a resource capable of profound innovation, one of preservationist historical necessity, and a logical further extension of scholarship dating all the way back to the early twentieth century and the Short-Title Catalogue. Yet also EEBO is a resource of human construction, and therefore must be approached with the same considerations one would the physical archive, giving careful thought to the intersection of material and print culture, and the ways in which they correlate.

Biography

Conner Wilson is a postgraduate student at the University of Birmingham studying Shakespeare, his contemporaries, and Early Modern theatre culture.

Over the course of the past two decades, the mass digitization of the archive has radically transformed the breadth of primary source material readily available to the modern scholar. Online archives such as *Early English Books Online* (*EEBO*), *Eighteenth Century Collections Online* (*ECCO*), *Manuscript Pamphleteering in Early Stuart England* (*MPESE*), and *the Old Bailey Proceedings Online*, along with numerous others, have become inundated with modern methodological approaches to historiography, with most, if not all, Masters and PhD programs requiring some compulsory module towards navigating these resources. On one hand, this "revolution"[1] of digitization as Tim Hitchcock describes it, represents a turning point for historians, as researchers embrace the advantages of immediacy and accessibility in the information age; yet, in a field where visual, material, and print culture so often coincide, how do we determine the accuracy in which these online archival substitutions can produce the unique phenomenological experience associated with resource tangibility? Or, for instance, how do researchers overcome the implicit bias of search bar algorithms in tandem with imperfect and outdated Optical Character Recognition (OCR)? While much of what has been written about *EEBO* tends to exist in a binary dialectic of good vs. bad, helpful vs. unhelpful, accurate vs. inaccurate, this article will aim to circumnavigate such finite categorizations, and acknowledge both the trepidations of scholars who fear misuse, and embrace the growing computational literacy of historical fields. This reciprocal analysis, alongside a detailed historical account of the creation of the database, presents *EEBO* as not too dissimilar to the physical archive: proposing that, with both the digital and the material, it is ultimately the historian's job to determine relevancy and overcome inherent bias.

EEBO's Beginnings

---

[1] T. Hitchcock, 'Confronting the Digital', *Cultural and Social History*, 10/ 1 (2013), p. 9.

The origins of *EEBO* can be traced all the way back to the early twentieth century. In 1918 on commission from The Bibliographical Society, scholars A. W. Pollard and G. R. Redgrave began the monumental task of creating a unified catalogue covering all extant books, printed between 1475 and 1640, across Great Britain and North America. It was a project which would take nearly 8 years of research and require an immense amount of interlibrary cooperation, however, by 1926 Pollard and Redgrave's work: *A Short-Title Catalogue of Books Printed in England, Scotland, & Ireland and of English Books Printed Abroad, 1475–1640*, was finally ready for publication.[22] This Short-Title Catalogue or *STC*, as it is frequently abbreviated, immediately proved to be an invaluable road map for scholars in the sourcing of rare and out-of-print books. The *STC* covered the holdings of a myriad of libraries, provided bibliographic information on nearly 26,000 extant texts, and managed a scope of information which was unprecedented. The scholars had successfully proved that the cross unification of resources and information was possible on a massive scale, so far as researchers were willing to acknowledge that it was "dangerous work for any one to handle lazily."[3] This cautionary caveat would come to permeate historical research well into the information age. Considering now the contemporary trepidations around *EEBO*, is it fitting here to include Pollard and Redgrave's initial caution of the *STC* that "in so large a work based on such varied sources, probably every kind of error will be found represented."[4] Perhaps, historians have always been cautiously self-aware of the dangers of mass bibliographic consolidation and the seductive illusion of an entirely comprehensive historical archive. Yet, despite this, the pairs' work has unequivocally become one of the most influential and enduring enterprises towards the sourcing of Early Modern texts. Fourteen years later, with the danger of WWII fast approaching, and the advent of a new technological system, Microfilm, the American Council of Learned Societies felt the processing and photographing of Early English vulnerable

---

[2] M. Gavin, 'How to Think about EEBO', *Textual Cultures,* 11 / ½ (2017), pp. 70-102.

[3] B. Nagle, 'Introduction', in D. Wing (ed.), *Short-title catalogue of books printed in England, Scotland, Ireland, Wales, and British America and of English books printed in other countries, 1641-1700* (New York, 1945) p. 10.

[4] Nagle, 'Introduction', 10.

texts was a project which could not be delayed, and the Short-Title Catalogue should become the bedrock from which the selection committee would work. Six million pages were prioritized for this microfilmic reproduction process with an ultimate objective of storing the facsimiles securely in America, farther from the increasingly volatile Western Front.[5] This decision to integrate the microphotographic imaging process with the *STC* would serve as the basis for what is now *EEBO*, with many of the original images captured by this commission populating the contemporary database today. It is imperative to understand that while much work has been done since the original publication of the *STC* (notably Donald Wing's subsequent, yet separate, catalogue expanding the breadth of titles from 1641 to 1700)[6] and on microfilm reproductions themselves (with *STC* titles continuing to be photographed well into the 1990s) the digital visual make-up of *EEBO* began nearly 40 years before the advent of the internet; suffice it to say, the microphotographic process was not designed with considerations towards its ultimate digital transference. *EEBO*, as we know it now, would finally come into existence with the birth of the Text Creation Partnership (TCP) in 1999. This interlibrary effort to "create texts to a common standard suitable for search, display, navigation, and reuse" is the process on which the second half of this article will focus more specifically, as it has come to define the contemporary successes and pitfalls of the database.[7]

OCR, Comprehensive Digital Archives, and Material Culture

Perhaps the most extraordinary feat the *EEBO-TCP* partnership has undertaken, is its avoidance of common OCR problems, by abandoning the technology altogether. The implementation of a "double-keyed"[8] transcription system, with human editors

---

[5] Gavin, 'How to Think about EEBO', pp. 70-102.

[6] I. Gadd, "The Use and Misuse of Early English Books Online", *Literature Compass,* 6/3 (2009): p. 683.

[7] 'Text Creation Partnership', < https://textcreationpartnership.org/>, accessed 31.3. 2021.

[8] J. Heil and T. Samuelson, 'Book History in the Early Modern OCR Project or, Bringing Balance to the Force', *Journal for Early Modern Cultural Studies,* 13 / 4 (2013), pp. 93-94.

coding from the original microfilm images, boasts a "99.995%"[9] accuracy rating per-text-entered, thereby enabling the current sophistication level offered in the simple and advanced search bar functions. While immensely expensive and labor intensive, this effort ensures a consistent accuracy which has previously proven difficult in Early Modern typeface transcriptions, yet it does, however, simultaneously shatter the illusion of an entirely comprehensive archive. As Ian Gadd notes, "*EEBO* does not include every copy of every edition published prior to 1701… nor even does it include a copy of every *surviving* edition published prior to 1701."[10] This is an important distinction in that the textual variances between subsequent editions of Early Modern books can prove to be drastic. One need look no further than Quarto 1 and Quarto 2 of *Hamlet* (both available on *EEBO*), to detect a noticeably alternate print of the infamous, "To be, or not to be, that is the question" (*Tragedy of Hamlet* 23)[11] which instead reads, "To be, or not to be, I there's the point." (*Tragicall Historie of Hamlet* 15)[12] Fortunately for Shakespeare, the infamy of his work secures an archival placeholder for the various editions of his plays, however, it is near impossible to discern a similar degree of canonical entirety for the multitudes of lesser-known authors present on *EEBO*. If a scholar either unknowingly or willfully ignores this fact, the dangers of misrepresentation, false negatives, and false positives are relatively high. Additionally, given that *EEBO* is computationally manual, the quick inclusion of a subsequent textual edition is seemingly non-existent, and the notion that *EEBO* could be entirely comprehensive rapidly falls away simply considering the sheer labor intensity of the archive, which is tremendous. This is not to suggest either that *EEBO* advertises itself as a comprehensive archive (as neither did Redgrave and Pollard consider their work entirely comprehensive), but instead give caution to the scholars who may be first

---

[9] 'The results of keying instead of OCR', <https://textcreationpartnership.org/using-tcp-content/results-of-keying/>, accessed 31.3.2021.
[10] I. Gadd, 'The Use and Misuse of Early English Books Online'*, Literature Compass,* 6/3 (2009), p. 686. (Italics mine).
[11] W. Shakespeare, *The Tragedy of Hamlet Prince of Denmarke*, Printed by George Eld for Iohn Smethwicke, and are to be sold at his shoppe in Saint Dunstons Church yeard in Fleetstreet. Vnder the Diall, (London, 1611), p. 23.
[12] W. Shakespeare, *The Tragicall Historie of Hamlet, Prince of Denmarke*, Printed [by Valentine Simmes] for N[icholas] L[ing] and Iohn Trundell, (London, 1603), p. 15.

using the resource. One would not assume a physical library could possibly contain every text on a single subject and the same principle must be applied to the digital.

Regarding the physical tangibility of primary source material, *EEBO* presents both clear advantages and disadvantages. Considering the preservationist origins of the online archive, the sheer volume of scholars who now have access to the texts without having to physically handle the pages is an immense victory for the longevity of Early Modern books. The reality that pages are turned less frequently, less exposed to light, able to maintain a consistent temperature, and are simply less prone to accidental human contamination, will keep these resources accessible to those who need them for many years to come.[13] Healthy shelf life in correlation with the Early Modern book was already a precarious relationship, and the digital archive aids in keeping these texts in the hands of those most qualified to handle their longevity. On the other hand, *EEBO* all but abandons the material culture of the printed book, as the researcher is, of course, not actually manipulating the original artifact. Books on *EEBO* all appear to be roughly the same size and dimensions, which is simply not the case.[14] Furthermore, microfilm does little to aid in the capturing of handwritten notes of previous owners, thereby potentially overlooking additional valuable historiographical evidence. For example, many of the digital reproductions of seminal works now existent on the internet today, do little to account for things such as transportability or mobility of the original object. An Early Modern book capable of fitting in its owner's pocket carries significantly different cultural weight than one which sits on the lectern of a library or lecture hall. Expanding on this work, the researcher may begin to unlock information such as the author's contemporary popularity or their cordiality with publishing companies. A book existent in multiple different contemporaneous languages may reveal an author's audience reach, their financial stability, or the sociological circle of which they were a member, all of which in turn can affect

---

[13] R. de But, 'Managing Risks: what are the agents of deterioration', <https://artsandculture.google.com/exhibit/managing-risks what-are-the-agents-of-deterioration-trinity-college-dublin-library/PQKyBVnbqWmqLw?hl=en>, accessed 8.4.2021.

[14] I. Gadd, 'The Use and Misuse of Early English Books Online*', Literature Compass,* 6/3 (2009), p. 682.

literary analysis.

While some of this information may be discerned from *EEBO*, the researcher must continue to be diligent and thorough with historiographical information beyond the text itself. This ultimately asks the important question at the intersection of material and print culture: can we consider the text of primary source material in a vacuum, or does removing the physical life of the object detract vital information which in turn can affect textual analysis? The answer to this, of course, depends on the author, the text, the book itself, the contemporary and historical associations of the material, the previous owner(s), the type of research being conducted, and a litany of other potential factors, yet still, the contemporary historian must not be swayed into
ignoring the material world which exists behind the digital reproduction, as there is certainly valuable information existent there.

Conclusions

In 2001 John Jowett and Gabriel Egan authored one of the earliest reviews of *EEBO*, writing that "the potential for generating new research in early modern studies is considerable indeed… and electronic products such as *EEBO*… enable new forms of scholarly study which were not possible using paper and film technologies."[16] Two decades later, this observation still holds true. *EEBO* has provided massive amounts of information to scholars over the years, ushering in exciting and new historical discoveries, which otherwise may have gone unrealized, overlooked, or been significantly delayed. Moreover, the access which current students now have to primary source material is unprecedented and evolving the very fabric of how academic arguments are conducted. In the midst of this exciting growth, it is vital for the researcher to remember that they must not rely solely on what is convenient. All archives, whether digital or physical, are ultimately human constructions and therefore contain certain limitations and biases both consciously and unconsciously. The above examples outlined are merely a few of the considerations scholars should take into account when conducting online research.

As always, the historian shoulders the burden of accuracy, thoroughness, and overcoming bias, but when used diligently, the potentialities of *EEBO* are immense.

Bibliography

De But, R., 'Managing Risks: what are the agents of deterioration',
<https://artsandculture.
google.com/exhibit/managing-riskswhat-are-the-agents-of-deterioration-trini
ty-college-dublin library/PQKyBVnbqWmqLw?hl=en
https://artsandculture.google.com/exhibit/managing-riskswhat
are-the-agents-of-deterioration-trinity-college-dublin-library/PQKyBVnbqWm
qLw?hl=en>, accessed 8.4.2021.

Gadd, I., 'The Use and Misuse of Early English Books Online', *Literature Compass,*
6/3 (2009), p. 680- 692.

Gavin, M. 'How to Think about EEBO', *Textual Cultures,* 11 / ½ (2017), pp. 70-102.

Heil, J. and Samuelson, T., 'Book History in the Early Modern OCR Project or,
Bringing Balance to the Force', *Journal for Early Modern Cultural Studies,* 13 / 4
(2013), pp. 93-94.

Hitchcock, T., 'Confronting the Digital', *Cultural and Social History*, 10/ 1 (2013), p.
9-23.

Jowett, J. and Egan, G., 'Review of the Early English Books Online (EEBO)',
Interactive Early Modern Literary Studies (2001), pp. 1-13.

Nagle, B. 'Introduction', in Wing, D. (ed.), *Short-title catalogue of books printed in*
England, Scotland, Ireland, Wales, and British America and of English books
printed in other countries, 1641-1700 (New York, 1945) p. 10.

Shakespeare, W., *The Tragedy of Hamlet Prince of Denmarke*, Printed by George
Eld for Iohn Smethwicke, and are to be sold at his shoppe in Saint Dunstons
Church yeard in Fleetstreet. Vnder the Diall, (London, 1611).

Shakespeare, W., *The Tragicall Historie of Hamlet, Prince of Denmarke*, Printed

[by Valentine Simmes] for N[icholas] L[ing] and Iohn Trundell, (London, 1603).

'Text Creation Partnership', <https://textcreationpartnership.org/>, accessed 31.3. 2021.

'The results of keying instead of OCR', <https://textcreationpartnership.org/using-tcp content/results-of-keying/>, accessed 31.3.2021.