How to Think about EEBO

Michael Gavin

Abstract

This essay narrates a brief history of the Early English Books Online (EEBO) corpus, describing a convergence between bibliography, document imaging, and information science. Whereas literary scholars and book historians tend to express skepticism about EEBO and emphasize its limitations, this essay offers a revisionist and more optimistic perspective in hopes of clearing the theoretical ground for quantitative research.

Introduction

When the Text Creation Partnership (TCP) publicly released its "Phase I" documents of the *Early English Books Online* (EEBO) collection in early 2015, the files posted to Github represented the culmination of decades of scholarly labor. By "decades", I refer not only to the now twenty-year history of the TCP itself, which began at the University of Michigan in 1999, but more fundamentally to the work that preceded the project and made it possible, some of which stretches back to the early twentieth century. EEBO sits at the intersection of several lines of disciplinary development and technological innovation and is one result of generations of work in bibliography, document imaging, and information science.

My purpose in this essay is take a brief walk along the paths of this winding history, with an eye toward the future. I'll focus on the creation of the Short-Title Catalogue and the microfilm collection it inspired, as well as on the invention of text-markup language, which structures the recently released EEBO documents. In particular, the EEBO-TCP sits at the nexus of three intellectual and institutional developments: 1) the creation of historically comprehensive bibliographies, 2) the remediation of large rare-book collections into microfilm, and 3) the conversion of that film into rigorously marked up transcriptions. These pieces of the project ("project", here, most broadly conceived) share several important assumptions and aspirations. Most importantly, they assume that knowledge about books

Textual Cultures 11.1–2 (2017 [2019]): 70–105. DOI: 10.14434/textual.v11i1-2.23570

can be held by proxy. Whether through bibliographies or digital interfaces, books are abstract objects with representable attributes, and information about those attributes can be regularized into forms that enable comparison across large collections. They also jointly aspire to limited comprehensiveness and qualified transparency. Although by definition the information held in catalogues and digital archives is incomplete and prone to error, nonetheless its compilers hoped to communicate something like the totality of surviving English print.

The short-title catalogues provided a panoramic view into libraries across the world by creating a searchable surrogate for rare-book archives, and the microfilm and digital transformations that followed winnowed the scope of that view down to the pages of the books and, eventually, to their very words. The transparency and comprehensiveness provided by this surrogate is, of course, not *real*, but neither is it an illusion. The EEBO-TCP is more like a simulation or model of extant print. Bibliographic metadata fold documents into history, connecting them to libraries and authors, to booksellers, printers, and patrons. Descriptive markup in turn teases out the formal structure of those documents, identifying texts' individual parts and the relationships among them. The result is a vast and sophisticated model of historicity, textuality, and sociality.

Many elements of this history will be known to readers in the fields of book history and digital humanities, but the themes I'll touch on and the perspective I'll offer are not typical of discussion surrounding EEBO, in particular, or digital archives more generally. My primary intended audience includes scholars who are currently engaged in quantitative analyses of the EEBO corpus, or those who might be considering such work. Since 2015, my own research has been completely dependent on this collection, and I'm not alone. Now that we have tens of thousands of early modern documents available for computational analysis, it's worth pausing to ask what interpretive demands this collection poses, and the best way I know of to do this is to review the history of its creation. However, existing surveys tend to be written from the perspective of literary scholars or book historians, and so tend to focus on what was lost in the digitization process. Instead, I want to offer a more sympathetic way of thinking about the long history of bibliography and digitization, a way of thinking that opens the collection up as an object of study in its own right. Now that we have EEBO, what do we do with it? But before we can ask that question, we need to know where the files came from, what theories informed their creation, what features they have, and how they came be what they are.

Reading Machines: Bibliography, Microphotography, and the Simulated Archive

Beginning in the early twentieth century, bibliographers like A. W. Pollard, G. R. Redgrave, and Donald Wing compiled unified catalogues of rare books in British and North American libraries, providing an unprecedented level of transparency to the total archive of early English print. At the much same time, Eugene Power, founder of University Microfilms International (now ProQuest), was developing and popularizing a film-based technique for preserving newspapers and out-of-print books. Inspired by the information futurist Robert C. Binkley, Power hoped that new imaging technology could preserve cultural history while increasing public access to archival materials. In the late 1930s, as war loomed over Great Britain, Power received grant funding to photograph thousands of books deemed to have research value, thus providing the foundation and impetus for the *Early English Books* microfilm collection. After 1998, when the page images were online and the supporting bibliography was made available as a searchable database, demand for full-text search inspired the Text Creation Partnership, which formed the next year and quickly began its first phase of transcription.

The Short-Title Catalogue was first and primarily conceived as a "finding-list" for scholars who were then expected to consult paper copies of rare books in libraries across Britain and North America. Work on the initial catalogue took 8 years to complete, and was published in 1926 by The Bibliographical Society as *A Short-Title Catalogue of Books Printed in England, Scotland, & Ireland and of English Books Printed Abroad, 1475–1640*. The collaborative and distributed nature of their work can be glimpsed in the book's subtitle, which announces that it had been "Compiled by A. W. Pollard & G. R. Redgrave, with the help of G. F. Barwick, Geo. Watson Cole, Ethel Fegan, F. S. Ferguson, W. W. Greg, W. Jaggard, Stephen K. Jones, F. R. D. Needham, H. R. Plomer, Cecie Stainer, E. V. Stocks and others". From the first, the editors were careful to warn their readers that the catalogue did not represent, in any absolute nor even tentative way, a full record of surviving English print. They describe the project instead as a record of their own activities, as "a catalogue of the books of which its compilers have been able to locate copies, not a bibliography of books known or believed to have been produced".[1] Although they express the hope that

1. A. W. Pollard and G. R. Redgrave, *A Short-Title Catalogue of Books Printed in England, Scotland, & Ireland* (London: The Bibliographical Society, 1926

they will be found to have included over 90% of extant titles and 80% of extant editions and issues, Pollard nonetheless reminds readers that "in so large a work based on such varied sources, probably every kind of error will be found represented, and those who use the book as anything more than a finding-list must be on their guard".[2]

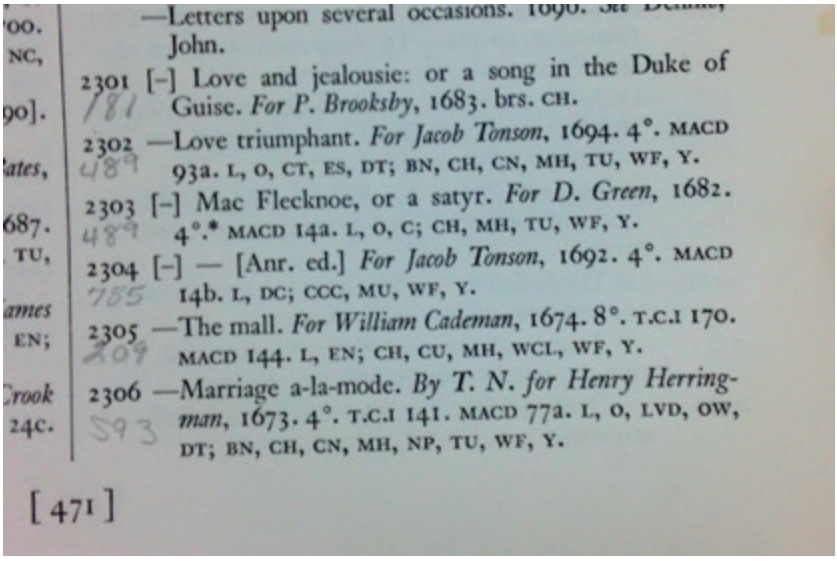
Donald G. Wing continued this work more or less single-handedly in his catalogue, which was published after the war between 1945 and 1951 (revised 1972 to 1998), and which extended the scope of the bibliography from 1641 to 1700. Like his predecessors, Wing admonished his readers to approach the materials with care. A curious feature of these bibliographer's prefatory remarks is a tension between their professed desire for comprehensiveness and their worry that scholars might take their catalogues too much at their word. *Caveat lector!*, they warn. Especially on the matter of his catalogue's comprehensiveness, Wing is careful to caution readers, "Because a library is included it does not follow that all that library's holdings are listed. *This is not a census of copies*, but only a guide to inform scholars where a given entry may most conveniently be consulted".[3] The emphasis is original. On the exact same page, Wing adds, "I should repeat here the warning that this is not a census of copies, but rather an effort to locate copies available in various geographic regions". In the General Introduction, editorial committee chairman Benjamin Nangle reaffirms the point, in case it wasn't clear: "The user must always bear in mind that it is a *short-title* catalogue . . . not a census of copies, but rather an effort to locate copies in various geographical areas and thus to inform the scholar where he can conveniently consult a copy".[4] Their ambition was to provide a synoptic view into archives around the world, but they insisted that it must not be mistaken as a representative survey of surviving print. They worried that, because their short-title catalogues under-represent the number of actually extant copies, unsuspecting librarians might be hustled into paying higher prices by unscrupulous book dealers. For scholars, they took for granted that information in the catalogue must never be substituted for direct consultation of library copies and comparisons among them.

[reprint 1946]), xiii. Cited in Mak, "Archeology of a Digitization", 1518.

2. Pollard and Redgrave, *Short-Title Catalogue*, xvi.
3. Wing, *Short-Title Catalogue of Books Printed in England, Scotland, Ireland, Wales, and British America and of English Books Printed in Other Countries, 1641–1700*. Printed for the Index Society (New York: Columbia University Press, 1945), ix, emphasis original.
4. Benjamin Nangle, "General Introduction", in Wing, *Short-Title Catalogue*, v, emphasis original.

Perhaps oddly, the bibliographic catalogues that form the basis for so much work in English literary history have from their inception been enshrouded with a curious rhetoric of their own disavowal. Pollard, Redgrave, and Wing all openly worried about the potentially deleterious effects their work could have on their unsuspecting peers. Expressions of this concern remain popular among book historians and other critics, for whom the myth of the uncritical scholar retains powerful appeal, though nowadays it's usually associated with the online version. Recent essays by Stephen Tabor, Ian Gadd, Bonnie Mak, and, most recently, Kathryn Sutherland and Marilyn Deegan have warned about how inattention to the history of the STC threatens future work. Echoing concerns voiced eighty years before by Pollard, Gadd cautions that "students and scholars often tacitly, but wrongly, assume that EEBO represents the printed record in its entirety".[5] Mak avers that readers are "encouraged to overlook as inconsequential the material history" of the archives.[6] Sutherland and Deegan insist that "a digital copy of a print copy is never more than a partial copy", while worrying that "most of us, seasoned scholars and students new to historical research, are blind to their inadequacies".[7] For Tabor, this lack of critical self-awareness manifests as a generational divide: "The younger generation of scholars in particular, lured by full-text images and ransacking the Web for illustrations for their books and articles, are using these utilities as de facto bibliographic databases".[8] *Kids today . . . with their Snapchat, their Tinder, and their EEBO!*

5. Ian Gadd, "The Use and Misuse of Early English Books Online", *Literature Compass* 6, 3 (2009): 680–92. Gadd continues, "EEBO is obviously aiming to provide a useful scholarly mechanism in terms of searching but by doing so are implying — albeit not deliberately — that the record and the copy are one and the same thing" (687). Such comments are not only wrong, they border on slander. Nowhere do the editors of EEBO make so patently false a claim, nor is it ever explained how they could possibly do so by implication. In this comment, Gadd deploys a strategy typical of EEBO's critics, which is to impute their own (presumptively superseded) naiveté onto the project they pretend to critique.
6. Bonnie Mak, "Archaeology of a Digitization", *Journal of the Association for Information Science and Technology* 65, 8 (2014): 1519.
7. Kathryn Sutherland and Marilyn Deegan, *Transferred Illusions: Digital Technology and the Forms of Print* (New York: Routledge, 2016), 133.
8. Stephen Tabor, "ESTC and the Bibliographic Community", *Library* 8, 4 (December 2007): 368.


—Letters upon several occasions. 1690. Sir [illegible] John.

2301 [–] Love and jealousie: or a song in the Duke of Guise. *For P. Brooksby*, 1683. brs. CH.

2302 —Love triumphant. *For Jacob Tonson*, 1694. 4°. MACD 93a. L, O, CT, ES, DT; BN, CH, CN, MH, TU, WF, Y.

2303 [–] Mac Flecknoe, or a satyr. *For D. Green*, 1682. 4°.* MACD 14a. L, O, C; CH, MH, TU, WF, Y.

2304 [–] — [Anr. ed.] *For Jacob Tonson*, 1692. 4°. MACD 14b. L, DC; CCC, MU, WF, Y.

2305 —The mall. *For William Cademan*, 1674. 8°. T.C.I 170. MACD 144. L, EN; CH, CU, MH, WCL, WF, Y.

2306 —Marriage a-la-mode. *By T. N. for Henry Herringman*, 1673. 4°. T.C.I 141. MACD 77a. L, O, LVD, OW, DT; BN, CH, CN, MH, NP, TU, WF, Y.

[471]

Figure 1. Wing's *Short-Title Catalogue* (1945). Notice the marginalia. In this copy, owned by the University of South Carolina, entries have been manually cross-referenced with each item's *Early English Books* microfilm reel number, leaving a physical trace of the indexing function Wing's catalogue would perform once it was combined with a searchable collection of images.

The irony here is that these latter-day skeptics impute onto the digital interface an act of abstraction that was already inherent to the catalogues in the first place. Unified bibliographies like the Short-Title Catalogue offered scholars a new and radically different way to experience library archives: not as a collection of rare books but as a compilation of metadata already powerfully abstracted from the paper, cardboard, and leather on the shelves. As Wing and Nangle insisted, short-title catalogues gather disparate information from multiple sources, each differing significantly from every other but nonetheless grouped into unified bibliographic entries: it is neither a census of copies nor a catalogue of library holdings. A short-title catalogue presumes that variations — among copies, re-prints, or re-issues — can be implied by each entry without being directly represented. For example, the Wing entry for John Dryden's *MacFlecknoe* takes up just a few lines, listing two entries, 1682 for D. Green, and another edition ("Anr. ed".) in 1692 for Jacob Tonson, while providing abbreviations for a few libraries around the world that held copies in 1945. (Figure 1.) The abbreviation

MACD refers to the still-definitive 1939 bibliography of Dryden's works, by Hugh MacDonald, where *MacFlecknoe* is registered as items 14a and 14b.

Echoes of Wing's bibliography can be found in the TEI headers of EEBO-TCP's documents. (Figure 2.) The EEBO-TCP edition of *MacFlecknoe* is composed of 12 kilobytes of XML data and 5 TIFF images, scanned from microfilm (position 8 in reel 785 of the *Early English Books* collection). Notice how thoroughly integrated into the bibliographic tradition the file is: corresponding identification numbers in Wing, ESTC, and EEBO allow users to cross-reference the metadata with print and digital sources. (A quick examination of the record in ESTC lists thirty repositories where physical copies might be found for comparison.) Information about the print source is included as well, drawn from Wing and ESTC. The film was photographed from a copy owned by Duke University Library, where it was bound with *Absalom and Achitophel* and *The Medal*, though that binding is not represented in the film nor, therefore, in the scanned images. The book was drawn from the later edition (14b in MacDonald) that was printed in London in 1692 for Jacob Tonson, though the work may have first appeared to the public as part of Dryden's works, published in 1693 and 1695.

Contra assertions that digital representation somehow occludes attention to physical realities outside itself, the metadata of the XML file is designed to be integrated with print resources. It highlights areas of uncertainty and it notes oddities in the source copy, at all times inviting users to be mindful of variations that may appear across the 30 copies known to exist. Of course, none of this can be assumed to be complete nor perfectly trustworthy. Despite the best efforts of cataloguers and archivists, errors certainly crept in, perhaps even in this very record.

Comprehensive enumerative bibliographies like the Short-Title Catalogue (or like the headers drawn from TEI collections) thereby invite a very strange reading practice, though perhaps it's been so naturalized since the information revolutions of the nineteenth and early twentieth centuries that it no longer feels strange. To read an entry from the STC is to project one's imagination outward to libraries across Britain and North America, where items (probably) exist that are likely to share many of the characteristics described, but which are also presumed to exhibit variations not represented in the entry. Bibliographic catalogues provoke a kind of sublime experience, an awareness of ambient textuality, whispering: *Books like this, but different, exist*. This sublimity is most clearly reflected in the terrestrial admonishments of the bibliographers themselves, who insist with raw certainty that the books to which their books refer are real, and fragile and

```
<fileDesc>
 <titleStmt>
  <title>MacFlecknoe</title>
  <author>Dryden, John, 1631-1700.</author>
 </titleStmt>
  <editionStmt>
   <edition>
    <date>1692</date>
   </edition>
  </editionStmt>
  <extent>Approx. 12 KB of XML-encoded text transcribed from 5 1-bit group-IV TIFF page images.</extent>
  <publicationStmt>
   <publisher>Text Creation Partnership,</publisher>
    <pubPlace>Ann Arbor, MI ; Oxford (UK) :</pubPlace>
    <date when="2003-01">2003-01 (EEBO-TCP Phase 1).</date>
    <idno type="DLPS">A36643</idno>
    <idno type="STC">Wing D2304</idno>
    <idno type="STC">ESTC R1438</idno>
    <idno type="EEBO-CITATION">13429983</idno>
    <idno type="OCLC">ocm 13429983</idno>
    <idno type="VID">99523</idno>
     <availability>
      <p>This keyboarded and encoded edition of the work described above is co-owned by the institutions providing
financial support to the Early English Books Online Text Creation Partnership. This Phase I text is available for reuse,
according to the terms of <ref target="https://creativecommons.org/publicdomain/zero/1.0/">Creative Commons
1.0 Universal</ref>. The text can be copied, modified, distributed and performed, even for commercial purposes, all
without asking permission.</p>
     </availability>
   </publicationStmt>
   <seriesStmt>
    <title>Early English books online.</title>
   </seriesStmt>
   <notesStmt>
    <note>(EEBO-TCP ; phase 1, no. A36643)</note>
    <note>Transcribed from: (Early English Books Online ; image set 99523)</note>
    <note>Images scanned from microfilm: (Early English books, 1641-1700 ; 785:8)</note>
   </notesStmt>
   <sourceDesc>
    <biblFull>
     <titleStmt>
      <title>MacFlecknoe</title>
      <author>Dryden, John, 1631-1700.</author>
     </titleStmt>
     <extent>8 p.  </extent>
     <publicationStmt>
      <publisher>Printed for Jacob Tonson,</publisher>
      <pubPlace>[London :</pubPlace>
      <date>1692]</date>
     </publicationStmt>
     <notesStmt>
      <note>A satire against Thomas Shadwell.</note>
      <note>Imperfect: title page wanting.</note>
      <note>Imprint from Wing.</note>
      <note>It is not certain whether this is a separate issue or whether it originally formed part of v. 4 of Dryden's
works, which were issued with collective title page by J. Tonson in 1693 and again in 1695.</note>
      <note>Originally published in 84 p. with Absalom and Achitophel and The medal, which are lacking in filmed
copy.</note>
      <note>Reproduction of original in Duke University Library.</note>
     </notesStmt>
    </biblFull>
  </sourceDesc>
</fileDesc>
```

Figure 2. The <fileDesc> element from the TEI header for *MacFlecknoe*. EEBO-TCP A3664

scarce and plentiful, and irreducible to the abstractions of their articulation. It's reiterated by EEBO's critics, who surrender to this sublimity while throwing stones at the digital artifacts that stimulate it.

After 1926 and 1945 when the first short-title catalogues were published, something like the total reality of the print record could for the first time be glimpsed through the panoptic gaze of the catalogue, even if this view was by proxy and therefore by definition partial, abstract, and potentially misleading (and limited to English books printed before 1700). We might call this gaze a kind of "distant reading", and indeed STC records have proved fruitful ground for statistical analysis.[9] But for Pollard, Wing, and later commentators, their awareness of the catalogues' extraordinary potential as a new form of reading-like knowing registered only as a concern about its limitations, its abstractions, and its propensities for error. Hence their insistence that the catalogue be used solely as a finding aid. Critical authenticity must continue to reside in the individual scholar's consultation of paper-based books in actual archives, not in the mere "topographic map" provided by bibliographies.[10]

The advent of microfilm promised (threatened?) to render such consultation obsolete, or at least unnecessary in all but the most specialized cases. The figure most commonly associated with this development in discussion of *Early English Books Online* is Eugene Power, an early proponent of microphotography for book preservation, dissemination, and on-demand printing. EEBO as we know it comes directly out of his work.[11] (Figure 3.) Power is an interesting figure. In many ways he represents the ideal of twentieth-century masculine subjectivity — even his name seems lifted from an Ayn Rand novel. Liberal Republican turned Democrat, Eugene Power was an entrepreneur and philanthropist who without irony promoted his private corporation as a public good.[12] He sat on many corporate boards and supported a wide range of liberal political causes. Always hoping to be in tune

9. ESTC data is often used by book historians as a proxy for book-trade activity, with all the usual caveats. See, for example, Steven N. Zwicker, "Is there such a Thing as Restoration Literature?" *Huntington Library Quarterly* 69, 3 (2006): 425–49.
10. Tabor, "ESTC and the Bibliographic Community", 368.
11. Power's work leading up to the creation of EEBO is usefully summarized in Mak, "Archaeology of a Digitization", 1517–19. See also his autobiography, *Edition of One* (UMI Press, 1990).
12. For Power's life, see his autobiography, written with Robert Anderson: *Edition of One: the Autobiography of Eugene B. Power* (Ann Arbor: University Microfilms International, 1990).

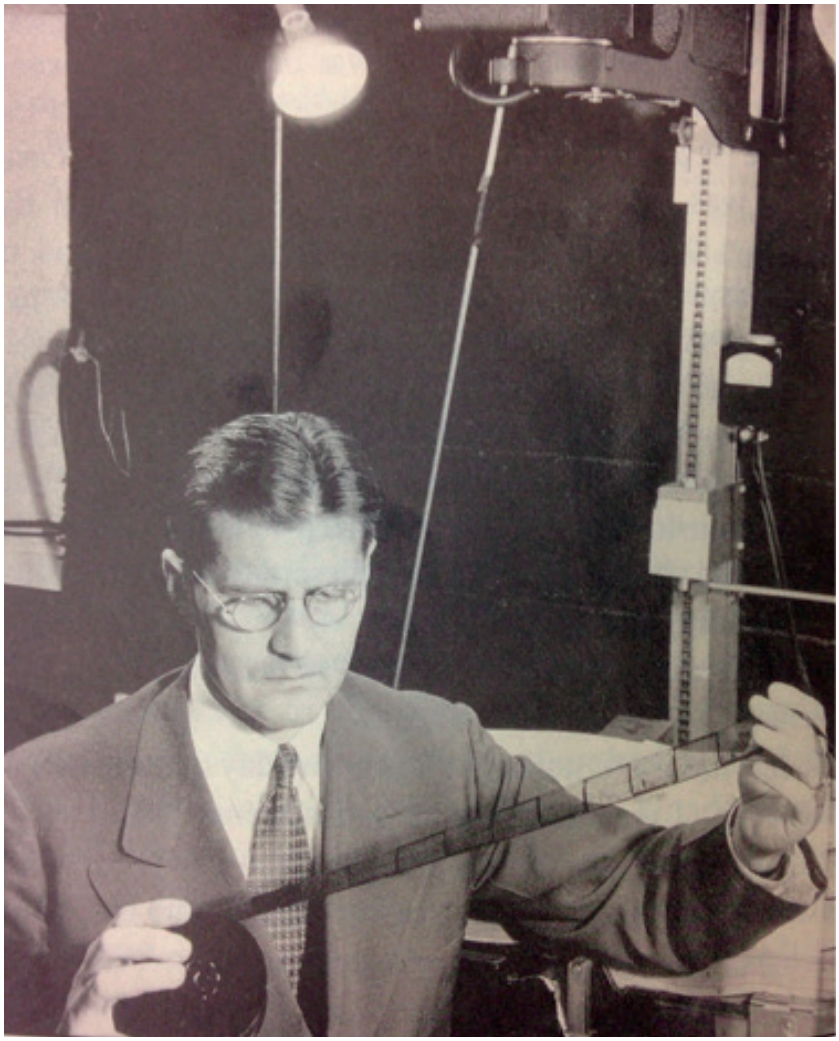


Figure 3. Eugene Power. *Edition of One*, p. 92.

with the needs of scholars and academic institutions, who were his most important customers, Power maintained a long association with the University of Michigan, and his publishing company, now named ProQuest, remains a major partner of university libraries.[13]

13. In fact, Power's business was so intimately connected to the University of Michigan that he briefly served on the Board of Regents until 1966, when the state's

But to understand Power's career requires tracing this history back one step further, to Robert C. Binkley. In the 1930s Binkley was a young Stanford-trained historian but he already had been appointed to an important post with the Social Science Research Council and the American Council of Learned Societies. Writing in 1958, Power recalls first meeting Binkley at a symposium in 1931, in which Binkley argued that the "deterioration of wood pulp and sulphite papers" would, within two centuries, leave "little by way of permanent records". Power describes Binkley's intellectual force in glowing terms: "For the small group of young men who sat at his feet and felt the force and excitement of his ideas, he is unforgettable, and through some of us his work goes on".[14] Concerned that nineteenth-century paper was dangerously ephemeral, Binkley urged his adoring audience to embrace microphotography as its replacement. We now know that this threat of decay and deterioration was exaggerated, but in the 1930s it motivated a massive effort to develop new technologies for document management.[15]

---

attorney general forced him to resign because of potential conflict of interest. Power describes these events in *Edition of One*, including an appendix that contains documents related to his resignation.

14. Eugene Power, "O-P Books, A Library Breakthrough", *American Documentation* 9, 4 (October 1958): 273.

15. Nicholson Baker has argued that this fear of imminent deterioration was not well founded, and in fact most forms of paper are actually easier to preserve than film. See *Double Fold: Libraries and the Assault on Paper* (New York: Random House, 2001). For an overview of Binkley's career, see Max H. Fisch's introduction to the *Selected Papers of Robert C. Binkley* (Cambridge: Harvard University Press, 1948); Kenneth Carpenter, "Toward a New Cultural Design: The American Council of Learned Societies, the Social Science Research Council, and Libraries in the 1930s", in *Institutions of Reading: The Social Life of Libraries in the United States*, ed. Thomas Augst and Kenneth Carpenter (Amherst: University of Massachusetts Press), 283–309; and Lisa Gitelman, *Paper Knowledge: Toward a Media History of Documents* (Durham: Duke University Press, 2014). Perhaps because of his influence on Power, whose work became so important to the humanities, Binkley is often mentioned in this regard, but he was by no means the only major proponent of microfilm in the early twentieth century. G. Watson Davis, whom Power mentions only in passing, is often cited as a more important figure in the sciences. Davis's parallel activities are described in Alistair Black and Dave Muddiman, "The Information Society Before the Computer", in *Early Information Society: Information Management in Britain before the Computer*, ed. Alistair Black, Dave Muddiman, and Helen Plant (Abingdon: Ashgate, 2012), 18–23.

The speech that Power heard likely included an argument that Binkley would publish a few years later in the *Yale Review* as "New Tools for Men of Letters".[16] According to Binkley, twentieth-century scholars faced a unique situation in the history of knowledge. The sheer volume of collected written matter meant that more information was available than ever before, but the increased specialization of academic labor meant that individual researchers needed access to smaller and smaller portions of this increasingly massive whole. Printing, which since the hand-press era enabled the creation of thousands of copies, needed to be replaced by a technology that could handle greater volume while targeting individual texts to much smaller audiences. His essay is worth quoting at length:

> The relation of the scholar-reader to the books on the library shelves has been changing. The body of documentation that was once the common ground of all learning and culture has lost its cohesion. And it has become a relatively unimportant element in the total bulk of publication. Today the Western scholar's problem is not to get hold of the books that everyone else has read or is reading but rather to procure materials that hardly anyone else would think of looking at. This is, of course, the natural consequence of the highly specialized organization of our intellectual activity. As a result, so far as Western culture is concerned, the qualities of the printing process that began in the fifteenth century to make things accessible have now begun in our different circumstances to make them inaccessible. When many if not all scholars wanted the same things, the printing press served them. In the twentieth century, when the number of those who want the same things has fallen in some cases below the practical publishing point (American Indian language specialists are an illustration), the printing press leaves them in the lurch. Printing techniques, scholarly activities, and library funds have increased the amount of available material at a tremendous rate, but widening interests and the three centuries' accumulation of out-of-print titles have increased the number of desired but inaccessible books at an

16. This essay has enjoyed a second life in the twenty-first century as an analog precursor to the techno-futurism of Internet enthusiasts. Lisa Gitelman writes, "More so than most of his peers, Binkley had a keen sense of living amid a continually accumulating and imperfectly preserved historical record, a sea of documents, the great recent accumulation of which was in jeopardy both because the necessary commitment to stewardship was lacking and because of the nineteenth-century switch from rag-based paper to less durable stock" (*Paper Knowledge*).

> even greater rate. Scholarship is now ready to utilize a method of book production that would return to the cost system of the old copyist, by which a unique copy could be made to order and a very few reproductions supplied without special expense.[17]

It's worth noting here an underlying similarity between Binkley's concerns and those of bibliographers like Pollard and Wing. The accumulation of out-of-print books promised greater accessibility to the world's knowledge, but this accumulation had coincided with institutional and economic developments that heightened demand for rare material. Demand exceeded libraries' ability to communicate information out, creating a bottleneck in research activity. Bibliographers approached this problem in a relatively narrow, tactical way, designing aids for scholars who hoped to wade into this great mass of documentation, while at the same time urging those scholars to stay mindful of their catalogues' inadequacies. For Binkley, the only conceivable solution was strategic and technological. With its comparatively cheap production and storage costs, microphotography promised to resolve print's contradictions and to meet the needs of institutions and individuals both.

The potential for microfilm to condense and cheaply reproduce massive amounts of information captured the imaginations of many writers during this time, when information science as an academic discipline (and IT as an "institutional desiring engine", in Alan Liu's phrase) was just beginning to gain public attention.[18] Librarian Fremont Rider argued in *The Scholar and the Future of the Research Library, A Problem and Its Solution* (1944) that libraries should replace their holdings entirely with micro-card readers.[19] Rider even invented a genre of film-based storage, called Microcard, that achieved modest success during the 1950s before being beaten out by rival formats.[20] In 1945, Vannevar Bush's futurist essay, "As We May Think",

17. Robert C. Binkley, "New Tools for Men of Letters", in *Selected Papers of Robert C. Binkley*, ed. Max H. Fisch (Cambridge: Harvard University Press, 1948), 182.
18. Alan Liu, "The State of Digital Humanities: A Report and Critique", *Arts and Humanities in Higher Education* 11, 1–2 (February/April 2012): 9. Kenneth Carpenter emphasizes microfilm's important place within the institutional ecology of the 1930s in "Toward a New Cultural Design".
19. Fremont Rider, *The Scholar and the Future of the Research Library, A Problem and Its Solution* (New York: Hadham Press, 1944)
20. For an overview of Rider's career, see Martin Jamison",The Microcard: Fremont Rider's Precomputer Revolution", *Libraries & Culture* 23, 1 (Winter 1988):

Figure 4. Recordak Microfilm Viewer, ca. mid-1960's. University Archives Photograph Collection, University of Wisconsin, Eau-Claire.

1–17. Rider comes in for special ridicule by Nicholson Baker, who quips that his "enduring achievement was to convince the heads of research libraries that it was somehow embarrassing to add more low-cost storage space" (*The Double Fold*).

appeared in *The Atlantic*, regaling readers with the tale of a "memex" machine that could condense millions of microfilmed records and easily retrieve them using a process of associative selection that would mimic human consciousness.[21]

Competition was fierce to bring this vision to reality, and reading and recording devices like Kodak's "Recordak" machine promised modern efficiency and style. (Figure 4.) When confronted by actual users, however, microfilm reading machines developed a reputation for being difficult to learn and straining to use. By the 1970s, one commentator remarked that the "reluctance of most readers to use microfilm or other microform is too well known to argue".[22] Nonetheless, development of microfilm was generously supported by government and other non-profit initiatives, who doled out hundreds of millions of dollars in grant funding for the creation of film-based archives and the installation of machines to access them.[23]

Back in 1935 in Michigan, Eugene Power's innovation was to repurpose the Short-Title Catalogue as an index for microfilm reproduction. According to Power, "It seemed to me that photographing STC books would be an ideal trial, since the collection was extensive, some 26,000 titles, and demand for them would be certain: American libraries, having been established relatively recently, were generally lacking in STC titles".[24] With 16 institutional subscribers, Power began microfilming select books, chosen for their likely research interest to American scholars.[25] As World War II approached, however, concerns about preservation became paramount, and in 1940 the American Council of Learned Societies declared that microfilming rare materials and storing those reproductions safely in America was an urgent priority. Power won a $30,000 grant from the Rockefeller Foundation to photograph six million pages of early English books, all selected from the Short-Title Catalogue. Although the microfilming process would continue from the mid-century heydays through the 1990s, it got its impetus during this moment of global conflict, when British libraries and the entirety of early English print faced very real physical danger.

21. Vannevar Bush, "As We May Think", *The Atlantic* (July 1945).
22. Rolland E. Stevens, "The Microform Revolution", *Library Trends* (January 1971): 388.
23. Details of this history are told, with a punchy and indignant tone, by Nicholson Baker in *Double Fold*.
24. Power, *Edition of One*, 28–29; partially cited in Mak, "Archeology of a Digitization".
25. These developments are described in Mak, "Archeology of a Digitization".

For Power, microfilm was bound up in Binkley's vision for micropublishing, and he long considered its most important application to be on-demand production of out-of-print books, what he called the "edition of one". In the 1950s, combining microfilm with Xerox printing made it economically feasible to singly produce bound copies of books. Power hailed this development as a major breakthrough in information technology. He writes:

> The history of the written word has gone full circle: from the single manuscript copy to the limited editions of the early printers; to the big editions of modern printing technology; to the smaller editions of offset and mimeograph; and back to the single copy edition of an O-P [out-of print] Book. For the first time, through the proper use of 35mm microfilm, the full history of our culture can be preserved and stored at low cost and, when copies are needed, they can be produced in any desired quantity at rates scholars can afford.[26]

The archival scope of microfilm preservation, Power believed, provided an information base for producing (and selling) relatively attractive paper copies of books that could facilitate familiar reading styles (Figure 5). This idea remained a small but important part of UMI's business model, and in 2010 ProQuest began offering bound prints from the Early English Books collection for sale on Amazon and other retailers. The publisher's blurb that accompanies all EEBO Editions prints could have been written by Power himself: "Imagine holding history in your hands. Now you can. Digitally preserved and previously accessible only through libraries as Early English Books Online, this rare material is now available in single print editions. Thousands of books written between 1475 and 1700 can be delivered to your doorstep in individual volumes of high quality historical reproductions".

On-demand printing may be a service that scholars and other readers sometimes find useful, but most readers of this essay, I suspect, will share my sense that it's rather beside the point. 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. When emphasizing its uses for on-demand printing, Power reduced the *Early English Books* collection to a marketing mechanism that simply brought difficult-

26. Power, "O-P Books, A Library Breakthrough", *American Documentation* (1958): 276.

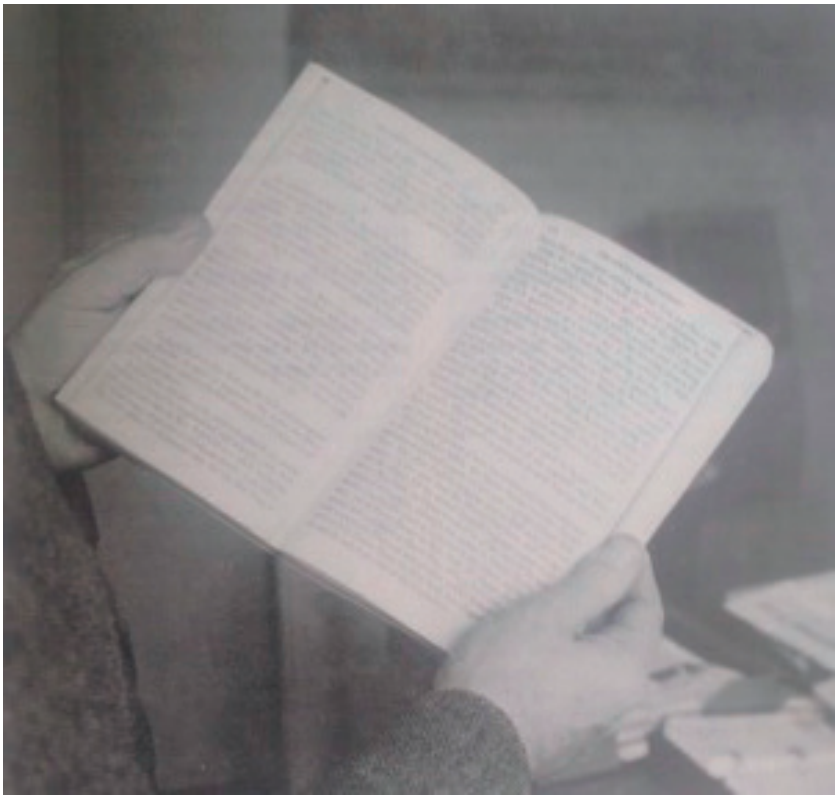


Figure 5. A single-copy edition of an out-of-print book, printed by Xerox from the *Early English Books* microfilm collection. From Power, "O-P Books, A Library Breakthrough" (1958).

to-find products into the hands of his customers. Power's mistake, if I may call it that, was to misunderstand the gulf that separates the reading styles facilitated by microfilm from those that involve the more familiar (and more physically comfortable) act of handling bound paper. In this sense, Power makes exactly the same mistake that bibliographers like Pollard and Wing made when describing the Short-Title Catalogue as a finding aid. All seem to assume that the most important outcome of their work would be to facilitate reading in conventional ways. They hoped to put human-shaped protein bags in direct physical contact with book-shaped rag pulp.

Much more important was the Short-Title Catalogue's capacity for giving scholars a sense of "what there was", even if that sense is, as they

insisted, ontologically incomplete. The catalogue provided a record system that made library holdings visible and therefore accessible. That visibility depended on translating archives to historical points of reference outside themselves: to authors, titles, imprints, catalog numbers, libraries, etc. Catalogues fold archives into history by layering them with historical metadata. Microfilm inserts page images into this data structure; it textures the sublime biblioscape with images, such that *Books like this, but different, exist* could be experienced anew as *Pages like this, but different, exist*. Catalogues abstract books from the shelf and treat metadata as proxies for them. Microfilm reconfigures this metadata as the index of a new archive, still pointing outside itself, but providing also an internally coherent proxy for rare books. This was the primary intellectual innovation of *Early English Books*: it re-purposed the STC as an index of an image collection. No longer a mere finding aid, the STC became an authoritative mechanism over which search queries could be performed from virtually any university library.

When the *Early English Books* microfilm collection was digitized in the 1990s, its index was transformed into a computer database, which whetted the scholarly appetite for more advanced search capabilities.[27] "EEBO's presentation of the ESTC metadata in database format made it possible to rapidly search citations for particular words or phrases and then access images of the texts indicated. Scholars soon sought to perform similar searches on the full texts of the works in this corpus".[28] ProQuest collaborated with the University of Michigan to solicit support from partner libraries, and in March 2000 a working group was tasked with deciding how the documents would be transcribed and encoded.[29] The group decided on a simple but rigorous descriptive markup, providing more than mere transcriptions, but leaving the documents open to adaptation. As we'll see, the move from page images to descriptive markup entailed yet another transformation in reading practices and textual form, as well as an altogether new conception of textuality as such.

27. The compilation of EEBO's metadata involved ingesting information from various sources. For a succinct review of this process, see "History of Early English Books Online", *Folgerpedia*. http://folgerpedia.folger.edu.
28. "History of Early English Books Online", *Folgerpedia*.
29. Rebecca Weizenbach, "Transcribed by hand, owned by libraries, made for everyone: EEBO-TCP in 2012". http://hdl.handle.net/2027.42/94307

The Invention of SGML and Its Adoption by the Text Encoding Initiative

Digitization rendered clunky microfilm machines obsolete while further easing search and the manipulation of page images. Even when digitized, however, the texts themselves — that is, the words actually printed in their sequences — remained invisible to the database's organizing structure, which was constrained by the physical forms of the books and their pages. The purpose of the Text Creation Partnership was to remedy this problem by remediating the books once more: this time by transcribing the page images and performing a basic descriptive markup that would enable full-text searching and online reading, while also generating tables of contents automatically. According to Rebecca Welzenbach, "For EEBO-TCP, the purpose of adding markup is to replicate the structure of the book, so that a user who does not have access to the page images or the original book will still be able to make sense of the text. Although of course all markup is interpretive, the aim has been to capture what is on the page, not to add new information".[30] In this comment, Welzenbach gestures toward a set of problems long familiar to digitally trained humanists, at least in the field of literary studies. Text markup puts into tension three tightly related bibliographic problems: first, to determine the overall structure of a text; second, to evaluate the relationship between that structure and the physical layout of printed or written pages; and third, to justify the sometimes uncomfortable interpretive decisions that need to be made while editing. She also hints at the tendency, again among literary historians, to defer to page layout as the primary guiding authority.

The EEBO texts were encoded in TEI P3, the standard for digital editing that prevailed in 2000 when the Text Creation Partnership began work. TEI is or should be deeply familiar to all readers of this essay, but it's worth reviewing some of its history and theory to explain how descriptive markup intervenes in the continuing transformation of historical textuality.

The development of TEI was a large, interdisciplinary scholarly project, but the person most directly responsible for laying its intellectual and technological foundations was Charles Goldfarb, inventor of SGML. I sometimes describe Goldfarb as the most important literary theorist English professors never heard of. Goldfarb's career spanned from the 1960s through the early 2000s, during which time he designed and helped popularize a form of markup that would be adopted as the International Stan-

30. Rebecca Welzenbach, "Transcribed by hand".

dard (ISO 8879) language for structured data and document representation. This markup language — which is really, as we'll see, a metalanguage of textuality — provides the basic information infrastructure of electronic communications.

Much like Power, Goldfarb worked in industry at the outskirts of academia. A Harvard-trained lawyer, Goldfarb left the law in 1967 to work for IBM on document management and production. Inspiration for markup language came when he was hired to install a computerized typesetter for a Boston-area newspaper. Because newspapers have to produce large amounts of text on a tight daily schedule, typesetters developed a practice of using style sheets to guide production. Rather than write out instructions for formatting every element in an article, editors would label articles' parts (headline, byline, caption, etc.) and then the typesetters would associate those categories with their appropriate formatting instructions.[31] Goldfarb's basic insight follows this practice to separate a text's formatting from the definition of its parts, allowing documents to be shared across software systems and, later, providing the basic structure of HTML and TEI. Since 2007 when Microsoft adopted a similar XML format for its Office suite, virtually all electronic texts have been built on the principles Goldfarb learned from watching newspaper typesetters.[32] (Book historians may find here a delicious irony. The centuries-long practice of composing type — the history of typesetting from early print compositors to twentieth-century newspaper editors — inspired the underlying design of the electronic applications that supposedly superseded print.)

In 1969, Goldfarb began a project that applied these principles to legal documents, not only to allow formatting instructions to be shared across printing systems, but also to expose various parts of each document to a common vocabulary for searching. Rather than maintain a separate database for case numbers, dates, plaintiffs and defendants, marked up case files allowed for direct searching across any of these variables. According to Goldfarb's biographer: "His idea was to treat different aspects of the document as data elements instead of as content. In this way, each legal document was actually a database of all its parts, with formatting code to

31. For Goldfarb's biography, see "Charles Goldfarb, Inventor of SGML", in *The Internet: a Historical Encyclopedia*, ed. Hilary W. Poole (Santa Barbara: ABC-CLIO, 2005), 1:126–31.
32. Goldfarb's idiosyncratic personal history is not the only important connection between print typesetting and computerized document management. See also the work of editor and typographer Stanley Rice, in particular his *Book Design: Text Format Models* (New York: R. R. Bowker, 1978).

describe each part of the text".[33] Working with Ed Mosher and Ray Lorie, they developed this idea into a schema they called Generalized Markup Language (GML), an acronym that "not coincidentally" corresponded to their own initials.[34]

Over the next decade, generalized markup was used increasingly by IBM for their own internal documentation and published materials. Goldfarb and others began developing GML into a rigorous standard for document production, and they submitted it to the American National Standard Institute (ANSI) in the late 1970s. GML was quickly picked up by the International Organization for Standards (ISO), and in 1986 it was published as ISO 8879/1986. GML became SGML, Standard Generalized Markup Language.[35]

Goldfarb's 1981 essay, "A Generalized Approach to Document Markup", should be required reading for all aspiring digital humanists, and indeed for anyone interested in text theory or computer remediation. Later adopted as the introduction to the ISO 8879 standard (Annex A), this essay lays out the basic theory that informs the design of virtually all electronic documentation. Though written in a flat style that emphasizes the theory's practical applications, Goldfarb's short essay puts forward a highly sophisticated model of textuality. He presupposes a separation, much like the distinction between "form" and "content", that distinguishes strings of words from the rules used to process those words for printing and display. Markup serves two purposes, he says: "it separates the logical elements of the document; and it specifies the processing functions to be performed on those elements".[36] What cultural theorists call "entextualization" Goldfarb calls "text processing": that is, the social and technological procedures that segment discourse into textual objects.[37]

33. Poole, "Charles Goldfarb, Inventor of SGML",128.
34. Charles Goldfarb, *The SGML Handbook*, ed. Yuri Rubinsky (Oxford: Oxford University Press, 1990), 568.
35. Goldfarb, *The SGML Handbook*, xiv–xv. See also Charles Goldfarb, "The Roots of SGML — A Personal Recollection", *Technical Communication* 46, 1 (1999): 75–83.
36. Charles Goldfarb, "A Generalized Approach to Document Markup", *ACM SIGPLAN Notices* 16, 6 (June 1981): 68.
37. In his account of oral transmission among indigenous Brazilian peoples, anthropologist Greg Urban defines "text" most generally as any "segmentable linguistic form": that is, a text is a sequence of linguistic objects that are differentiated from their context and marked out as a common unit. "Entextualization, Replication, and Power", in *Natural Histories of Discourse*, ed. Michael Silverstein and

```
.sk 1
Text processing and word
processing systems typically
require users to intersperse
additional information in
the natural text of the document
being processed.
This added information, called
'markup,' serves two purposes:
.tb 4
.of 4
.sk 1
1.¬it separates the logical
elements of the document; and
.of 4
.sk 1
2.¬it specifies the processing
functions to be
performed on those elements.
.of 0
.sk 1
```

```
:p.
Text processing and word
processing systems typically
require users to intersperse
additional information in
the natural text of the document
being processed.
This added information, called
:q.markup::q., serves two purposes:
:ol.
:li.it separates the logical
elements of the document; and
:li.it specifies the processing
functions to be
performed on those elements.
::ol.
```

Figure 6. Procedural and descriptive markup compared. From Goldfarb, "A Generalized Approach to Document Markup".

Goldfarb begins by noting that all computerized documents require additional language, hidden from human readers, to allow machines to recognize documents' formal features. This supplementary code, or "markup", conventionally contained instructions for how texts should be displayed. Goldfarb refers to this as "procedural" markup, because it contains instructions for text formatting. The code on the left in Figure 6 shows his example

---

Greg Urban (Chicago: University of Chicago Press, 1996), 27. Entextualization is a process best exemplified by oral transmission, as stories are told, memorized, and re-told in a dynamic interplay between the metadiscursive features of the text (e.g., author, genre) and the social formations that police such features and give them meaning. Entextualization is therefore a process by which linguistic objects representing tradition are replicated through mechanisms of power. This most general view of textuality conforms broadly with histories of authorship in the field of print, specifically having to do with questions of copyright and censorship. See Mark Rose, *Authors & Owners: The Invention of Copyright* (Cambridge: Harvard University Press, 1993); Joseph Loewenstein, *The Author's Due: Printing and the Prehistory of Copyright* (Chicago: University of Chicago Press, 2002); William St. Clair, *The Reading Nation in the Romantic Period* (Cambridge: Cambridge University Press, 2004); and Jody Greene, *The Trouble with Ownership: Literary Property and Authorial Liability in England, 1660–1730* (Philadelphia: University of Pennsylvania Press, 2005). It also describes very well editorial and digitization projects like EEBO.

of procedural markup: .sk 1 means to insert one blank line; .tb 4 means to insert a tab of four spaces, and .of 4 creates a matching hanging indent for the numbered list of items. By contrast, the markup on the right includes no such processing instructions. Instead it is "descriptive" markup that identifies the text's basic structural features: :p. means that what follows is a paragraph, :q. segments a quotation from the main text, while :ol. and :li. identify, respectively, an ordered list and its list items. Notice how some things in the main text have been removed. The quotation marks that surround the word "markup" on the left are taken out, as are the numbers in the list itself. This has the practical benefit of allowing for greater flexibility: "The list items", Goldfarb explains, "might be numbered in the body of a book, but lettered in an appendix".[38]

The other important feature of GML is its rigorous generalization. Textual features like paragraphs and ordered lists are familiar and common, and so the :p. and :ol. elements described in 1981 have remained more or less unchanged in most applications over the intervening 35 years. However, authors and editors can't be constrained by existing textual forms, and so they often require different elements or wish to define existing features differently. Rather than attempt to define an exhaustive set of possible text elements, GML makes it "possible to advise the system about the attributes of any type of element the user creates. This is done by creating a formal definition, or 'model'. . . . While the markup in a document consists of descriptions of individual elements, a GML model defines the set of all possible valid descriptions of a type of element".[39] What this means is that an element like an ordered list (:ol.) can be defined as an element that contains list items (:li.), and list items can be defined as elements that contain words, or, in GML-speak, "character data". All elements can be assigned attributes that the editor defines — chapters might be numbered, sonnets might be required to contain exactly fourteen lines, images might have height and width. The markup in any individual document is "rigorous" because it's validated against these rules. The markup is "general" because the rules themselves are user-generated. GML isn't a language of document markup, really, but a guiding framework for editors to create their own textual schemes.

In GML, both the model that defines a text's features and the rules that guide its format are abstracted from the character data, from the ostensible "content" of the text. I use scare quotes here because Goldfarb does as well.

38. Ibid., 70.
39. Ibid., 71.

He uses the term very circumspectly. Containment is the defining metaphor of markup: documents contain lists, lists contain items, items contain words. And so on and so on. . . . The formal features of a subelement are the content of the parent element. Goldfarb explains,

> "Content" is, of course, a primary attribute, and is the one that the secondary attributes of an element describe. The content consists of an arrangement of other elements, each of which in turn may have other elements in its content, and so on until further division is impossible. One way in which GML differs from generic coding schemes is in the conceptual and notational tools it provides for dealing with this hierarchical structure.[40]

According to this definition, the string of words that make up a line of poetry aren't what that line is. Instead, those words are the value of an attribute of an object, called <line>. (I switch now to using modern angle-bracket < > notation, which is far easier to read.) Just as a <person> might have attributes like <height> and <birthplace> with values like "1.94 meters" and "Winnipeg", so too a <line> element might have attributes like <number> and <character data> with values like "14" and "I am not I; pity the tale of me". The <line> isn't the string of words that humans read as a line of poetry, but a data object that bears as one of its attributes the fact that it contains characters, and that has this particular string of characters for the value of that attribute. Those words don't add up to a <sonnet> except through the intervening formal features determined by the editor, who configures these elements in a hierarchical tree structure, for which the rules are set in a separate file and which, after processing, is invisible to the reading eye.

In the wake of ISO 8879's 1986 publication, SGML was quickly recognized by scholars as a potentially valuable tool for building electronic editions and as an illuminating theory of textuality in its own right.[41] By

40. Ibid., 70.
41. See Joan M. Smith, "The Standard Generalized Markup Language (SGML) for Humanities Publishing" *Literary & Linguistic Computing* 2, 3 (1987): 171–75. Though she gestures quite broadly toward a wide range of possible applications, Smith understood SGML's value primarily in its capacity for supporting different output, and her description of its potential impact is reminiscent of how Binkley and Power described microfilm: "The important thing is that the text is retained, in a data base, where it may be updated at will (without affecting cross-references since it is the application software that specifies these at the output

the early 1990s, SGML had already accumulated a significant academic following.[42] Scholars like Michael Sperberg-McQueen at the University of Illinois and Allen Renear at Brown University worked to reconcile markup with text theory and scholarly editing.[43] According to Renear, the fact that generalized markup provided "so many different kinds of advantages, seemed to some people to suggest that it was not simply a handy way of working with text, but that it was rather in some sense deeply, profoundly, *correct*".[44] Under descriptive markup, the text is reimagined as an "ordered hierarchy of content objects" in which any text can be defined as an ordered sequence of parts, each of which is composed hierarchically of constituent parts. A play is made up of a certain number of acts that occur in a certain order; each of those acts is made up of a certain number of scenes, and each of those scenes by a certain number of speeches; the speeches are made up by words, which in turn are made up of characters.[45] The sequence and the hierarchy determine the structure of the text — indeed, any text. Descrip-

---

stage). Different sheets may be applied to it; it can be used for different purposes and output on different media (including microfiche and compact disk). Books can be published in accordance with different house styles, both European and American if different editions are required. Subsequent editions can be brought out as and when necessary or desirable, and there can be extractions of certain elements" (173).

42. See Robin Cover, Nicholas Duncan, and David T. Barnard, "The Progress of SGML (Standard Generalized Markup Language): Extracts from a Comprehensive Bibliography", *Literary & Linguistic Computing* 6, 3 (1991): 197–209.
43. For probably the best early overview of markup as a tool for literary editing and analysis, see C. M. Sperberg McQueen, "Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts", *Literary & Linguistic Computing* 6, 1 (1991): 34–46. For a description of the OCHO model, see Steven J. DeRose, David G. Durand, Elli Mylonas, and Allen Renear, "What is a Text, Really?" *Journal of Computing in Higher Education* 1, 2 (1990): 3–26; as well as the revision to the theory in Allen Renear, Elli Mylonas, David G. Durand, "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies", in *Research in Humanities Computing*, ed. Nancy Ide and Susan Hockey (Oxford: Oxford University Press, 1996), 263–80.
44. Allen Renear, "Text Encoding", in *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, John Unsworth (Oxford: Blackwell, 2004), original emphasis.
45. These "characters" differ starkly, of course, from the characters of the play, who according to TEI are represented either using <speaker> elements or by providing @who attributes for <sp> elements. In either case, and in most cases, the TEI represents personality (and, arguably, personhood) as metadata to text

tive markup works as well as it does because it identifies that hierarchy, making it explicit and available for systematic processing.

Identifying the "logical structure" of discourse and differentiating that structure from the "processing instructions" that render it as a document, to return to Goldfarb's original formulation, was not always so easy. Once the task shifted from authoring new documents with customized elements to using SGML as a metalanguage for describing already existing documents, the practical problems involved threw new light on the underlying theory. Think of it this way: when scholars edit historical texts into TEI, they have to choose tags for documents that are already published, and so they are essentially reverse-engineering those documents, trying to imagine the best-possible descriptive markers that might have informed their original production, had SGML existed at the time. (Digital editing really is a gloriously absurd intellectual activity!) Typically, and this has been the case with the EEBO-TCP, it means choosing elements that will reflect back something like the formatting of the page layout of the source copy.

There's no reason why print format has to be the guiding authority, however. The whole premise of GML was to allow users to define their own textual models. For Goldfarb, this meant that the system could be universally applicable and interoperable: "text processing" named a set of protocols for converting character data into documentation. For literary critics, however, this extensibility meant something very different: markup promised a textual system explicitly sensitive to the unbounded possibilities of interpretation. Whereas a literary scholar might choose to tag a play's act and scene divisions, a linguist might leave those out and focus instead on a grammatical analysis of each sentence. (More on this below.) In either case, the structure of the text is determined by an element set that is chosen by the editor and defined in a separate file. In SGML (now XML), every text becomes an archive of its own parts, but the principle of differentiation among those parts is necessarily extrinsic to the text. A file containing marked up character data is therefore never identical to the text, as such. The "text" in a literary sense, or the "document" in Goldfarb's, is the result of a process that imposes structure from the outside and realizes discourse through that structure. For this reason, markup language is not so much a theory of text as a theory of entextualization.[46]

---

sources. See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/examples-sp.html.

46. Though they don't put it in quite these terms, Renear et al. make a similar argument in "Refining Our Notion of What a Text Really Is".

Within months of its adoption as an international standard, humanists began to think about how they might use SGML to set guidelines for editing electronic texts. In November 1987, Nancy Ide, a computational linguist at Vassar College, led a meeting sponsored by the Association for Computers and the Humanities and the National Endowment for the Humanities that established a set of protocols for these guidelines, what came to be known as the "Poughkeepsie Principles":

> The Poughkeepsie Principles define two functions for the guidelines: to recommend a format for interchange of texts, and to recommend principles and practices for the encoding of new texts. Existing archives have large investments in their existing schemes and will have no motive for converting the storage format of their holdings. But they are keenly interested in reducing the number of other formats from which and into which they must translate their texts, by helping develop and support a single common format for interchange. Scholars working to encode new texts — many of them novices in computing with no investment at all in any existing scheme — will benefit from having some guidance about what textual features to encode and how to encode them.[47]

The Text Encoding Initiative was formed to address these needs. Its initial advisory board was international and interdisciplinary. Computational linguistics was represented by Donald Walker, Clifford Lynch, Antonio Zampolli, Scott Deerwester, and others (including Ide). Representatives of major academic associations included Joseph Hollander and Randall Jones (MLA), Peggy Brown (AHA), Susan Hockey (ALLC), Anne-Maria di Sciullo (CLA), and so on. Michael Sperberg-McQueen joined the project as the TEI's editor.

When the first board meeting convened in Chicago in 1989, Lou Burnard presented SGML to the group. After walking his audience through the basic structure of SGML's metalanguage, he pointed out several problems that would face the initiative: "all existing SGML applications are for producing, not for analyzing, documents", even though scholars are most interested in analysis, and "real documents have more than a single document hierarchy", and so will present numerous analytical and interpretive challenges. "But despite its problems SGML is a great step forward for markup and a solid base for our work", he concluded, adding, with char-

47. C. M. Sperberg-McQueen, "Minutes of Advisory Board Meeting, Chicago, 18–19 Feb 89". TEI ABM1. <http://www.tei-c.org/Vault/AB/abm01.gml>

acteristic flair, "God did not bring us into this world and give us minds in order to choose, while writing, between roman and italic type".[48]

God, it seemed, had brought them into the world to design element tags (or, at least, God brought them to Chicago). This task was divided among several working groups. The Text Documentation group was formed to design a scheme for metadata about the files and their sources, which was the relatively easy part of the project, because they could imitate protocols developed for library catalogues. More challenging were the tasks of the other two main groups. The Text Representation group sought to develop a set of standard tags for representing the formal and bibliographic features of paper-based documents, and the Text Analysis and Interpretation group designed tags for linguistic and thematic markup.

The Text Representation working group was chaired by Stig Johansson and included Burnard and Sperberg-McQueen, as well as David Chesnutt, Steven DeRose, Susan Hockey, Elli Mylonas, William Ott, and Manfred Thaller. (Basically, a who's who of 1990s markup theory.) They were charged with considering "techniques for encoding all the information explicitly present in a copy text on the physical or graphetic level", such as quotations, topographical and layout information, figures and captions, and lineation, as well as structural features like chapters and paragraphs.[49] In its first meeting, the working group noted "the need to cater for the description of existing printed or manuscript texts", while acknowledging that "some users of the scheme would be interested in the physical description of a source, others in its logical structure and yet others in the relationship between the two".[50]

The Analysis and Interpretation group was "responsible for all interpretive material not conventionally represented physically in an edition", with an emphasis on transcriptions of spoken language, syntax, and interpretive features like style, theme, and content.[51] In 1990 they distributed a survey among literary scholars (receiving about 40 responses), who strongly agreed on the importance of marking up bibliographic information and the basic structural features of a text, but who were hesitant to endorse other kinds of annotation. When asked whether tags should include gram-

48. TEI ABM1. <http://www.tei-c.org/Vault/AB/abm01.gml>
49. Minutes of the meeting of the temporary Steering Committee. Pisa, 12–13 December 1987". TEI SCM01. <http://www.tei-c.org/Vault/SC/scm01.txt>
50. Burnard, Lou, "Minutes of the First Meeting of the Text Representation Committee of the Text Encoding Initiative Held at the University of Toronto, 6 June 1989". TEI TRM1. <http://www.tei-c.org/Vault/TR/trm01.tex>
51. TEI SCM01. <http://www.tei-c.org/Vault/SC/scm01.txt>

matical information, the respondents were ambivalent. One person averred that "Emphasis must be on flexibility, and avoiding any hint of prescription which would encourage the tail to wag the dog". Another remarked, "While this is not important to the ways in which I see myself using an electronic text, it might be important to a linguist". When asked whether markup should include interpretive information about the texts, noting features such as "narrative vs. expository passages, direct and indirect discourse, point of view, themes, images, [or] allusions", more than half responded with an emphatic "no". One person acknowledged the potential but explained the challenges such an effort would face:

> Any coding should work to loosen the web of the text and encourage its multivalence to be exploited in a non-print medium. Exciting potential here, I'd have thought. How can coding facilitate the exposition of multiple levels of, say "point of view" or the complexity of "themes", without constricting them? Can coding be sufficiently sensitive to maximize the examination of tensions between, say, overt levels of meaning, and, perhaps covert or subverted/-sive levels caused by dislocations within the varying "points of view" (authorial intended, historically and culturally conditioned, skewed by time, class, gender, race, etc. or the sheer slipperiness of the signifiers themselves) or between such semantic levels and acoustic/semiotic/paralinguistic levels?[52]

The interpretive possibilities of generalized markup promised to "loosen the web of the text and encourage its multivalence", but it was hard to imagine any actual coding scheme that wouldn't feel constraining.

These considerations were echoed in 2000, when a new task force was formed, this time by the Text Creation Partnership, to evaluate the TEI guidelines and choose a basic coding scheme for the Early English Books transcriptions. They noted that "if number of texts, length of project, and amount of money available are fixed, the level of encoding is constrained".[53] For this reason the coding scheme for EEBO-TCP tends to emphasize page-format information and basic document-structure features, concluding that "it is better to do less, than to do wrong or mislead" and that "all encoding decisions should allow for enhancement and avoid tag abuse". The files

52. Paul Fortier, "Literature Needs Survey Results", 22 January 1991. TEI AI3 W4. <http://www.tei-c.org/Vault/AI/ai3w04.txt>
53. DTD Working Group Notes", Text Creation Partnership. <http://www.textcreationpartnership.org/dtd-working-group-note/>

would be divided into a <teiheader> and <text> elements that formed the core structure of each document, putting metadata and textual information in the same file but keeping them distinct. Major divisions in the text would be marked with numbered elements, including poetry, which would organize all lines into line groups using <lg> tags. Page breaks would be noted with <pb>, and all marginal notations would be designated with <note> elements. All font shifts would be noted using the <hi> highlight element, without specifying among italic, gothic, and other fonts.

By sidestepping some of the more finely grained possibilities offered by the TEI guidelines, the EEBO-TCP editors avoided the complex interpretive decisions markup sometimes provokes, but by deferring to page layout as the guiding authority they also made things simpler for the coders themselves, who were working from scans of the Early English Books microfilm. Over the next fifteen years, EEBO files were sent in monthly batches to two third-party vendors, Apex CoVantage and SPi Global, whose employees performed the actual transcription and markup.[54] Files were prepared and reviewed by editorial staff at the University of Michigan, whose work was overseen by the project director, Paul Schaffner.[55] As texts were selected for transcription, the goal was to provide as comprehensive a sample of EEBO as possible, covering all major periods and genres.[56]

By 2010, the first phase of 25,000 texts had been transcribed and marked up. Sponsoring organizations enjoyed a five-year embargo on the tran-

54. Both Apex CoVantage and SPi Global are media companies that provide government, academic, and corporate clients with "content solutions", which include digital publishing and producing XML documentation. See <http://apexcovantage.com/content-solutions/solutions/> and <http://www.spi-global.com/content>. Transcriptions were performed by anonymous coders working in India.
55. Welzenbach, "Transcribed by hand, owned by libraries, made for everyone: EEBO-TCP in 2012".
56. Some bias may have crept into the selection. Because the vendors charged by the page, not by the title, there was a consistent bias towards documents that were comparatively short, as well as toward documents that were in English. Very long books with less obvious research value to historians — like long legal dictionaries — tended to be excluded to allow for a greater variety of shorter titles. Just as was true of the short-title catalogues, the EEBO-TCP should never be confused with a complete model of actually extant print. Schaffner explained this process in a meeting at the Folger Shakespeare Library, during their 2015 NEH-funded institute, "Early Modern Digital Agendas: Advanced Topics". <http://folgerpedia.folger.edu/EMDA2015_Curriculum>.

scriptions, which were released to the public on Github in January 2015, along with small sample collections from Eighteenth-Century Collections Online and Evans. Work has continued on the "Phase II" titles, which, when released from copyright in 2020 (hopefully), will bring the total collection to a little over 60,000 files, representing about one-third the total number of surviving titles. Pending some extraordinary breakthrough in OCR software, this is likely for the foreseeable future to be the corpus through which English history is measured.[57]

Conclusion: The Death of the Document

When I began my PhD program in 2005, it was still common to speak of "the death of the author" as a relevant event in intellectual history. My professors sometimes said things like "Since the death of the author . . ." in the same way journalists would say, "Since 9/11 . . ." The phrase evoked a sense of traumatic, epochal shift that divided time into a before-and-after — the present became a shared *Neuezeit* marked by crude emblems of affiliation.

On October 26, 1992, Charles Goldfarb proclaimed an altogether different death. The SGML conference was held in Danvers, Massachusetts that year, attracting a record attendance of 275 participants. Goldfarb delivered the keynote address. He began by expressing cautious optimism about SGML's success but warned that vendors of proprietary software would always have an incentive to push for system-dependent data representations. According to Michael Sperberg-McQueen:

> Moving to his main theme, Goldfarb proclaimed the death of the "document", which he said may in fact never have been anything more than a makeshift to enable the use of computer technology. The future of SGML lies in its use to link both within and between documents . . . He showed medieval pages (from the Winchester Bible) and discussed the division of labor among scribes, rubricators, illuminators, and applicators of gold leaf, which corresponds closely to the division of labor, in presenting a hypermedia document today, among the text displayer, the graphics presentation software, and other specialized modules.[58]

57. Laura Mandell has spearheaded the Early Modern OCR Project, which seeks to address this problem. See <http://emop.tamu.edu>
58. C. M. Sperberg-McQueen. "Trip Report: SGML '92, Danvers, Mass". <http://cmsmcq.com/1992/edr2.html>

Another attendee summed up Goldfarb's thesis like this: "The world of the isolated single document is dead".[59] What literary theory proclaimed as a shift from "text" to "hypertext" was figured rather differently by Goldfarb. The "document" was the organizing unit of discourse for information technology, but the very architecture that made documents visible to computer systems profoundly undermined their coherence, offering not so much a new form of documentation as a platform for multiple information streams. By comparing hypermedia to illuminated manuscripts, Goldfarb seems to be envisioning something like a webpage with streaming content.

Yet, the death of the document can be seen in the EEBO-TCP corpus too, where "titles" are a similar kind of makeshift. As a unit of discourse, "titles" are inherited from library catalogues but are used to render character data up for manipulation and analysis, even though such analysis often dissolves the boundaries of the titles themselves.

It is the combination of enumerative bibliography and text transcription and markup that makes the EEBO-TCP a resource of such unique power. A TEI-encoded file is a textual form unlike any other; the invention of this genre was an extraordinary intellectual accomplishment that remains under-appreciated. When combined with the archival research of the short-title catalogues (themselves scholarly projects of the highest quality) the result is a collection of files — files, not documents nor texts — that fold discourse into history in a remarkable way, combining "real life" sociological information about names, books, places, and dates, with the formal and lexical features of the texts that record that information. Scale is absolutely essential here. It matters that these projects aspired to comprehensiveness. The most interesting applications of corpus linguistics depend on a sufficiently large word base to get real interpretive traction, and EEBO-TCP provides a very large base. But, again, it isn't just a big corpus. The EEBO-TCP files provide a highly structured body of data that make possible analysis over any number of social or textual configurations.

The archive of early print is now remediated as a collection of networked particulars. Everything (that is, everything included in the model) is connected to everything else, at both the supra-textual levels of biographical and geographical metadata, as well as at the sub-textual levels of parts, down to the individual words and characters. Each item in the collection exists in relation to every other and is therefore available for re-formulation as data. This structure allows words, persons, and places to be represented

59. Michael Popham, "SGML '92 Conference Report, by Michael Popham". <http://xml.coverpages.org/sgml92.html>

in commensurable numeric forms that navigate elegantly among history's conflicting and overlapping ontological registers. Sometimes the data can be used to represent the career of an author (or a printer or bookseller or politician, or any group thereof). Other times it can stand in for large epistemic shifts. Still other times it can be made to represent the books themselves, or the places where those books circulated, or the readers who read them.

Of course, just as with any form of study, asking different questions involves provisionally accepting different assumptions and navigating different pitfalls. The mistake that literary historians make most routinely is to assume that explanation requires a consistently applied metaphysics — that words and matter and time exist in a knowably true relation, and so ideas that violate one's favored ontology are therefore simplistic, ideologically dubious, or just plain wrong. Computational textuality dispenses with this comforting but debilitating assumption. If I may venture to speculate: such metaphysical rigidity is quite possibly the real reason quantification makes so many scholars uncomfortable.

But to return to the story, in a nutshell. Catalogues took books off the shelves. Microfilm took pages out of books. Transcription and markup freed words from the page. Collection and standardization dissolved those words into data. Early print's realization as data opened a new horizon of study that we're still just beginning to survey.

The horizon itself was glimpsed early on. Among the members of the first advisory board of the TEI was Scott Deerwester, who represented the Association for Computing Machinery's Special Interest Group on Information Retrieval (ACM/SIGIR). At the first meeting in Chicago in 1989, as the members of the board went around the room introducing themselves, Deerwester described his own interest in TEI as an extension of his work designing algorithms to search over bibliographic records.[60] Full-text search, he said, raises a new question for information retrieval: "What are we retrieving?"

Now that we have EEBO in full-text form, what do we do with it?

60. TEI ABM1. <http://www.tei-c.org/Vault/AB/abm01.gml> Deerwester would go on to be known as one of the inventors of latent semantic analysis — a technique that teases out the major themes in a collection of documents, much like "topic modeling".

Works Cited

Baker, Nicholson. 2001. *Double Fold: Libraries and the Assault on Paper*. New York: Random House.

Binkley, Robert C. 1948. "New Tools for Men of Letters". *Selected Papers of Robert C. Binkley*, ed. Max H. Fisch. Cambridge: Harvard University Press.

Black, Alistair, and Dave Muddiman. 2012. "The Information Society Before the Computer". *Early Information Society: Information Management in Britain before the Computer*, eds. Alistair Black, Dave Muddiman, and Helen Plant. Abingdon: Ashgate.

Bush, Vannevar. 1945. "As We May Think". *The Atlantic*.

Carpenter, Kenneth. 2007. "Toward a New Cultural Design: The American Council of Learned Societies, the Social Science Research Council, and Libraries in the 1930s". I*nstitutions of Reading: The Social Life of Libraries in the United States*, eds. Thomas Augst and Kenneth Carpenter. Amherst: University of Massachusetts Press.

Cover, Robin, Nicholas Duncan, and David T. Barnard. 1991. "The Progress of SGML (Standard Generalized Markup Language): Extracts from a Comprehensive Bibliography". *Literary & Linguistic Computing* 6.3: 197–209

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen Renear. 1990. "What is a Text, Really?" *Journal of Computing in Higher Education* 1.2: 3–26.

Fisch, Max H, ed. 1948. *The Selected Papers of Robert C. Binkley*. Cambridge: Harvard University Press.

Gadd, Ian. 2009. "The Use and Misuse of Early English Books Online". *Literature Compass* 6.3: 680–92.

Gitelman, Lisa. *Paper Knowledge: Toward a Media History of Documents*. Durham: Duke University Press.

Goldfarb, Charles. 1990. *The SGML Handbook*, ed. Yuri Rubinsky. Oxford: Oxford University Press.

———. 1981. "A Generalized Approach to Document Markup". *ACM SIGPLAN Notices* 16.6: 68–73.

———. 1999. "The Roots of SGML — A Personal Recollection". *Technical Communication* 46.1: 75–83.

Greene, Jody. 2005. *The Trouble with Ownership: Literary Property and Authorial Liability in England, 1660–1730*. Philadelphia: University of Pennsylvania Press.

Jamison, Martin. 1988. "The Microcard: Fremont Rider's Precomputer Revolution". *Libraries & Culture* 23.1: 1–17.

Liu, Alan. 2012. "The State of Digital Humanities: A Report and Critique". *Arts and Humanities in Higher Education* 11.1–2: 8–41.

Loewenstein, Joseph. 2002. *The Author's Due: Printing and the Prehistory of Copyright*. Chicago: University of Chicago Press.

Mak, Bonnie. 2014. "Archaeology of a Digitization", *Journal of the Association for Information Science and Technology* 65.8: 1515–1526.

Nangle, Benjamin. 1945. "General Introduction", in Wing, *Short-Title Catalogue*, v, emphasis original.

Pollard, A. W. and G. R. Redgrave. 1926. *A Short-Title Catalogue of Books Printed in England, Scotland, & Ireland*. London: The Bibliographical Society.

Poole, Hilary W. 2005. "Charles Goldfarb, Inventor of SGML". *The Internet: a Historical Encyclopedia*. Santa Barbara: ABC-CLIO.

Popham, Michael. 1992. "SGML '92 Conference Report, by Michael Popham". http://xml.coverpages.org/sgml92.html

Power, Eugene. 1958. "O-P Books, A Library Breakthrough". *American Documentation* 9.4: 273–276.

Power, Eugene with Robert Anderson. 1990. *Edition of One: the Autobiography of Eugene B. Power*. Ann Arbor: University Microfilms International.

Renear, Allen. 2004. "Text Encoding". *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell.

Renear, Allen, Elli Mylonas, David G. Durand. 1996. "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies". *Research in Humanities Computing*, ed. Nancy Ide and Susan Hockey. Oxford: Oxford University Press.

Rice, Stanley. 1978. *Book Design: Text Format Models*. New York: R. R. Bowker.

Rider, Fremont. 1944. *The Scholar and the Future of the Research Library, A Problem and Its Solution*. New York: Hadham Press.

Rose, Mark. 1993. *Authors & Owners: The Invention of Copyright*. Cambridge: Harvard University Press.

Smith, Joan M. 1987. "The Standard Generalized Markup Language (SGML) for Humanities Publishing". *Literary & Linguistic Computing* 2.3: 171–75.

Sperberg-McQueen, C. M. 1991. "Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts". *Literary & Linguistic Computing* 6.1: 34–46.

———. 1992. "Trip Report: SGML '92, Danvers, Mass". http://cmsmcq.com/1992/edr2.html

St. Clair, William. 2004. *The Reading Nation in the Romantic Period*. Cambridge: Cambridge University Press.

Stevens, Rolland E. 1971. "The Microform Revolution", *Library Trends* January: 379–395.

Sutherland, Kathryn and Marilyn Deegan. 2016. *Transferred Illusions: Digital Technology and the Forms of Print*. New York: Routledge.

Tabor, Stephen. 2007. "ESTC and the Bibliographic Community", *Library* 8.4: 367–386.

TEI Consortium, eds. 2019. *Guidelines for Electronic Text Encoding and Interchange*. 29th January 2019. http://www.tei-c.org/P5/.

———. 1987. "Minutes of the meeting of the temporary Steering Committee. Pisa, 12–13 December 1987". TEI SCM01. http://www.tei-c.org/Vault/SC/scm01.txt

———. 1989. "Minutes of Advisory Board Meeting, Chicago, 18–19 Feb 89". TEI ABM1. http://www.tei-c.org/Vault/AB/abm01.gml

———. 1989. "Minutes of the First Meeting of the Text Representation Committee of the Text Encoding Initiative Held at the University of Toronto, 6 June 1989". TEI TRM1. http://www.tei-c.org/Vault/TR/trm01.tex

———. 1991. "Literature Needs Survey Results, 22 January 1991". TEI AI3 W4. http://www.tei-c.org/Vault/AI/ai3w04.txt

Text Creation Partnership. 2000. "DTD Working Group Notes". *Text Creation Partnership*. http://www.textcreationpartnership.org/dtd-working-group-note/

Urban, Greg. 1996. "Entextualization, Replication, and Power". *Natural Histories of Discourse*, eds. Michael Silverstein and Greg Urban. Chicago: University of Chicago Press.

Welzenbach, Rebecca. 2012. "Transcribed by hand, owned by libraries, made for everyone: EEBO-TCP in 2012". http://hdl.handle.net/2027.42/94307

Wing, Donald. 1945. *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: Columbia University Press.

Zimmer, Erica and Meaghan Brown. 2017. "History of Early English Books Online". *Folgerpedia*. https://folgerpedia.folger.edu/History_of_Early_English_Books_Online.

Zwicker, Steven N. 2006. "Is there such a Thing as Restoration Literature?" *Huntington Library Quarterly* 69.3: 425–49.