The Readex Corporation, the American Antiquarian Society, and the Brave New World of Electronic Text: A Librarian's Perspective

ROBERT SCOTT

To many a scholar coming of age in an academic library in the late twentieth century (and I draw here not only on my observations as a librarian but on my own experience as a graduate student at Columbia), the work of the Readex Corporation[1] naturally seemed not only as American as apple pie, but as librarian (in the adjectival sense) as the National Union Catalog. Very quickly, those learning to use the cutting-edge information tool of their day—the card catalogue—would stumble across one of those printed cards advising them that works by a given author were also likely to be available in the

[1] 'In June of 1950, New York publishing legend Albert Boni, co-founder of the Modern Library, formed Readex in Chester, Vermont. Boni initially launched the publication of the entire series of British Parliamentary Papers, under the sponsorship of the American Historical Association. Soon the company was filming the *New York Times* for the New York Public Library, as well as the Annual Subject Catalog for the Library of Congress and the declassified papers of the Atomic Energy Commission. Five years later, the American Antiquarian Society (AAS), an independent research library founded in 1812 and located

Robert Scott is Head, Electronic Text Service, as well as a member of the Reference Department of the History and Humanities Division of Columbia University Libraries.

Readex microform collection of *Early American Imprints*. Slowly it began to dawn on the reader that somewhere out there—actually just a few steps away in the microform reading room—lurked an enormous and comprehensive collection of American books. The ubiquity of those references and the form of their presentation—on preprinted cards like those distributed by the Library of Congress—suggested something transcending the purely commercial, that one was dealing here with an objective, almost disinterested arbiter of scholarly information.

And, indeed, there was a great deal of validity to that initial impression. It had something to do with the fact that this first classic set of microform material was based upon the bibliographic labor of Charles Evans, who had sought to be as broad and objective as possible in listing all that been published in this country prior to 1820.[2] It also owed much, as we have learned today from August Imholz's presentation, to the very idealistic vision of the Readex

in Worcester, Massachusetts, invited Readex to produce in microprint the incomparable *Early American Imprints, Series I, Evans, 1639–1800*. . . . This partnership led to the publication of such essential collections as *Early American Imprints, Series II, Shaw-Shoemaker, 1801–1819*, and *Early American Newspapers, Series I, 1690–1876*. Other special partnerships have enabled Readex to publish an incomparably rich digital edition of the U. S. Congressional Serial Set, 1817–1980 with the American State Papers, 1789–1838. The Readex digital *Serial Set* is a cornerstone collection for understanding American history, culture and daily life, and the effort behind its creation befits this national treasure. The Library of Congress is providing colored scanned images of its color maps, Dartmouth College its hardcopy volumes for digitization in original color of all text pages and illustrations, and Stanford University its Dublin Core records. Finally, Readex is providing all-new bibliographic records created by experts who examine and index each and every document to create the most comprehensive collection of its kind. Readex is also widely known for its high-quality Scholarly Microform Collections in such areas as American and English literature, women's studies, and the history of science. In addition, its essential Documents Collections include *United Nations Documents*, which offer comprehensive coverage of the United Nations from its founding to the present, along with several others. As Readex entered its sixth decade, Readex President David Braden said, "At its heart our goal today remains unaltered: to be leaders in fostering a greater understanding of American history and the role of the United States in world history for scholars, researchers, teachers and students at all levels" ' (http://www.readex.com/aboutred/aboutred.html).

[2] Charles Evans, *American Bibliography: A Chronological Dictionary of All Books, Pamphlets, and Periodical Publications Printed in the United States of America from the Genesis of Printing in 1639 Down to and Including the Year 1820*, vols. 1–12, (Chicago: privately printed, 1903–59).Volume 13 (1955), edited by Clifford K. Shipton, and volume 14 (1959), the index prepared by Roger P. Bristol, were published by the American Antiquarian Society.

Corporation's founder. There is little question, however, that it also reflected the unique, collaborative partnership of that company with the American Antiquarian Society, an organization dedicated to the broad collection and guardianship of the record of the American past, and to making that material available for the research needs of the broader scholarly community.

As we have already heard today, the creation of that microform set was an enormous achievement. Building on the work of Evans and then, in its continuation, of Shaw and Shoemaker,[3] it succeeded in bringing together the classic early American corpus for comprehensive consideration. It assured the preservation of rare and scattered works and brought them to countless institutions that could never have dreamed of owning even a small portion of these riches in the original. While a scholar might have once had to spend a lifetime running down the sources of his or her research at many institutions, he or she could now consume them in a single setting. As a result, hosts of outstanding historians were able to build their research on this collection.

Following upon that classic achievement came many other collections of documents, of later American bibliography, of early newspapers, of broadside and ephemeral literature. As a librarian, I am deeply appreciative of the value of microfilm (something Nicolson Baker would describe in more sinister terms),[4] recognizing the essential role this format has played in the preservation of and access to the record of human society. Indeed, its historic role, particularly in the context of classic corpus building, with the capacity the latter provided for a synthetic view of the heritage of individual societies and cultures, is probably insufficiently appreciated. That work also set the stage

[3] Ralph R. Shaw and Richard H. Shoemaker, *American Bibliography: A Preliminary Checklist for 1801–1819*, 22 vols. (New York: Scarecrow Press, 1958).

[4] For an earnest but ultimately wrong-headed assault on libraries' efforts to microfilm the deteriorating print heritage of the past, see Nicholson Baker, *Double Fold: Libraries and the Assault on Paper*, 1st ed. (New York: Random House, 2001).

for a much more exciting and to my mind—you will excuse me as head of the Electronic Text Service—even more important and vital step forward into the scholarship of the future when it served as the basis for vast online electronic collections, a development that could probably never have taken place without the initial collection development involved in the microfilming projects.

That said, as Ed Gray has graphically illustrated to us today, microform was a medium that separated the sheep from the goats, at least where scholars did not have ready access to the texts in their original form. Access to these broad collections was a remarkable thing, but it was not the same thing as reading a printed book. An individual willing to poke his head into a metal box and turn a crank to read a text was clearly one very interested in that text. Scholars for whom that title was only of tangential interest were less likely to make the pilgrimage upstairs to the microform collection, to say nothing of undergraduates doing research on some transient paper topic. On the Readex website, historian Carol Varnum has also eloquently described her complex love-hate relationship with the green gnome-like microform reader that lurked in the corner of her study.

In short, there were clearly some problems to be addressed. The collection's vision of a broad access was in fact hampered in some ways by the difficulty of using that medium. The resolution, or at least a partial resolution, was to come nearly half a century later, in the marriage of this very, very solid medium (very small, as August has indicated, but nonetheless very heavy as any librarian who has had to think about the load-bearing weight of the floor under the microfilm collection knows) with the much more ethereal medium of electronic information, the byte stream of the electronic text world.

Hearing this afternoon about the 'ancient' roots of microfilm, I am happy to be able to say that the roots of electronic text are almost as ancient. Arguably, it was in 1949 that the first electronic text project was born, when Father Roberto Busa contacted Thomas Watson, president of IBM, to explore the possibilities of

Watson's emerging new computational technologies for realizing the Jesuit's dream of a comprehensive linguistic study of the works of Saint Thomas Aquinas. Although it took until 1967, the project ultimately came to fruition. Along with the print publications generated by this project, you can find the database itself on CD-ROM in many research libraries today. Users can quickly recognize it as an artifact of an earlier era—you can practically see the punch cards—but it nonetheless excites respect as a monument of scholarly achievement (and continues to be a very handy way for tracing word usage in the extensive oeuvre of this medieval philosopher).[5] The year 1949, I should also note, was the year of my birth, and so I stand before you today as visible evidence that electronic text is indeed, sadly, no spring chicken.

However, as is true of most inventions, the early trajectory of e-text was long a slow, flat, barely rising line until a moment of takeoff, when the change in direction took such a rapid turn that the line of development began to approach the vertical. Despite Busa's early vision, it was really only in the mid-to-late 1980s that libraries began to appreciate the coming change that this new format for their textual resources was about to have. Moreover, for a few years still, it was a question of keyboarded text, much of it produced offshore to accommodate the enormous costs that such an approach implied, and the body of material that could be produced continued to be fairly limited. Significantly, too, major collections first appeared in appeared in areas with relatively narrow, focused bodies of source material—Classics, Medieval Studies, and Biblical Studies—fields that were also, not coincidentally, predisposed to appreciate close textual research.

From the beginning, there were hints of a similar interest in early American studies, in part, once again, probably because of the relatively more manageable body of sources involved—key early works included titles from the Library of America and early

[5] Thomas Aquinas and Roberto Busa, *Thomae Aquinatis Opera omnia cum hypertextibus in CD-ROM* (Milano: Editoria Elettronica Editel, 1992).

constitutional documents in the *WordCruncher* collection,[6] a few classic texts included in InteLex's *Past Masters* Series,[7] or the *Pennsylvania Gazette* and *African American Newspapers* produced by John Nagy at Accessible Archives in Philadelphia.[8] The enthusiastic reader response to the little world contained on the pages of the Accessible Archives products in particular gave some a first taste of what one might be able to do with a resource of this kind once some truly critical mass had been achieved. Suddenly, one could pull together disparate details to study such problems as crime in early America, the image and treatment of indentured servants, the sartorial codes of the American republic, or the commercial geography of the African American community in early New York.[9] Nonetheless, that realm of fuller kinds of textual riches seemed very far away, even as we began to see larger corpora entering the field as a result of the work of Chadwyck Healey, a company that, significantly enough, had made its original success as a microfilm company.[10]

[6] Electronic Text Corporation, *The WordCruncher Disc* (Orem, Utah, 1990).

[7] 'Since 1989, InteLex has been publishing the Past Masters® series of full-text humanities databases. We are focused on the needs of the scholarly community in the human sciences. InteLex assembles and publishes cohesive collections of excellent editions, in both original language and English translation, using meticulous text conversion processes. Combined with powerful CD-ROM- and web-based search and reference tools, the Past Masters series provides scholars with significantly enhanced and highly-flexible access to the classic texts' (www.nlx.com/pstm/index.htm).

[8] 'Accessible Archives was founded in 1990 with the goal of utilizing computer technology to make available vast quantities of archived historical information, previously furnished only on microfilm. In pursuit of this vision, primary source material has been selected to reflect a broad view of the times, and has been assembled into databases with a strict attention to detail allowing access to specific information with pinpoint accuracy. Our ON-LINE full text search capability and digital imaging permits the user to search and manipulate information in ways never before possible' (http://www.accessible.com/aboutus.htm).

[9] For some examples of these early studies, see T. K. Hunter, '"So that his Master may have him again": Pursuit of Indentured Property in the Pages of the *Pennsylvania Gazette*,' *Columbia Library Columns* (1997) 65:1 (1997):11–19, or Michael Zakim, 'What Is a Political History of Clothing?' *Columbia Library Columns* 65:33–39, and James Delbourgo, 'Electricity, Experiment and Enlightenment in Eighteenth-Century North America' (Ph. D. diss., Columbia University, 2003).

[10] 'Chadwyck-Healey™ is recognized worldwide as a leading source of high-quality publications in the humanities and social sciences. Founded in 1973 and acquired by ProQuest Information and Learning in 1999, Chadwyck-Healey is reputed as a leading publisher of innovative scholarly reference and full-text materials. Today Chadwyck-Healey™ continues the

Little did we realize that sweeping changes were just around the corner, as a result of new advances in scanning technology and the accumulated technological expertise and resource base of the great microfilm collections created in the preceding decades. Indeed, in just a few years, the floodgates have burst, as three major microfilm producers have rapidly moved to the conversion of those filmed page images into online text.

The first of these companies to enter the fray was ProQuest (formerly UMI), with its release of Early English Books Online, a digital rendering of its Early English Books microfilm set in 1999.[11] At the time, while convinced that Columbia had to acquire a digital collection of this magnitude and authority, I was frankly a little disappointed, having had a taste of the potential of searchable online text, to think that this resource would offer page images only. Putting them online would of course be a handy improvement, I suspected, but only a small step forward. I think my reaction was typical of most of my e-text colleagues. We underestimated the great leap forward that this change of venue would really make—the broad new circulation of texts that had heretofore been the province of a small group of dedicated scholars, putting them immediately before the whole academic community, available for undergraduate teaching and for use by

tradition of excellence in both content and the quality of their indexing and abstracting to provide students, scholars and teachers with extensive digital resources in the arts, humanities, reference and social sciences' (http:// www.proquest.co.uk/brands/CHbrand.html).

[11] 'From the first book printed in English by William Caxton, through the age of Spenser and Shakespeare and the tumult of the English Civil War, Early English Books Online (EEBO) will contain over 125,000 titles listed in Pollard and Redgrave's *Short-Title Catalogue* (1475–1640), Wing's *Short-Title Catalogue* (1641–1700), the Thomason Tracts (1640–1661), and the Early English Tract Supplement—all in full digital facsimile from the Early English Books microfilm collection.' EEBO—Early English Books on Line (http://eebo.chadwyck.com/home). To do full justice to the publishing history, one should note the role of an earlier company, Primary Source Microfilm, which subsequently became a component of Gale, in the early creation of these collections. In terms of our discussion, it also worth noting that the *Eighteenth Century* microfilm set on which *ECCO* was built was also based on a major bibliographic undertaking, the *Eighteenth-Century Short-Title Catalogue*, an electronic database now part of the broader online *English Short-Title Catalog*. To provide full orientation to the history of this process, one must also give full credit to ProQuest's predecessors, University Microfilms International and Bell and Howell, under whose names much of the early microfilm work was produced.

scholars for whom these texts had only been of secondary interest and hence, untouched in the microfilm area, available for browsing, searching, and exploration into areas where one would never have ventured before.

All the more revolutionary, then, was the impact when Readex and Gale[12] followed suit with their imprint catalogue-based microfilm sets, *Early American Imprints* and the *Eighteenth Century* respectively, collections that lent themselves to the delivery not only of page images but also of a not-perfect but often quite good rendering of the underlying text using a technique somewhat irreverently known as 'dirty ASCII' or 'uncorrected OCR,' in which powerful scanning programs were followed by quality control procedures but not proofread.

The first reaction of many librarians to the idea that the underlying searchable text included many errors was, particularly in the early days, often one of shock. Nonetheless, it usually took little effort to demonstrate to them that such an approach was not simply a necessary evil but a genuine good; that such an approach was utterly essential to making such a vast body of literature available to us now and at a price that any of us could dream of being able to pay. And lest one be misled by the words 'uncorrected' or 'dirty,' I recommend a trip to Readex's digital printing house in Chester, Vermont, for a chance to witness the impressive quality control applied at every step of the digitization process by the microfilm publishers to ensure that the text is as good as it can be.

As a result of the digitization of those major collections, accompanied by a number of supplementary ones as well, the scholarly community now possesses an unbelievably broad and, in some areas, nearly comprehensive corpus of the printed heritage of the

[12] 'Thomson Gale (www.gale.com), a business of The Thomson Corporation, is a world leader in e-research and educational publishing for libraries, schools and businesses. Best known for its accurate and authoritative reference content as well as its intelligent organization of full-text magazine and newspaper articles, the company creates and maintains more than 600 databases that are published online, in print and in microform' (http://www.gale.come/about/index.htm).

English-speaking world down to the early years of the nineteenth century. For American history and culture, the lion's share of this material is what has been produced by the collaboration of Readex and the American Antiquarian Society.

Recognizing the crucial importance of this material, Columbia has made it its aim to acquire as many of those collections as possible as I suspect have most of its peers. (Thanks to the firm commitment of our University administration to a growing library-materials budget, I think that in the case of the Readex material we are currently just one collection short, and are working on adding that one as well.)

These resources open up a breathtaking range of new possibilities for scholarship. One can quickly search out material on individuals or organizations that might have taken months or years of work in the past. One can do new kinds of close studies of language and ideas over time, examine the history of social groups, particular localities, customs, and commodities. With the array of data—books, newspapers, government documents, broadsides and ephemera of various kinds—there is an increasingly real possibility of working at reconstructing the past that comes closer than ever before to the literal meaning of those words. I anticipate that we will see in coming years a growing number of attempts to model the societies of the past and to explore closely their inner dynamics in ways that were never before possible. These ventures will no doubt begin with studies of earlier periods, where the relatively smaller amount of source material makes comprehensive coverage more of a realistic goal (and where the gaps to be covered by such reconstruction are also more in evidence).

Our scholars are increasingly aware of these possibilities, although I think we can fairly say that the scholarly community is still taking only its first tottering steps toward what will eventually be possible. Some recent encounters with Columbia users come to mind. For example, a professor who had used some of our first-generation electronic text tools in his research for a history of the concept of freedom in America, wistfully reflected to me not long ago about the good use he might have made of collections like the ones Readex has since produced, while noting at the same time the very real data management issues that such a vast array of information can present. Or the recently graduated Ph.D. student, who eagerly checked for updates of early American imprints to search for new material of relevance to her study of legal challenges to slavery in antebellum America. Or the scholar about to embark on a study of the role of statistic keeping on changes in the thought and culture of early nineteenth-century America, eagerly looking forward to the ways he will mine *Early American Imprints* and the *Congressional Serial Set* for useful data.

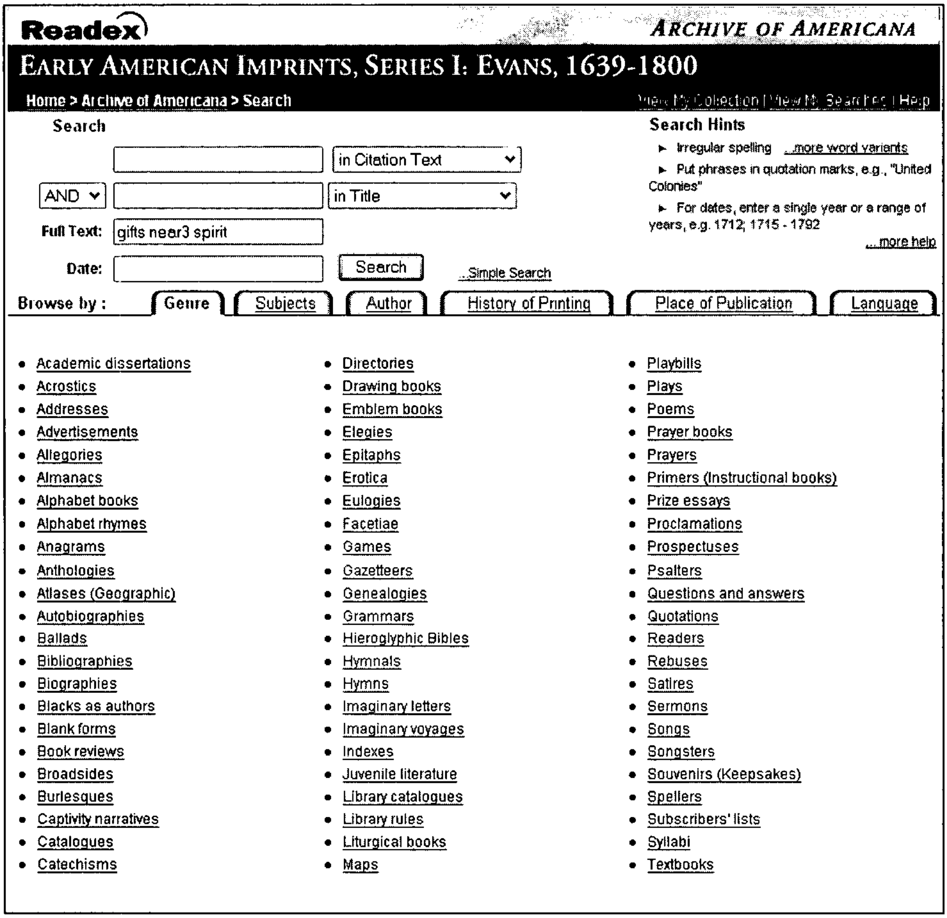
However, every great success inevitably imposes new tasks on the achievers. As the first of the scholars mentioned above noted, the very vastness of the feast before us constitutes a literal 'embarrassment of riches.' Thanks to creation of these enormous sets of information, we are often awash in data, desperately in need of tools to find our way to the best material deftly and effectively.

Navigational tools are needed to help even the seasoned scholar find his or her way across this great reservoir of data, to say nothing of enabling the incautious undergraduate visitor to sip safely and effectively from this figurative fire hose. And here the collaboration of Readex with the American Antiquarian Society and its gifted librarians has paid off well. The interface to the Early American Imprints collection, far more than any of its counterparts, is supplied with rich cataloguing—classification by genre, by subject—that can point the way to some of the most relevant material for their research (fig. 1).

However, cataloguing can only go so far. For many collections, such as *Early American Newspapers*, it is obviously not practical or realistic to expect such detailed description of countless individual entries. It would be an unending task to create comparable headings for a growing collection of that kind. For such resources, and equally as much for collections where better cataloguing is available, finding increasingly better and better tools for mining, navigating, and working with the full text is the real key. Readex (as well as ProQuest and Gale) is certainly aware of the developmental agenda that lies before us, and I can say from my own experience of working with them that they are impressively committed to an ongoing dialogue with their user community to try to find better ways to make all of this work more effectively. Not long ago, for example, we at Columbia and a number of other institutions were engaged in discussions with Readex about how best to cross search the whole collection, and still more recently we have been asked to consult on the best ways of providing catalogue access. Yet another expression of Readex's commitment to interaction with its scholarly public has been its recently inaugurated series of annual Digital Seminars, which are designed to bring together individuals involved in digital research from various disciplines and professions to discuss the common issues we face and to think seriously about the types of new resources and tools we may need. [13]

Readex has listened to the needs expressed in those various consultations, with a keen sensitivity to scholarly needs that no doubt owes much to its long collaboration with the American Antiquarian Society. The interface it has provided to its collections includes many features designed to make these resources function most effectively for their users—highlighted search terms on the page image, for example, which you can see in the accompanying printout (fig. 2; note the search term 'rattle-snake' in the right-hand column, somewhat less visible in a black-and-white presentation than in the actual screen display, where color immediately draws the eye to the term)—something still not available in many other important page-image collections. The navigation system on the left is also a very helpful means of jumping into the text, making pages with hits immediately available while enabling easy browsing of adjacent areas of the text as well.

[13] See http://www.readex.com/features/digitalist.html.


Readex
ARCHIVE OF AMERICANA
EARLY AMERICAN IMPRINTS, SERIES I: EVANS, 1639-1800
Home > Archive of Americana > Search
View My Collection | View My Searches | Help
Search
Search
..Simple Search
Browse by :
Genre
Subjects
Author
History of Printing
Place of Publication
Language

Search: [ ] in Citation Text

AND [ ] in Title

Full Text: gifts near3 spirit

Date: [ ]

Search Hints

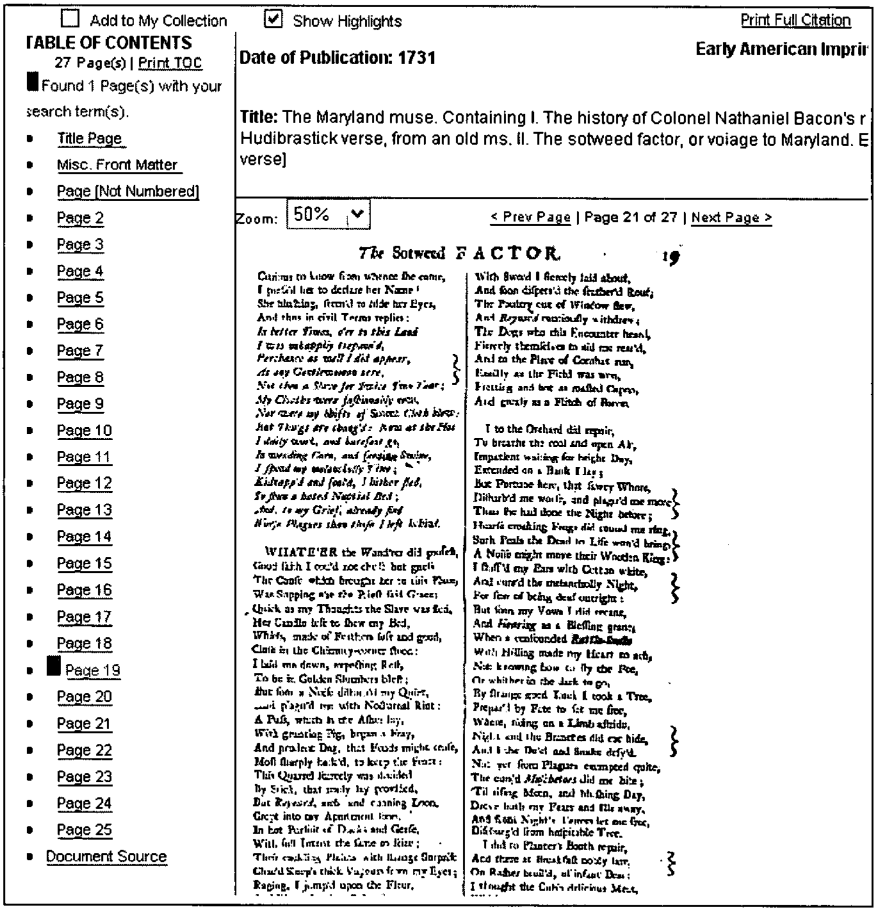
- Irregular spelling ...more word variants
- Put phrases in quotation marks, e.g., "United Colonies"
- For dates, enter a single year or a range of years, e.g. 1712; 1715 - 1792

...more help

- Academic dissertations
- Acrostics
- Addresses
- Advertisements
- Allegories
- Almanacs
- Alphabet books
- Alphabet rhymes
- Anagrams
- Anthologies
- Atlases (Geographic)
- Autobiographies
- Ballads
- Bibliographies
- Biographies
- Blacks as authors
- Blank forms
- Book reviews
- Broadsides
- Burlesques
- Captivity narratives
- Catalogues
- Catechisms
- Directories
- Drawing books
- Emblem books
- Elegies
- Epitaphs
- Erotica
- Eulogies
- Facetiae
- Games
- Gazetteers
- Genealogies
- Grammars
- Hieroglyphic Bibles
- Hymnals
- Hymns
- Imaginary letters
- Imaginary voyages
- Indexes
- Juvenile literature
- Library catalogues
- Library rules
- Liturgical books
- Maps
- Playbills
- Plays
- Poems
- Prayer books
- Prayers
- Primers (Instructional books)
- Prize essays
- Proclamations
- Prospectuses
- Psalters
- Questions and answers
- Quotations
- Readers
- Rebuses
- Satires
- Sermons
- Songs
- Songsters
- Souvenirs (Keepsakes)
- Spellers
- Subscribers' lists
- Syllabi
- Textbooks

Figure 1. Advanced Search Screen for Early American Imprints. Note the many genre categories and tabs enabling limits by a variety of other metadata categories, as well as the proximity search statement in the Full Text search window.

Yet another admirable feature of this interface is its flexible and many-featured search syntax. An area where special praise is due is Readex's explicit encouragement of the use of proximity operators (for example, syntax permitting searches for terms within a specified distance of one another). Proximity operators are absolutely



☐ Add to My Collection ☑ Show Highlights Print Full Citation

TABLE OF CONTENTS

27 Page(s) | Print TOC

▇ Found 1 Page(s) with your search term(s).

- Title Page
- Misc. Front Matter
- Page [Not Numbered]
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8
- Page 9
- Page 10
- Page 11
- Page 12
- Page 13
- Page 14
- Page 15
- Page 16
- Page 17
- Page 18
- ▇ Page 19
- Page 20
- Page 21
- Page 22
- Page 23
- Page 24
- Page 25
- Document Source

Date of Publication: 1731

Early American Imprir

Title: The Maryland muse. Containing I. The history of Colonel Nathaniel Bacon's r
Hudibrastick verse, from an old ms. II. The sotweed factor, or voiage to Maryland. E
verse]

Zoom: 50% ▾

< Prev Page | Page 21 of 27 | Next Page >

The Sotweed FACTOR. 19

Curious to know from whence ſhe came,
I preſs'd her to declare her Name:
She bluſhing, ſeem'd to hide her Eyes,
And thus in civil Terms replies:
In better Times, e'er to this Land
I was unhappily trepann'd,
Perchance as well I did appear,
As any Gentlewoman here,
Not then a Slave for twice ſive Year;
My Cloaths were faſhionably new,
Nor were my Shifts of Scotch *Cloth blew;*
But Things are chang'd: Now at the Hoe
I daily work, and barefoot go,
In weeding Corn, and feeding Swine,
I ſpend my melancholy Time;
Kidnapp'd and fool'd, I hither fled,
To ſhun a hated Nuptial Bed;
And, to my Grief, already find
Worſe Plagues than thoſe I left behind.

WHATE'ER the *Wand'rer* did profeſs,
Good-faith I cou'd not chooſe but gueſs
The Cauſe which brought her to this Place,
Was Supping e'er the Prieſt ſaid Grace;
Quick as my Thoughts the Slave was fled,
Her Candle left to ſhew my Bed,
Which, made of Feathers ſoft and good,
Cloſe in the Chimney-corner ſtood:
I laid me down, expecting Reſt,
To be in Golden Slumbers bleſt;
But ſoon a Noiſe diſturb'd my Quiet,
And plagu'd me with Nocturnal Riot:
A Puſs, which in the Aſhes lay,
With grunting Pig, began a Fray,
And prudent Dog, that Feuds might ceaſe,
Moſt ſharply bark'd, to keep the Peace:
This Quarrel ſcarcely was decided
By Stick, that ready lay provided,
But *Reynard*, arch and cunning Loon,
Crept into my Apartment ſoon,
In hot Purſuit of Ducks and Geeſe,
With full Intent the ſame to ſeize;
Their cackling Plaints with ſtrange Surpriſe
Chas'd Sleep's thick Vapours from my Eyes;
Raging, I jump'd upon the Floor,

With Sword I fiercely laid about,
And ſoon diſpers'd the feather'd Rout;
The Poultry out of Window flew,
And *Reynard* cautiouſly withdrew;
The Dogs who this Encounter heard,
Fiercely themſelves to aid me rear'd,
And to the Place of Combat run,
Exactly as the Field was won,
Fretting and hot as roaſted Capon,
And greaſy as a Flitch of Bacon.

I to the Orchard did repair,
To breathe the cool and open Air,
Impatient waiting for bright Day,
Extended on a Bank I lay;
But Fortune here, that ſaucy Whore,
Diſturb'd me worſe, and plagu'd me more
Than ſhe had done the Night before;
Hoarſe croaking Frogs did 'round me ring,
Such Peals the Dead to Life wou'd bring,
A Noiſe might move their Wooden King:
I ſtuff'd my Ears with Cotton white,
And curs'd the melancholly Night,
For fear of being deaf outright:
But ſoon my Vows I did recant,
And *Hearing* as a Bleſſing grant;
When a confounded *Rattle-Snake*
With Hiſſing made my Heart to ach,
Not knowing how to fly the Foe,
Or whither in the dark to go,
By ſtrange good Luck I took a Tree,
Prepar'd by Fate to ſet me free,
Where, riding on a Limb aſtride,
Night and the Branches did me hide,
And I the Devil and Snake defy'd.
Not yet from Plagues exempted quite,
The curs'd *Muſketoes* did me bite;
'Til riſing Morn, and bluſhing Day,
Drove both my Fears and Ills away,
And from Night's Errors ſet me free,
Diſcharg'd from hoſpitable Tree.
I did to Planter's Booth repair,
And there at Breakfaſt nobly fare,
On Raſher broil'd, of infant Bear:
I thought the Cub delicious Meat,

Figure 2. Result Page from Early American Imprints Online. Note highlighted search term, navigational menu at left, and the difficult typography that the OCR engine had to read.

essential for mining full-text data, but Readex is the only publisher I have seen to emphasize this need in its help guides. (Most database providers do not even want you to think about proximity operators, because they impose a lot of work on the search engine.) Readex also provides admirable guidance to readers for working with the 'dirty ASCII' and variously spelled material it contains, and backs it up with excellent wildcard search capabilities.

Certainly much still remains to be done in terms of search capability. This should not be seen as criticisms, but simply the recognition of the new challenges and opportunities that leap up as we move ahead with this rapidly developing medium. These same challenges face all of the electronic publishers, as well as the electronic text research community as a whole.

First, with all of the data these projects are bringing online, it is critical to ensure sufficient server power to permit quick and easy use. I am sure that they worry up in Chester every day about the computing power that is going to lift all of this up.[14] I think I find I do my best searching of all of these publishers' databases at three o'clock in the morning—I had a chance when I got up this morning to take the bus to Worcester to test that out again—when everybody else in the country is still asleep and the folks in California are not even thinking about getting up. Making sure that we can rapidly process these massive collections is essential if we are really to make use of them, and I think that this is really a community issue. As the body of material available grows, I suspect that we will need to look increasingly to collaborative server options, to multiple sites, and to shared hosting.

There are two other very important ways in which users can be assisted in moving quickly to the texts they need. One is a little more daunting than the other, but both are ultimately doable. The first has to do with the multiplicity of spellings found in early texts. Even the most advanced researcher is unlikely to know, or at least be able to remember off the top of her head, the various ways in which a word might be rendered in the works of that period. Add to that, for the Readex and Gale texts, with underlying uncorrected OCR, the possibility that the electronic text version

[14] It should be noted that at the time when the symposium was held, Readex was implementing a new, faster set of search software, testimony to the continuing challenge to provide ever more powerful engines for accessing growing online collections.

of a word is slightly misspelled. Various strategies have been proposed to date. As already noted, Readex offers some very good guidance here, and robust wildcard syntax to support searching for variant forms, although this largely assumes that the searcher already knows what the variant forms are likely to be. ProQuest has introduced a feature that allows one to search for typographical variants, but this will not usually capture true orthographic variants, but rather instances where one letter is substituted for another (say, a 'v' for a 'u'). Fuzzy searching is employed by Gale and ProQuest as a way of capturing variant forms (and one should single Gale out for praise for providing users an option of turning that feature on or off). Index searches are another possible strategy, too, but what is really needed here is a mechanism that would enable a searcher to type in a modern spelling and receive all variants (unless that was not what he wanted). My own personal vision of the solution here, informed by the new collaborative environment around so much of this early English-language corpus-building, would be a joint approach by the publishers and the Text Creation Partnership (described a little later in this paper) to Oxford University Press to somehow license (and then harness) the authoritative variant spelling information that their *Oxford English Dictionary Online* contains.

The other useful step here is a surprisingly easy one, but one that almost no commercial electronic publisher has adopted (one of the few exceptions being Alexander Street Press).[15] That would

[15] 'Alexander Street Press brings together the skills of traditional publishing, librarianship, and software development to create quality electronic collections. We believe that an electronic publication should be carefully crafted by expert editors around a specific subject or discipline; detail all materials relevant to the subject, whatever their original form or ownership; contain as many of these materials as possible, in multiple formats if necessary; be indexed with controlled vocabularies for precise, exhaustive searching; provide unique ways of searching, viewing, exploring and analyzing the material; facilitate contributions from scholars and librarians; and be priced to enable unlimited exploration by users. We are creating a series of products using these values.' As another footnote to electronic publishing history, it is worth noting that the leadership of this company, in personal terms, represents a continuation of the original Chadwyck Healey North America (http://www.alexanderstreetpress.com/about/index.htm).

be to display the results of a search immediately in a standard, concordance-like, 'keyword-in-context' format (often called KWIC). Instead, the typical text database lists the titles that contain a search term and perhaps the number of hits they contain. To see if the works contain a relevant hit, users must open the text and sometimes take yet another step before seeing exactly what it is that their search has found. When one is dealing with vast collections, that is almost a guarantee that many, if not most searchers, will fail to examine all their results. A KWIC display would allow users to quickly zero in on the passages of relevance, increasing the effectiveness of their research enormously.

Presenting a KWIC display would also inevitably force 'uncorrected OCR' publishers to take another courageous step, namely, to show the 'dirty ASCII.' Their reluctance to do so is understandable. As I suggested earlier, a naïve user is likely to be shocked at first glance to see that the text underneath has a number of errors, sometimes many errors in it. However, as I have also suggested, serious researchers quickly realize upon reflection that this approach is the very thing that has made the existence of such a resource possible. Indeed, I would argue that ultimately it would be good if publishers were willing to make the full 'dirty ASCII' text visible (albeit a couple of clicks away in order not to horrify the uninitiated). Knowing what is 'under the hood' can only help sophisticated researchers to make better use of these collections

There is also a need for even more search guides of the type toward which Readex's exemplary help-screens point the way. These are complex collections, with texts whose language, orthography, and modes of expression are very different from those of today and there is much room for suggesting appropriate strategies for enabling users to make the best use of this material. Indeed, it would be good to enlist the services of scholars at the cutting edge to describe some of the truly advanced uses to which these resources may be put, in the interests of pushing us all further forward. Even more, perhaps, there is a need for guides to help less experienced users—interested members of the general public, undergraduates or even high-school students—for all of these collections, particularly the ones that Readex produces, if properly presented, have an enormous potential for enriching the educational process at all levels in this country.

And this brings us to an even more burning issue, that of the digital divide. These are expensive collections to build and maintain. An enormous investment in able staff, technology, production, quality control, and bibliographic description has gone into their making. In the days of microfilm collection building, the financial model was clear. A fairly limited community, the ARL libraries and a few other key institutions, would be able to purchase these sets, and others who needed to use this material would find their way, one way or another, to the institutions that held them on card, film, or fiche. The pricing of these collections was established accordingly. Once the collections had been sold to the usual suspects, there would probably not be another great body of potential customers. The investment had to be recouped in those first sales, and the value that was received by the institutions that could afford it made that price acceptable to them. With the advent of the web, however, the paradigm has changed. Traveling across the city or across the state to access an otherwise inaccessible title no longer has the appeal it once did for a generation accustomed to find what it wants just a couple of clicks away. We librarians working in the humanities and history constantly fret about the difficulty of getting today's undergraduate to accept that the best journal article needed for their research is not necessarily the one online but another just a few steps away in the stacks. Moreover, I would argue that the electronic format also ultimately makes possible less of an all-or-nothing approach to the product than had to be characteristic of microform.

These are issues that scholars, librarians, and publishers need to work together to resolve, recognizing that we must all proceed carefully to ensure that the undertaking continues to be commercially feasible for the companies that can make this happen. I know that all of the publishers have sharply tiered pricing schemes, depending on the size of user community and acquisition budget of acquiring institutions, but I think I speak for many of my ARL colleagues in the development of digital collections that we would welcome even more aggressive efforts to ensure that these key materials are put into the hands of as many students and researchers as possible. The successful use of these collections at major research institutions will be all the more successful and quick to develop if the students who come to us are already familiar with these resources and accustomed to using them. Moreover, in the wide-open world of the web and in the face of freely accessible collections such as Google BookSearch, there is a real danger that better-quality text may be pushed aside by lower-quality but freely accessible material. A great deal has been accomplished in these publishing ventures, and it is of critical interest to scholarship that that forward position be preserved.

Hence, I would strongly urge the consideration of some new approaches by all three of these publishers. One strategy, for example, might be smaller, very cheaply priced subsets of the main databases, designed for the public school, small public library, or community college markets, which have limited resources for acquisitions, and whose users in any case might find the vast, multi-edition holdings of the full corpora too rich a meal. Another tactic might be to consider individual researcher subscriptions. I was recently speaking with one of our alumni, who now teaches at an institution in another country. He described how precious his periodic visits to Columbia were, since they enabled him to take advantage of the newest additions to our online Americana holdings, but noted how much more productive his research would be if he had regular access to this material. His institution may never have the means or motivation to acquire them for all of its users, he said, but every year he receives individual research funding that he would eagerly use to subscribe to Readex's databases. Finally, I would urge consideration of the sale of individual texts in the collection to individual scholars, as a move toward a more granular approach that may be ultimately the best strategy for all partners in the changing world of electronic publishing, making possible revenue streams that are not solely dependent on the large institutions which will continue to want to acquire large databases, that will enable the continued coexistence of smaller, specialized niche publishers alongside the larger companies, and will enable libraries to begin looking at their collective acquisitions of electronic text material as so many books alongside one another on the same electronic shelf, regardless of origin ultimately lending themselves to collective treatment using search engines and conceivably other tools of analysis .

That latter vision raises one other issue that must concern us all, and that is the need to ensure compatibility of standards—standards for metadata, standards for the markup of the underlying electronic text that can ultimately make possible the kind of collective treatment I have just described.

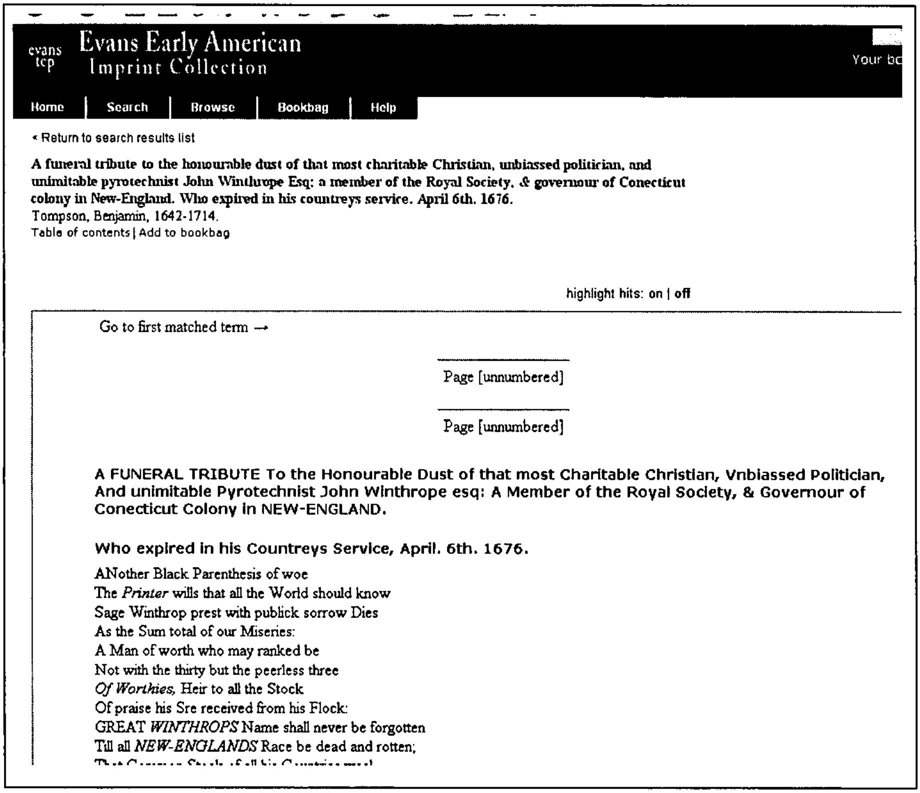
And this brings me to a final point of great relevance to our discussion today, to an admirable collective undertaking of the library and electronic publishing community that seeks to address many of the challenges I have raised above. It is the Text Creation Partnership, a collaborative venture of academic and research libraries and the scholarly publishing community aiming 'to support the creation of accurately keyboarded and encoded editions of thousands of culturally significant works in all fields of scholarly and artistic endeavor.'[16]

[16] The aims of the partnership are further elaborated on its website, located at http://www.lib.umich.edu/tcp, which also offers a link to the demonstration database described a little later in this paper: 'The underlying principles of the TCP are mindful of the long-term needs of libraries, scholars and the larger society. TCP projects are notable for the quality and cost-effectiveness of their content, as well as for the underlying principles of the Partnership that: convey robust rights of use to scholars; protect the public domain rights of the larger society to access out-of-copyright materials; present the user with accurately keyed, modern texts that are faithful to the spellings and organization of the original works; ensure that this content will migrate forward through shifts in technology to represent editions of enduring value to libraries. The net effect of the TCP initiatives has been to maximize the respective strengths of commercial and academic digital library development for the long-term benefit of researchers and students.'

The original inspiration for its creation came with the release of Early English Books Online, which, as I have noted, was understandably limited to the delivery of page images of that vast collection of pre-1700 English imprints. A number of the acquiring libraries, eager to see those images supplemented with full text, formed a partnership with the publisher to underwrite the full keyboarding of a subset of twenty-five thousand titles in the collection. In drawing up the standards for that text creation, they adopted the guidelines of the Text Encoding Initiative, a consortium promoting the implementation of a uniform standard of markup for scholarly texts.[17] The texts themselves, as they are created, are incorporated into the original EEBO database, but the TCP also maintains its own searchable collection (generally containing a little more text than EEBO itself, since the newly created text is added only at selected intervals.). At the completion of the project, the partners will be full owners of this material to make available as they see fit.

So successful was this undertaking, that when the Readex and Gale collections became available, participants immediately recognized the value of producing fully keyboarded and marked-up subsets to supplement the other 'dirty ASCII' text that they contained, enabling scholars to use them as fully accurate material for scholarly research. The new Evans partnership, involving the same kind of collaboration between library partners and Readex as in the EEBO partnership, is already well underway. Ultimately, the plans call for a full keyboarding of some six thousand texts from *Early American Imprints Part I.* Not surprisingly, when the partnership sought to come up with a way of selecting the texts to be included, it turned to the American Antiquarian Society, which arguably knows this collection better than anyone else, and the Society has cheerfully taken up the task. Digitization is well underway, and a small test database is already available. I offer here, as an example, a screen from *A Funeral Tribute to . . . John*

[17] For a full description of the TEI consortium's activities and goals see its website at http://www.tei-c.org.



« Return to search results list

A funeral tribute to the honourable dust of that most charitable Christian, unbiassed politician, and unimitable pyrotechnist John Winthrope Esq: a member of the Royal Society, & governour of Conecticut colony in New-England. Who expired in his countreys service. April 6th. 1676.
Tompson, Benjamin, 1642-1714.
Table of contents | Add to bookbag

highlight hits: on | off

Go to first matched term →

Page [unnumbered]

Page [unnumbered]

A FUNERAL TRIBUTE To the Honourable Dust of that most Charitable Christian, Vnbiassed Politician, And unimitable Pyrotechnist John Winthrope esq: A Member of the Royal Society, & Governour of Conecticut Colony in NEW-ENGLAND.

Who expired in his Countreys Service, April. 6th. 1676.

ANother Black Parenthesis of woe
The *Printer* wills that all the World should know
Sage Winthrop prest with publick sorrow Dies
As the Sum total of our Miseries:
A Man of worth who may ranked be
Not with the thirty but the peerless three
Of Worthies, Heir to all the Stock
Of praise his Sre received from his Flock:
GREAT *WINTHROPS* Name shall never be forgotten
Till all *NEW-ENGLANDS* Race be dead and rotten;

Fig. 3. Result page from the Evans TCP Demonstration Database. Note the fully keyed text and link to page image.

Winthrope. . ., one of the texts resulting from a search for the word 'peerless.' What you see on the screen here is fully keyboarded text. As you can see, the inputters have striven to match important changes in the typography. Not visible here, but present behind the scenes, is TEI markup identifying the key features of the text and sometimes providing the basis for more focused searching, say for captions with the word 'Jefferson' in them, as one way of searching for pictures of Thomas Jefferson. A link on the page permits one to jump back to the original page image as well.

The Eighteenth Century Collections Online project, a collaboration with Gale, aiming to rekey some ten thousand texts from that collection, is in its very early stages. When the three current projects are completed, they will have produced a collective searchable corpus of more than forty thousand of the most important works produced in the English language from the fifteenth through the eighteenth centuries, with uniform markup of structure and major features making possible a variety of research strategies. It is to be hoped, moreover, that this database can become the nucleus of a much broader undertaking to create a common English-language digital library, with other companies, research institutions, and even individual scholars contributing their work in the creation of a single resource. The materials currently being produced are being added to all three of the original databases to enhance their functionality, but at the end of the process they will also be the property of the academic partners, and I feel confident that ways will be ultimately found to make this material available to a much broader community.

The effective, creative collaboration of academics and publishers working together in a broad partnership represented by the TCP clearly grows out of the kind of earlier partnership so well exemplified by the collaboration of Readex and the American Antiquarian Society. It vividly reflects the shared values that such a collaboration can foster.

Earlier today, I had an opportunity to view the very interesting exhibition here documenting the history of the American Antiquarian Society. I would like to suggest that the display might be improved by the addition of one or two more images documenting the latest phase in the Society's impressive story: the creation of the electronic resources that proved only possible because of the impressive work done in the last century to create a pioneering microfilm collection, and that may do more than anything else to bring the richness of this organization's collections to the whole American scholarly community.