



EVANS EARLY AMERICAN IMPRINT COLLECTION

Text Creation Partnership

ESSENTIAL CONTENT

Charles Evans' American Bibliography, covering the years 1639 to 1800, is without question the most important source for the study of 17th and 18th century America. The University of Michigan, NewsBank/Readex Co., and the American Antiquarian Society are cooperating in a Text Creation Partnership that aspires to create 6,000 fully searchable SGML/XML text editions out of the approximately 36,000 titles represented in the Evans Digital collection. Selected works will cover a broad range of subjects and genres and will include such notable authors as:

History

- George Washington
- John Hancock
- John Locke
- John Adams
- Alexander Hamilton
- Samuel Adams
- James Madison
- John Marshall

Literature

- Benjamin Franklin
- Mary Rowlandson
- Jonathan Swift
- Phillis Wheatley
- Daniel Defoe
- Anne Bradstreet
- Samuel Johnson
- William Bradford

Social Life

- Robert Fulton
- William Penn
- Isaac Watts
- Jonathan Edwards
- Samuel Davies
- Cotton Mather
- Increase Mather
- Thomas Paine

EVANS TEXT CREATION PARTNERSHIP

Evans Digital subscribers are encouraged to support a cooperative library effort to convert a selected subset of the corpus as accurately keyboarded (99.995 standard) and SGML/XML tagged text. The Text Creation Partnership (TCP) has reached an agreement with Newsbank/Readex to use their images to create this accurately keyed and tagged subset of Evans. With adequate community support, we anticipate converting 6,000 texts over a five-year period, ending December 2007. The American Antiquarian Society has agreed to work with the TCP to coordinate the selection process for the keyboarded texts. The University of Michigan Digital Library Services will initially shape the production process based on the proven standards and procedures established for the Early English Books Online-Text Creation Partnership. As with other TCP initiatives, partner libraries will co-own the texts at the completion of the project (and allowing a reasonable sales window for Readex) with very robust rights of use, adaptation, and distribution.

ACCURATELY KEYED AND ENCODED TEXTS

It goes without saying that the Evans content is top-notch and of enduring scholarly value and interest. Return to any of our campuses in two hundred years and it is guaranteed that you'll find scholars and students working with this corpus. Evans images will be accessible by fully searchable bibliographic records and OCR running behind the page images. The reliability of the OCR for searching purposes remains an open question because of the nature of the images being used. Newsbank/Readex is doing their best of course, to optimize OCR accuracy, and report positive results in the early going. Nonetheless accurate OCR on materials of this nature is not easily accomplished, and even when the average results are good, it is bound to be uneven (very good sections or texts and some very poor texts). It is recognized that OCR provides cost effective access to a large corpus like Evans the library community stands ready to support efforts at Newsbank/Readex. That being said, we also see benefit in creating an accurately keyboarded and SGML/XML tagged subset of Evans text that allows more certain searching and other features that fully support research and instructional uses of this important corpus.

Libraries are very aware of how many times we've paid for this content in different formats. Therefore, we should explain what value-added we get from keyed and tagged text. In the OCR world-- and let's assume for the moment good and even OCR for searching purposes—a user enters a keyword search term and the system returns a listing of titles that contain matches or “hits.” The better systems (like

ACCURATELY KEYED AND ENCODED TEXTS

Evans Digital) return the pages where the hits can be found. You can click on the page numbers to go to an image of that page and then look for your keyword. Some systems will not take you directly to your page but to the beginning of an article, chapter, section or book and the reader then has to browse through multiple pages looking for their search term. Most OCR based products will not show the text they are searching because the prevalence of errors is thought to be distracting to users.

Compared to the OCR presentation, (which we recognize has been used to great advantage in projects like JSTOR and Making of America), accurately keyed and tagged text allows for the following benefits for users:

- The texts can be presented and read—and reading the text version is often easier for students than reading original page images of early printed books. The text can be cut, pasted, edited and manipulated.
- Search and retrieval is much improved, both because it is accurate and because searching can be limited to tagged elements like titles or verses or notes.
- Search results are returned not just as a page number but with a line or two of text to give context and help the user select appropriate hits among a large number of returns.
- Search terms are highlighted and easily located within the text
- The tagging of titles, chapter and section heads creates a browsable table of contents for every book with the ability to link to appropriate sections of the book.
- Accurate and tagged texts can be used as a basis for creating new and enhanced scholarly editions.
- Word indexing allows the user a rudimentary way to find or eliminate spelling variants.
- Standards based—production of these texts—both for character accuracy and mark-up— give the texts enduring value with assurance that they will migrate forward through new systems and be accessible via multiple platforms.

COST EFFECTIVENESS

Nobody should question that accurately keyed and tagged text adds value, but it is legitimate to ask *how much value* in an environment where resources are limited. If any individual institution attempted to keyboard a few volumes for a course or research project, it might cost between \$500 to \$1,000 each—over and above the investment already made in the commercially distributed image product. Generally speaking, most academic institutions could not absorb those costs for more than a few volumes. The Text Creation Partnership model, however, offers a very cost effective way to accomplish some important goals for our community.

1. In the context of the Text Creation Partnership, keyed and tagged texts are costing individual partner institutions between \$2 and \$6 each.
2. Partner institutions own the texts in perpetuity, and in the case of both EEBO and Evans, we have successfully negotiated for institutions to distribute the texts beyond their campus community if they should so choose. In other words, the keyboarded text will return to the public domain after a reasonable sales window for the vendor is met.
3. The texts from EEBO and Evans will be produced to a single standard and the files can be combined into a single searchable entity where hits on text pages would link to image counterparts delivered respectively from the Readex or ProQuest servers.

Based on the experience of EEBO, and having now laid the groundwork for an Evans project, we are assured that it offers libraries a cost effective acquisition of fully owned, very functional, high quality editions of culturally significant works. The editions cost a few dollars each, and the nation's research libraries—especially the publicly funded libraries—are protecting the access rights of communities beyond their authenticated campus users.

WEBSITE

<http://www.lib.umich.edu/evans/>