

Academic and Commercial Roles in Building “The Digital Library”

Mark Sandler

SUMMARY. The University of Michigan and the University of Oxford have encouraged the research library community to fund creation of structured SGML/XML text-files for a significant portion of the Early English Books Online (EEBO) corpus of digital images created by ProQuest Information and Learning Company. These full-text editions, linked to the corresponding digital facsimiles of the works, enable word or phrase searching across the corpus along with the display of the corrected modern texts for reading and editing. Creating thousands of these text editions will significantly extend intellectual access to this historically important content, making possible new avenues of research across a broad range of disciplines. The significance of the EEBO Text Creation Partnership is to be judged not only by the usefulness of resulting product but also by the success of the business plan under which the text editions have been created. These successful experiences have not yet sufficiently raised consciousness among libraries and consortia—even those participating in the EEBO-TCP—that would cause them to question lower quality production methods and less library- and user-friendly models still being promulgated by the vendor community. [Article copies available for a fee from *The Haworth Document Delivery Service*: 1-800-HAWORTH. E-mail address: <docdelivery@haworthpress.com> Website: <<http://www.HaworthPress.com>> © 2003 by The Haworth Press, Inc. All rights reserved.]

Mark Sandler is Collection Development Officer, University of Michigan, University Library, 205 Hatcher North, Ann Arbor, MI 48019 (E-mail: sandler@umich.edu).

[Haworth co-indexing entry note]: “Academic and Commercial Roles in Building ‘The Digital Library.’ ” Sandler, Mark. Co-published simultaneously in *Collection Management* (The Haworth Information Press, an imprint of The Haworth Press, Inc.) Vol. 28, No. 1/2, 2003, pp. 107-119; and: *The New Dynamics and Economics of Cooperative Collection Development* (ed: Edward Shreeves) The Haworth Information Press, an imprint of The Haworth Press, Inc., 2003, pp. 107-119. Single or multiple copies of this article are available for a fee from The Haworth Document Delivery Service [1-800-HAWORTH, 9:00 a.m. - 5:00 p.m. (EST). E-mail address: docdelivery@haworthpress.com].

<http://www.haworthpress.com/web/COL>

© 2003 by The Haworth Press, Inc. All rights reserved.
Digital Object Identifier: 10.1300/J105v28n01_09

107

KEYWORDS. Digital libraries, Early English Books Online (EEBO), cooperative collection development, ProQuest, electronic publishing–cooperation

I find it interesting that so much of the talk in academic circles of “*building the digital library*” excludes the role of the commercial sector in the production of content to populate its virtual shelves. When we were all about building print libraries, it was understood that commercially motivated publishers supplied us with content. With the switch to electronic resources, however, it seems that many of our colleagues think that the digital library of the future will be the sum of small scale, high cost conversion efforts to capture the images of Wisconsin circuses, a few issues of an early bicycle magazine, a couple of volumes of Scottish or American verse, the work a 16th century Italian playwright (richly tagged) and many other odds and ends from hundreds of grant-funded projects. While we try to keep up with these myriad digital projects—and try with even less success to keep our readers abreast of what is available to them through elaborate schemes of metadata creation and harvesting and registries—we seem to be denying the point that the real digital library, like the print library before it, is being created by Elsevier, the scholarly societies, JSTOR, ISI, Gale, ProQuest (including Chadwyck Healey), netLibrary, etc. While there are exceptions, a sober assessment suggests that our federations, forums, and grant driven initiatives are simply nibbling at the margins of a larger digital library. The scaleable projects, the important content, the current publications—i.e., the resources with “legs”—are most likely being produced by commercial publishers with one eye on user needs and the other on the bottom line. So, if this assessment is close to reality, what becomes the role of the research library community in shaping the digital library?

THE EEBO EXPERIENCE

Before attempting to answer the question of the role of research libraries, let me establish my credentials as someone who has had the opportunity to gaze upon the golden calf of commercially produced content. In 1998, ProQuest Company (then Bell & Howell) brought to market digital images of approximately 100,000 early English texts that many research libraries already held in microfilm (STC I, STC II and the Thomason Tracts). Titled Early English Books Online (EEBO), the

corpus was significant both because of the formative place of the content in western heritage as well as for the critical mass of material involved. As measured in bytes (or terabytes) EEBO was the largest digital project ever brought to the library market. As such, the overall cost was high, and at a time when budgets were being squeezed by a slackening economy and heightened demands for access to current electronic journals in science, technology and medicine. How would the library market react to such a massive project as that undertaken by ProQuest?

To keep a long story short, we can summarize the reaction of the library market as three-fold: some libraries had money (including some surprisingly small libraries that had year-end or other special funds) and purchased EEBO early on with little fuss. More still thought the price tag was too steep to warrant consideration. And, a third group began pushing for enhancements to the product as a condition of purchase and to gain greater control of the content through participating in this process of enhancement. It's primarily this latter response that I want to discuss in this paper but it could be helpful to say a few words about the early adopters and early naysayers of new library content.

TO BUY OR NOT TO BUY

Large purchase decisions in libraries are not always a rational process of assessing the match between needs and resources. Rather, decisions are often driven by a single—and not always informed—advocate (faculty or librarian). Sometimes it is because support for early purchase can be the result of a relationship to content, a relationship to a company, or even an individual sales representative. Sometimes it is because of a desire to do something (read *anything*) for a particular discipline or department. Sometimes it is because someone else bought it, or maybe even because someone else didn't buy it. My point is that not *every* library—or perhaps it would be more correct to say *any* library for *every* purchase—completely and carefully evaluates both the content of a new collection (regardless of format) and the actual quality and effectiveness of the representation of that content. “Pre-pub” or early adoptions reflect (or should reflect) confidence in the publisher, and a reasonable expectation based on past performance, that projects underway have a high likelihood of success.

In the area of humanities text corpora, Chadwyck-Healey built numerous collections on pre-pub investments because customers knew more or less what they would be getting—or at least that they would be

getting *something*. With some of the newer players in this market space—and that would have included ProQuest when EEBO was being developed—it is unclear that one could have a very clear sense of their capacity to deliver high quality digital images that could be called forth in real time by a sufficiently sensitive and user-friendly search engine. In any case, my larger point here is that some libraries on some days purchase resources—even those involving very large commitments—with little actual scrutiny.

Decisions *not* to purchase large collections can be equally casual. In many libraries, and among many consortia, it is simply easier to pursue smaller, bite-sized purchases. Often, it might seem a better use of limited resources to buy several small collections that would support myriad constituencies than to invest in one “super-sized” collection that provides a surfeit of resources to only one user community. In still other cases, it is difficult for librarians to get beyond sticker shock of a hefty price tag to assess the value of the collection in terms of unit costs or potential use. With more than 100,000 volumes, the EEBO collection was priced (list price) at less than one dollar per volume. How much further can we as consumers expect to beat down the unit costs of production for such work—especially when we look at the unit costs incurred by many academic libraries engaged in similar efforts? Cost effective or not, I realize that it is not easy to fund resources in the neighborhood of \$100,000—even for the country’s largest and best funded libraries. Purchases of the magnitude of EEBO invariably involve planning and coalition building—no individual bibliographer or collection manager can simply say “yes” and fund it out of pocket. So, a decision not to buy, or a non-decision that results in not buying, can also reflect a shortage of the time and team-building skills needed to bring together a group that could move such a purchase forward.

PRESSING FOR BETTER PRODUCTS WITH MORE ADVANTAGEOUS TERMS

The notion of libraries organizing to make resources better as a condition of purchase is an offshoot of the digital library environment. In the print world, books—or even reels of microfilm—came to us complete, and while we had the option of rejecting these resources, it was generally not feasible to modify a resource once it was available for distribution. E-products, on the other hand, are dynamic, often released incrementally, and maintained and upgraded by the publisher over time. In the case of

electronic resources, I would argue that the library community not only has an opportunity to shape a product, it has an obligation to present and future users to insist on desirable changes. Likewise, given the size of the required investments (and the corresponding risks) one would imagine that publishers and vendors would want to stay very close to market needs and perceptions, seeking input early and often. Midstream changes, and product rescue efforts after the fact, can be expensive and difficult to implement while in full-scale retreat. Better to encourage systems of collaboration at the early stages of development so that market expectations are clearly understood by the sellers, and so that the buyers, in turn, understand the cost and quality trade-offs underlying product decisions.

The EEBO Text Creation Partnership (EEBO-TCP) entered the scene in 1998 to offer an early response to the EEBO marketing effort. We believed then, and believe now, that making the EEBO content searchable was an enhancement that our users would demand and was a way to move beyond the limited access provided by cataloging records and accompanying microfilm images to take full advantage of the power of digital access. ProQuest never denied this, but simply and correctly felt that such an undertaking at their end would make the product unaffordable to all but a literal handful of institutions. When it was suggested that Michigan believed it could rally the community to bear some this cost if the funding libraries would be eventual owners of the text files, ProQuest began working to see if they could find a way to support such a joint venture. Over time, the details were worked out and this unusual partnership between a commercial vendor and the academic community was codified.

From the outset, the EEBO-TCP lead-institutions (the Universities of Michigan and Oxford) talked about the vision of creating a subset of 25,000 accurately keyed and tagged editions of works. This “product” however, was always less important than the principles and practical library truisms that dictated how and why the texts would be developed.

- Libraries, as a more or less unified market, need to establish realistic standards for digital products. Such standards should be cost effective, rather than an abstract ideal, and when producers meet these standards, libraries should be willing to pay. When products fail to conform to standards, libraries should withhold support regardless of the attractiveness of the content.
- Library staff have certain skill-sets, and access to campus users who have yet other skill-sets and expertise, neither of which can we routinely expect to find in the employ of commercial vendors.

This is an important basis for partnership that would make better information resources than either libraries or commercial vendors are likely to produce working on their own.

- When dealing with historic texts, or even current scientific journals, it seems axiomatic that users would benefit from greater uniformity of search and retrieval protocols. While publishers compete to produce ever better interfaces, and this competition undoubtedly does lead to improvements, the net effect for users is a level of complexity and jarring dissimilarity that works against successful retrieval across collections.
- Even if there were convergence of interface, it would in many instances be desirable to integrate collections behind a single interface rather than expecting users to move from one publisher's collection to the next to retrieve resources on a single search entry. Expecting users to understand the relationships between EEBO and the Eighteenth Century Short Title Catalogue, or Kluwer journals as opposed to Elsevier journals, is to burden them with market distinctions that bear no relationship to discipline or subject boundaries. While it is hard to imagine commercial vendors transcending their own branding, libraries could provide a neutral ground for bringing together content in aggregated files—as they always have on their bookshelves.
- Research libraries have an obligation to preserve content across generations, and have an excellent track record for meeting this responsibility.
- Finally, research libraries need to own and manage key resources for the long haul. Important collections with durable content should be part of the capital investment of the nation's research libraries and we shouldn't countenance the desire of commercial interests to move materials out of the public domain by virtue of converting them from one format to another. Commercial firms are entitled to a return on their investments in conversion but libraries need to own their collections and make them broadly available as is consistent with their educational mission. Rights to important cultural resources should ultimately revert to purchasing libraries with few if any fetters on their use beyond general conventions that regulate intellectual property rights.

These principles outlined above have been an underpinning of the EEBO Text Creation Partnership and are the basis for ongoing discussions with publishers other than ProQuest to extend their acceptance as the basis for creating text editions of other historic corpora.

The EEBO-TCP is at this point a successful cooperative effort supported by almost seventy academic libraries in the U.S. and abroad to fund the creation of full-text editions of works in the Early English Books Online Corpus. The text collection now stands at two thousand accurately keyed and tagged texts online and we've not abandoned the goal of converting 25,000 texts over the next five years. For partner libraries, the text files will be made available for local load and management and eventually (within the next twelve years) any library could make the texts available to any constituency that they choose to serve.

The unit costs of this effort are also an attractive endorsement for such collaboration. At this early stage of the project, ARL partner libraries have paid approximately \$25.00 per converted EEBO-TCP text. With the cash on hand, and existing library commitments, the cost will be approximately \$5 or \$6 per TCP volume and this could drop to \$2 per text if more partner libraries sign on. This is clearly a bargain rate for the permanent addition of quality resources to a library collection.

EEBO has likewise proven to be a very successful product for ProQuest. Approximately 150 institutions have purchased access to EEBO images—a number that in just three years surpassed the number of STC standing-order subscribers recruited over a sixty-year development and sales effort. All involved would say that at least some of this success should be attributed to the unusual partnership that ProQuest struck with the academic community. The partnership has allowed ProQuest to supplement its image product with the opportunities that full-text conversion offers for new kinds of research. Creating text editions allows discovery and innovation by opening up the corpus to truly granular retrieval of words and concepts, with links to the corresponding page images. The market appeal of this arrangement has not been lost on competitors as they seek to bring similar content online.

EEBO and EEBO-TCP (image and text) have developed as complementary representations of these historically significant works, and each approach adds value to the other. While the EEBO-TCP effort is independent of ProQuest—the work being done by staff at the Universities of Michigan and Oxford, with partnership fees collected by the Council on Library and Information Resources—we recognize that the project benefited greatly from the support offered by ProQuest. The company has been instrumental in marketing the text effort and is contributing far more money for the privileges of partnership than any single academic library. Nonetheless, our view of the EEBO Text Creation Partnership is that it is a project being led and financed by academic libraries, which must ultimately control the standards and production

methods. Most important for the long-term, libraries will retain the rights to own and distribute the text collections in accord with the goals and mission of their institutions.

EXTENDING THE EEBO/EEBO-TCP MODEL

The commercial success of EEBO has caused other publishers to overcome risk-aversion and commit to moving forward with mega-conversion projects of their own. Most closely aligned to EEBO are Gale's Eighteenth Century Collection and Early American Imprints produced in microform by Readex. Both of these collections provide important extensions of EEBO content by time and place. The two publishers have been considering conversion of these collections for many years but are now moving forward, likely based on business plans justified by the market acceptance of EEBO.

In the case of the Eighteenth Century, Gale has already issued a press release indicating its intention to work with the library community along lines established by the EEBO-Text Creation Partnership. For users, this would mean that the EEBO text files extending from 1473-1700, could be combined with thousands of 18th century texts, and searched as a single corpus, with corresponding page images "grabbed" from the appropriate publisher server. This level of integration is a compelling vision for scholars. Early American Imprints (the Evans/Shaw-Shoemaker bibliographies) could likewise be incorporated into such a research environment if the publisher can come to grips with the benefits of this level of collaboration. From the perspective of partnering institutions, it would be a great service to our users if we could integrate the texts from all three of these cornerstone collections behind a single interface, allowing the user to choose if their search term should apply to one, two or all of the collections keyed to a single standard and tagged in conformance to a unified Document Type Definition (DTD). It is also the case that libraries that own these text-files in perpetuity will be able to use them in conjunction with other projects and collections.

While the publishers will have to decide if they can accommodate the principles underlying the Text Creation Partnership, their decision will largely be driven by the reaction from the marketplace. Will libraries accept page images and uncertain OCR searchability, or will they insist that some of the texts in these culturally significant corpora be accurately keyed and tagged to allow for more precise searching, browsing of section heads, retrieval of keywords in context, and displayable and

printable text? Obviously, keyboarding and tagging adds cost so it is a significant collection decision to determine how much libraries can afford to invest in these collections. Assuming that the market expresses a desire to have some texts accurately produced, will our colleagues insist that they be granted ownership rights to these texts, including the right to integrate them with other collections, or would they be satisfied to pay a commercial vendor to produce, house, and control these texts along with the corresponding page images? In yet another possible outcome, would libraries and repositories settle for a royalty on commercial sales rather than claim ownership and control of the content? And finally, would users prefer to pay a commercial vendor for texts as product (even if the vendor subcontracts with an academic institution to actually produce the texts) or would they prefer to pay into a cooperative initiative to produce as many texts as possible depending on the rate of community participation—i.e., the number of paying partners?

While many libraries have approached the EEBO Text Creation Partnership output as a *product*, it was conceived by the initial advocates at Michigan, Oxford and even ProQuest, as a set of principles that could be applied in such a way as to support the goals of both commercial publishers and the academic library community. Because of the library-friendly aspects of this TCP arrangement, it is always disappointing to us when libraries compare EEBO/EEBO-TCP to traditional products and end up deciding in favor of the latter. In some cases, these decisions are program driven and sensibly reflect for the needs of a particular campus. For the library community as a whole, however, it is important that our collection librarians pay attention to principles that ultimately advance our relative standing in the marketplace.

Although EEBO-TCP can claim a measure of success, I'm not sure that the project has done enough to underscore the importance of liberal terms of use, and the benefits of ownership to research libraries. It would be disappointing if this were seen as an isolated opportunity specific to EEBO that can't or shouldn't be applied to other similar projects. We would not like to see libraries too willingly cede the gains that we feel were made in our relationship with ProQuest. There can be no argument that the terms of EEBO-TCP favor libraries and their users. But, they have also provided tremendous opportunities for ProQuest to realize a strong return on its investment. Others vendors that would attempt to circumvent the issues of standards, rights and ownership should be held to account.

THE BENEFITS OF PARTNERSHIP

While my greatest fear, post-EEBO, is that the library community will accept disadvantageous sales approaches and licensing terms for collections of core content and enduring value, there is also the possibility that the EEBO-TCP model could be undermined from a different direction. Some librarians, when presented with the TCP principles, seem emboldened to promote the idea that libraries could develop these collections totally independent of the commercial information providers. While I have some empathy for these views based on confidence in library skills, and can think of some significant successes, my own experience with EEBO and other projects makes me dubious that this level of coordination is feasible among libraries. To my mind, it is highly unlikely that repositories would cooperate to release their content to the larger library community without significant remuneration. It is further unlikely that the community could develop a viable funding model that would transfer considerable funds to one or several production sites. Finally, the kinds of investments required to market/sell an international project are considerable, and very unlikely to be carried out successfully by one or several libraries. While I don't think EEBO would have accomplished its excellent sales track record were it not for the Text Creation Partnership, I'm equally assured that the TCP would not have achieved anywhere near its present level of success were it not for its connection to the EEBO image product and the support of ProQuest.

ProQuest (or other commercial conversion vendors) brings to the table the following assets in support of partnerships with academic institutions:

- Image product with rights of distribution
- Production and business credibility
- Sales force and marketing staff—international market penetration
- Relationship with content providers
- Capital—both ongoing contributions to support conversion and front money in advance of revenues
- Drive to bring a product to market—time-bound goals
- Some content knowledge based on managing the collection in other formats.

EEBO-TCP could not have reached its current levels of success without early support and funding from ProQuest. It was also helpful that the EEBO image product was viable even without the searchable text,

allowing TCP a protracted period in which to seek funds and ramp up production. Finally, the national and international sales effort would not have been carried out on an effective level if left to the devices of an academic library. It is unfortunate, but a fact nonetheless, that most institutions assessing a project like EEBO/EEBO-TCP require ongoing contacts over an extended period of time to make a purchase/partnership decision. Even as a library colleague, it frequently involves unabashed “sales” calls to actually bring a decision to closure. While we can quibble about the extent of value added by commercial sales and marketing efforts, it is naive to overlook the work involved in distribution and its relationship to funding creation.

Conversely, libraries offer vendors the opportunity to create far better products than they could build on their own, without having to assume all the risks of missing the mark on production schedules, quality standards, or other aspects of a project. When the TCP set out to convert EEBO texts, we announced an intention to do so at a minimum accuracy standard of one error in 20,000 characters (99.995). The fact was that, given the difficulty of the texts with which we were working, we didn’t know for certain that the keyboarding vendor community could meet this standard, or do so in a cost effective manner. While happily this has worked out, we always believed that had it not we could come back to the community and say that it doesn’t appear possible to meet this standard for the material in hand. We believed that a lower but achievable standard would still offer a tremendous resource for most users. I was confident that our “colleagues”—as opposed to “customers”—would accept such a change as a good faith decision by a peer, rather than as a contract breach or profit-driven effort to cut back on quality and service. *Trust* is a tremendous asset that libraries can bring to the process of resource creation.

In the case of EEBO-TCP, the lead libraries brought

- Production capability/credibility in dealing with humanities texts
- Content credibility—in the library and on campus
- Concept investment that translated into will to work and to commit resources
- Director/Collection Development Officer/bibliographer marketing connections
- A cooperative spirit reflected in trusting collaborative relationships.

This was not a project that could have been done by a commercial vendor because of the skill sets required, because of the need to draw

feedback and direction from end users, and because of the untested expectations set for the conversion vendor community.

While ProQuest quickly recognized the value of adding searchable text to their records and page images, some vendors and some librarians might question the need for such a project to be done—at additional cost—when OCR searchable text is available. This will be the crux of the issue for future efforts at forging such commercial/academic partnerships. Does the full-text really add enough value to warrant the additional costs? For me, the answer is, “yes, for a limited subset of texts.” Conducting a simple keyword search over a corpus of 100,000 texts, and then following “hits” when no structure is displayed (chapter heads and section heads) and no surrounding text can be viewed, seems an exercise fraught with danger and frustration. The corpora are huge, and the cues for tracing hits are all but non-existent. By virtue of their significance and likely use, some texts deserve to be more fully represented; displayed and readable in modern fonts, browsable by virtue of rendering tagged chapter, section, and feature headings, and reproducible for incorporating the text into other work. With many new corpora on the horizon, it will be interesting, and I believe important, to see if the library marketplace affirms this judgment about the value of keyed and tagged editions of digital facsimiles. Then it remains to be seen if the marketplace will likewise affirm the value of full ownership of these texts—and the right to freely distribute them in support of the principle of public domain access—as opposed to accepting the limited rights of ownership supported by traditional product licenses.

CONCLUSION

We started by asking about the role of the research library in shaping a digital library that is being created first and foremost by commercial publishers. I think the answer is that libraries need to find a way to leverage their skills, resources, and most significantly, buying power, to work in partnership with commercial firms, giving clear messages as to the nature of the products they want and the terms under which they are willing to access them. In the instance of EEBO, it was argued that libraries could claim ownership of the text file because it was primarily their money that was used to create it. In a sense, library money (and library content) drives all commercially created resources. Accordingly, we should be mindful of our role in the creation process and advocate for arrangements that support our long-term interests and those of our

users. The importance of the EEBO Text Creation Partnership will be reflected in how it influences other likely conversion efforts—e.g., the Eighteenth Century, the Nineteenth Century, Goldsmith-Kress, and Evans-Shaw/Shoemaker. Will the publishers of these collections adopt production/marketing/licensing models based on the EEBO experience or will they seek to adopt more insular approaches to production and restrictive licensing? If so, will the market accept such an approach given the expectations developed around the EEBO-TCP? If libraries are to remain true to their oft stated values, it is hard to see how they could accept licensing terms that restrict a reader's right to access culturally significant texts in the public domain.