Bedfellows in Mass Digital Conversion: Ten Years of Text Creation Partnership(s)

Aaron McCollough

TCP Project Outreach Librarian, University of Michigan Library, Ann Arbor, MI, USA
amccollo@umich.edu

Abstract. This paper offers a brief account of some senses in which "partnership" and collaboration have been and continue to be fundamental to the Text Creation Partnership's mandate. Additionally, by examining the kinds of collaborative engagements in which TCP has participated, it raises and addresses questions about some frictions produced by unlikely partnerships between the private and public sectors as well as about benefits afforded by the same. Finally, it tenders some suggestions about future collaborative efforts the TCP might help to foster and, in turn, be fostered by. Ultimately, this paper stresses cooperative protection of the digital cultural commons as a cardinal virtue of digital humanities collaborative effort. It seeks to refocus attention on this aspect of the TCP's role in the field.

Keywords: Text Creation Partnerships, project collaboration, information management, scholarship, digital humanities

Introduction

"Mass Digitization" has been a hot – often uncomfortably hot – topic, at least since 2004 when Google unveiled its secret "books project" to the world. Somewhat less widely discussed and debated, although certainly no less significant, has been the mass conversion of idiosyncratic, OCR-baffling historic corpora into TEI-compliant markup. As the project I work on, the Text Creation Partnership has drawn its first phase of EEBO work to a close; however, now seems like a sensible time to revisit some accomplishments and challenges of such mass conversion.

This paper offers a brief account of some senses in which "partnership" and collaboration have been and continue to be fundamental to the TCP's mandate. Additionally, by examining the kinds of collaborative engagement in which TCP has participated, it raises and addresses questions about some frictions produced by unlikely partnerships between the private and public sectors as well as about benefits afforded by the same. Finally, it tenders some suggestions about future collaborative efforts the TCP might help to foster and, in turn, be fostered by.

Context

With 25,000 books converted from ProQuest's Early English Books Online, EEBO-TCP is the largest TEI-encoded text collection in the world and one of the most important fully-searchable text corpora for the humanities in existence. From the outset and in its most fundamental assumptions, the TCP has always been a collaborative effort. In fact, collaboration has always been a function of necessity for us. The need began with EEBO. Once ProQuest had produced the EEBO digital image product, they deemed the prospect of unilaterally converting it into text to be prohibitively expensive. Librarians and scholars, for their part, considered the flexibility of hypertext editions of this material to be revolutionary and necessary. Both also wanted to protect the archives' place in the digital public domain. A unique strategy was called for, one based on partnerships across the domains. 25,000 converted books later, this strategy still seems unique and successful. If 25,000 books is roughly only 1/3 of the unique content in the EEBO collection, it still stands as a noteworthy landmark to the potential for collaboration between very different enterprises.

EEBO and EEBO-TCP have significantly extended potential access to early modern archival treasures, and this potential has drawn comparison to the Model-T's impact on access to automotive ownership. Whereas EEBO production resembles the straightforward automation of the assembly line, though, the TCP model might bear better analogy to a public works initiative like the WPA's Federal Writers' Project, which sought in the depths of the Great Depression to produce a five volume *American Guide*, a "geographical-social-historical portrait of the states, cities, and localities of the entire United States" (Yetman, 1970). The point being, the initiative to create hypertext versions of every book printed in English before 1700 is a massive, costly undertaking, and one quantitatively and qualitatively significant to preserving a huge part of European and American cultural heritage for everyone. Also, it sometimes looks impossible.

All told, the first phase of the EEBO-TCP cost around 6.8 million dollars. Each book, therefore, cost $272 to produce. If that seems steep, consider how the per-book cost breaks down through the collaborative funding strategy. The famous American Express slogan exhorts us, "membership has its privileges." In the TCP's case the ideal "privileges" of wide membership have always been imagined in terms of access as much as cost. The more cooperation we've gotten in bearing the overall financial load, the more secure the goal of converting a significant subset of the Short Title Catalog has become. Likewise, the more secure the conversion goal, the safer the promise of protecting cultural heritage materials from restrictive terms of use dictated by digital licensing agreements. With over 150 libraries sharing costs and with 20% matching funds from ProQuest, the cost per text per partner institution has ended up being less than $2.

Public Domain Priorities

In 2003 — already four years into the initial phase of the TCP — Mark Sandler articulated ambient anxieties about the fate of collective culture in the digital age thus: "In the current licensing environment, large bodies of creative works once in the public domain are being returned to commodity status—and this time a commodity status that will never expire ... Denying access to these culturally significant works is an affront to the values of libraries, an affront to the mission of our universities, and flies in the face of Anglo-American law that justified both copyright and public domain as a means for advancing social progress. I'm hopeful that librarians, scholars, and publishers can begin an active dialogue about ways to encourage digital conversion without contravening the public's right to share in our collective cultural heritage" (Sandler, 2003). Certainly, the dialog Sandler hoped for has come about. It has become quite boisterous at times. But solutions to the problems that dialog identifies have tended to be elusive. The simplest strategies for protecting the commons are probably the best. Every library can create digital collections from its own special holdings or in collaboration with one or two other libraries. Grant money for such projects has been relatively abundant, and project overhead has been manageable. In the case of unique, discrete collections this has proven fairly successful. An early case in point is the Making of America (MOA) project coordinated by the University of Michigan and Cornell University in the mid-nineties. MOA had Mellon funding and converted approximately 1,600 monographs and ten serials in its Michigan effort. The Cornell effort converted another 109 monographs and 22 serials. These materials were generated using OCR with minimal document structuring and low-level indexing added post-conversion.

The STC is not held in one or two libraries, of course. The volumes that comprise it are scattered across the collections of the world. Additionally, the human labor required to capture millions of pages of early modern print as modern character encoded text is beyond the reach of libraries working on their own or in small clusters. Recent advances in distributed (or "crowd sourced") correction and editing do suggest promising possibilities for making this labor more organically shareable, which I will explore a bit below, but these strategies are still theoretical. The TCP aims to employ them in a piece-meal way as the relevant pieces can be identified and deployed. For the time being (and likely for some time to come), the established methods of conversion (using keyboarding vendors for basic capture and a centralized production staff for review and TEI-light markup) remain the most reliable means to pursuing the TCP's chief priority: preserving the early print archive as a digital archive for future generations.

Of course, the established methods of conversion are only as reliable as the institutional support they can sustain. The second phase of EEBO-TCP production has been fortunate so far to receive commitments from over sixty individual institutions and to reach a consortial arrangement through the JISC in Britain. These are encouraging signs that the library and scholarly communities continue to support the TCP effort and philosophy. At the same time, in my short tenure as TCP Project Outreach Librarian, I've found that many institutions that have supported the TCP in the past are now skeptical about the project's virtues. Difficult economic circumstances are easy to point to as a cause, but a more worrying element — the loss of perspective on the project's real aims — is the manifest expression of this doubt. Lingering suspicions persist regarding the role of commercial publishers in TCP initiatives. Also, some partners have complained that the TCP arrangement does not feel truly collaborative — that they feel their role has chiefly been as consumers rather than as a co-creators and co-beneficiaries. Finally, I've encountered what might be called a kind of institutional exceptionalism. Collections officers have averred that usage numbers are not very high among their particular patrons, that their particular faculty don't seem to be interested in the enhancements to EEBO afforded by searchability and/or local loading. Some scholars, too, are content with the present text subset of the STC, because that subset is adequate for the type of digital humanities research they are currently doing. All of these views are reasonable, but they ignore the priority on the public domain informing the most basic sense of *partnership* in the Text Creation Partnership.

Tight collection budgets oblige libraries to retreat from longer-range initiatives because the necessity of such initiatives is harder to prove (and thus the expense is harder to justify) in the short-term. For many, more TCP text seems like a luxury and one they can do without. As tempting as this view may be to cash-strapped librarians doing budgetary triage, it is dangerously narrow, and it undercuts the cooperative consensus necessary to make an unimpeachably valuable resource truly free.

Exceptional Partnerships

To date, the most conspicuous sense in which the Text Creation Partnership can be said to be collaborative is in its work with key research initiatives in the digital (and analog) humanities. This list includes INKE, of course, as well as the English Broadside Ballad Archive at UC-Santa Barbara, The Spenser Archive at Washington University, the National Library of Wales' corpus of Early Welsh printing, Renaissance Cultural Crossroads at the University of Warwick, The Complete Shirley project at Anglia Ruskin University, WEME: Witches in Early Modern England at Simon Fraser University, Virtual Research Environments at the University of Hull and East Anglia, the Oxford English Dictionary, NORA, CIC CLI Virtual Modernization, Word Hoard, and MONK at Northwestern and UIUC, and the recently published digital Holinshed at Oxford. Similar, equally promising new projects seem to emerge daily, including linguitic and genre-oriented projects at the University of Helsinki and the University of Wisconsin, as well as efforts to improve OCR-technology, including 18thConnect and IMPACT.

In each of these cases, a scholar or group of scholars at a TCP partner institution (or institutions) has requested a local load of the texts produced through the TCP workflow. My predecessor Shawn Martin and I, along with the TCP production staff, have then liaised with these scholars to facilitate the loading and incremental updating process so that the texts can be edited, mined, mashed, and otherwise processed at will through local installations. We have always been keen to act on special text requests by such projects (see, for example, the Holinshed Project's recent note that "the edition would have been impossible without the co-operation of EEBO-TCP who undertook the keying of the 1577 edition (in addition to the 1587 edition already on their site)"), but the majority of our effort has been aimed at producing a resource that is uniformly useful to any and all partners interested in enhancing or otherwise exploiting the texts comprising it (Archer, 2009). What the various scholars do with their texts (or, more accurately, the texts they co-own with all TCP partners) is basically none of our business. That's not to say we don't care about what these projects are up to. On the contrary, the first wave of TCP-related digital projects has done as much as anything could to reinforce the value of the TCP resource. Additionally, projects like those spearheaded by Martin Mueller, for example, continue to expand our sense of what digital text can be coerced into teaching us about our own history.

When it comes to these *exceptional* scholarly projects, TCP collaboration is mechanical but — we like to think — workmanly. The rewards are those afforded by playing some part in innovative endeavors, most crucially they come in the accumulation of cultural capital necessary to justify further text production. For the TCP (and for its partners, even if they sometimes lose sight of it), progress toward complete (and completely free) digital archives is a fundamental goal of each and every collaborative effort.

Where to, Text Creation Partnership?

While the EEBO-TCP has been successful by nearly any measure, its siblings ECCO-TCP and EVANS-TCP have seen less thorough institutional adoption. Unsurprisingly, local loading of the Eighteenth Century and Early American texts has been rare, and the perception of these resources' utility has suffered. These shortcomings are in no small part due to the OCR text accompanying the Gale and Readex image products. Although the OCR error rate for these collections is unacceptably high for most scholarly purposes, the mere existence of some searchable text seems to be enough to sink the effort to market high-quality keyed editions. Unfortunately, keying costs for ECCO-TCP and EVANS-TCP have outstripped revenues. Keying has therefore been suspended until additional funding can be identified.

The TCP has not given up on ECCO and EVANS. We are hopeful, in fact, that collaborative work with projects like 18thConnect and IMPACT will lead to new encoding and reviewing strategies that build on OCR's strengths while compensating for its weaknesses. As these groups use TCP text to improve OCR capture quality for early print, distributed correction models — like those currently used by Project Guttenberg for proofreading and employed by the German TextGrid project — should be flexible enough to address many of the remaining gaps in the conversion process through small, distributed analytical tasks. The expertise of the TCP production staff would position it well to collate, evaluate, and implement corrections delivered from the cloud.

This type of collaborative editing represents another partnership model on the horizon. Theoretically, it could save enormous sums of money currently committed to keying vendors. It could speed production, and it could put much of the editorial responsibility back in the hands of scholars.

After ten years, the Text Creation Partnership is still evolving. Although our production techniques are labor intensive and methodical, we remain nimble-minded in our approach to the future. At this moment, the most important definitional questions regarding "partnership" for the TCP have to do with collective responsibility to the cultural commonwealth. In this sense, TCP projects represent a major test case for collaborative priorities in the digital humanities community. Whether scholars and librarians can mobilize one another to keep text creation going, especially for ECCO and EVANS, but also for EEBO, will indicate a great deal about how collaborative the field can really afford to be in the years to come.

Works Cited

Archer, Ian. *Holinshed's Chronicles*. Web. 18 Sept. 2009. <shakesper.net>.

Sandler, Mark. "The Public Domain: To Be Or Not To Be." *The Charleston Advisor 5*.1 (2003): 56. Print.

Yetman, Norman. *Life Under the 'Peculiar Institution.* New York: Holt, Rhinehart, and Winston, 1970. 346. Print.