Transcribed by hand, owned by libraries, made for everyone:

EEBO-TCP in 2012

What is the state of the Early English Books Online-Text Creation Partnership (EEBO-TCP) in 2012? It is...old. Or is it? EEBO-TCP is now in its 12th year as a project, and its 11th year of keying and encoding early English books. In context, this means:

- EEBO-TCP appeared on the scene at just about the same time as the University of Nebraska's Walt Whitman Archive,[1] the University of Sussex's Newton Project,[2] the University of Virginia's Complete Writings and Pictures of Dante Gabriel Rossetti,[3] and Dickinson Electronic Archives.[4]
- It is younger by a few years than The William Blake Archive (1996), Indiana University's Victorian Women Writers Project (1995),[5] and the University of Victoria's Internet Shakespeare Editions (1996).[6]
- None of these can touch the Brown University Women Writers Project (1988),[7] Tufts University's Perseus Digital Library (1985),[8] and Project Gutenberg (1971!).[9]

EEBO-TCP is much younger than the oldest digital text archives, and came into being at the same moment as many other projects focused on building collections of TEI-encoded documents representing specific people, subjects, and corners of history. Like the projects above, EEBO-TCP focuses on a rather narrow area—English books 1475-1700—and is concerned with providing structured electronic text.

Around 2005 there seems to be a shift in focus from narrowly focused, boutique digitization projects, to preserving, presenting, and managing large amounts of digitized material. Two distinct approaches emerge: The project might gather together massive amounts of data from many sources, implementing large-scale production workflows in order to do so. Here I am thinking of the Google Books Project (2006),[10] Hathi Trust (2008),[11] and the JISC Historic Books platform (2011).[12] Although EEBO-TCP is working on a smaller scale by several orders of

---

[1] http://whitmanarchive.org/about/history.html
[2] http://www.newtonproject.sussex.ac.uk/prism.php?id=28
[3] http://www.rossettiarchive.org/about/index.html
[4] http://www.emilydickinson.org/about_us.html
[5] http://webapp1.dlib.indiana.edu/vwwp/projectinfo/technical.do
[6] http://internetshakespeare.uvic.ca/Foyer/about.html
[7] http://www.wwp.brown.edu/about/
[8] http://www.perseus.tufts.edu/hopper/about
[9] http://archive.org/details/gutenberg
[10] Announced under the name Google Print in December 2004 (http://googlepress.blogspot.com/2004/12/google-checks-out-library-books.html).
[11] Founded in 2008, the first newsletter is here: (http://www.hathitrust.org/updates_march2008).
[12] http://www.jisc-collections.ac.uk/jiscecollections/jischistoricbooks/

magnitude—tens of thousands, rather than tens of millions, of books—its aims align closely with this type of project. Its mission is to create a corpus that approaches comprehensive coverage of English publishing before 1700.[13]

Alternatively, this new generation of project gathers metadata in order to serve as a portal for cross-searching and discovery, rather than an archive in its own right. In this category, we see NINES (2004) and the ARC initiative that it spawned, the CERL Portal (2008), and Europeana (2008). Along the same lines, projects such as Dariah (2004)[14] and TextGrid (2009)[15] build infrastructure for the discovery and exchange of data. EEBO-TCP is quite clearly not this kind of project, but the amount of data it has produced—close to 5 million pages and 10 billion words!—certainly calls for engagement and collaboration with frameworks that can connect this data with other related content, and get it into the hands of those who wish to use it.

Why this "Who's Who" of digital archives? I draw this timeline in part to offer some context for the EEBO-TCP project, which is very much a product of its time, and to highlight its place among these projects—a position that is quite unique. More mass-produced than the Women Writers Project, smaller and more thoughtfully curated than Google Books, and with ties both to commercial publishers and the open access mission, EEBO-TCP has elements in common with many projects, but no one else is doing quite the same kind of thing. The uniqueness of the EEBO-TCP project comes down to three key points. The texts produced by this project are:

- transcribed by hand,
- owned by libraries, and
- made for everyone.

These are the three legs of the EEBO-TCP mission and—strangely enough—also the three most common misconceptions that I encounter when speaking to people about the project. These pillars, then, provide a useful framework from which to reflect and report on the TCP's past, present and future.

I: Transcribed by hand

The Process:

According to the product's website, Early English Books Online (EEBO):

> "now contains more than 125,000 titles listed in Pollard & Redgrave's *Short-Title Catalogue (1475-1640)* and Wing's *Short-Title Catalogue (1641-1700)* and their revised editions, as well as the*Thomason Tracts*

[13] EEBO-TCP focuses on English language texts and on first editions, so cannot truly claim to have covered everything published in England before 1700. However, the approach emphasizes broad coverage produced on a large scale, rather than carefully selected individual texts.

[14] http://www.dariah.eu/index.php?option=com_content&view=article&id=7&Itemid=119

[15] http://www.textgrid.de/en/ueber-textgrid.html

(1640-1661) collection and the *Early English Books Tract Supplement*."[16]

EEBO provides access to digital images—in most cases derived from microfilm—of books owned by many libraries around the world, although collections of the British Library, the Folger Shakespeare Library, and the Huntington Library are represented with particular prominence. EEBO is published by ProQuest and available to institutions or consortia for subscription or purchase.

EEBO-TCP adds value to EEBO, and simultaneously builds an independent archive of encoded text, by producing an accurate, searchable, XML-encoded electronic text file for each unique work represented in EEBO. Because of the challenging fonts used in early printed books, variation in emerging printing conventions and techniques, and unpredictable quality of the microfilm from which EEBO's digital images are derived, it was not possible in 1999 (when the project was first conceived), and it is not yet possible today, to use optical character recognition (OCR) to automatically generate electronic text for these works. (New research being done in this area will be addressed below).

Instead, in monthly batches, the page images for each book are delivered directly via FTP to our vendors, Apex CoVantage and SPi Global. There, each text is manually keyed by multiple people, and the differences reconciled. Structural markup, based on a customization of TEI P3, is added. The files are posted to a server, and monthly, a project manager at Michigan retrieves them. The first task of the TCP editors at Michigan and Oxford is to ensure the accuracy of the transcription, and therefore of the corpus as a whole. Editors work one book (or one batch of books) at a time, proofreading a 5% sample of each book, letter by letter, accepting files that meet the 99.995% accuracy specification and sending those that do not back to the keyers for revision and resubmission. Defects in the source image, or in the book itself, which may have been badly printed or preserved, make this a somewhat subjective process, requiring that the editors distinguish between 'excusable' (or 'forced') errors and 'inexcusable' (or 'unforced') errors, forgiving but correcting the former, counting the latter against the specified maximum error rate of 1:20,000, and alotting credit for particularly percipient capture of hard-to-read material.

Books that pass the accuracy test are then edited for correctness of markup: the editor's job at this stage is to ensure that all essential parts of the book's structure have been recognized and aptly tagged. Editors are expected to work at speed, balancing the cost (in time) against the benefit of every prospective revision, with 'benefit' always interpreted with a view to the user and reader, the objective in every case being intelligible display, intelligent navigation, and informed search.[17]

The History:

Planning for this workflow began as early as 1999 and with a few major exceptions—early drafts of the workflow called for shipping printouts of the pages to the vendor!—it has not changed

[16] http://eebo.chadwyck.com/marketing/about.htm

[17] Thanks and credit are due to Paul Schaffner for this description of the EEBO-TCP editorial workflow.

much. The very first text edited by EEBO-TCP was finished March 2, 2001, and the first batch of searchable, encoded text was unveiled at the American Library Association's Annual Meeting in 2001.

Any measure of the TCP's progress must take into account both the number of books and the amount of data (measured in pages, images, or kilobytes) processed: very large books may give the impression that little work was done in a given time period, while very small books may inflate the number of books completed, while also reducing the amount of data processed because of the additional per-book overhead (which can be addressed to some extent by batching small books together for processing at once). Accounting for all of these factors, the project gradually increased in productivity each year through 2007, when EEBO-TCP Phase I began to wind down. There is a dip in overall productivity in 2008—perhaps due both to discussion of whether or not the project would continue on with a second phase, as well as preference at that point for smaller books. Productivity picked up again in 2009, with only small ups and downs since then. Through 2007, the project averaged about 4,500 typical books per year.[18] From 2008 to the present, the project has averaged about 5,450 typical books per year.[19]

The Future:

By the end of 2012, EEBO-TCP will have keyed and encoded around 48,000 books. This is 80% more works than the original project goal, to convert a selected group of 25,000 texts. However, it is just 64% of the way toward completing the EEBO-TCP Phase II goal: to produce one electronic text for every unique work in the EEBO corpus.

The most important changes on the horizon that have the potential to impact our "transcribed by hand" assertion are new tool and platform developments going on outside the TCP. The Improving Access to Text (IMPACT) project, funded by the European commission, finished its research into improving OCR for early texts in 2011 and now hosts a Centre of Competence for libraries, archives, and museums seeking guidance in the digitization of historical materials.[20] Across the Atlantic, projects are in the works at Texas A&M University, the University of Maryland,[21] and elsewhere to both improve OCR software, and optimize the balance between automated and human creation, correction, and annotation of text.

If OCR can be trained to "read" EEBO page images, at a level of accuracy that makes it more efficient to clean up the automatically generated text than to key it from scratch, huge amounts of money and time could be saved, and human effort directed instead to correction and smart markup. We are eagerly watching these developments, and have offered support in the form of

---

[18] A typical or average book consists of roughly 100 pages, 50 images (each EEBO image includes two facing pages), or about 200 kilobytes (each image averages 4KB of text). These numbers are based on the number of titles and the number of page images available to be selected from EEBO for conversion.

[19] Thanks are due again to Paul Schaffner for calculating these production numbers. Any errors in representing them are mine.

[20] http://www.impact-project.eu/news/ic2011/

[21] http://mith.umd.edu/ocr-xml-topic-modeling-and-braille-accessibility-coming-soon-thanks-to-the-neh

providing TCP texts to be used as a "ground truth" against which to test these efforts. Since the inception of EEBO-TCP, "transcribed by hand" has been a core part of our work, but this is a means to accurate text, not an end in and of itself.

II. Owned by libraries

The Process:

The rights to the tens of thousands of EEBO-TCP text files are jointly held by all the stakeholders in the project, mostly academic libraries that have joined the project as partners. EEBO-TCP Phase I had around 150 independent library partners (nearly all in North America). Phase II has just about 100 partners.[22] In both Phase I and Phase II, JISC Collections has provided generous support to the project, garnering access to the texts for most academic libraries in the United Kingdom. These commitments make the work of EEBO-TCP possible, and directly fund the conversion of more texts. Early access to the texts is the main immediate benefit of joining.

In the United States, the introductory fee for an ARL[23] library to join EEBO-TCP Phase I was $50,000. The cost of producing a typical book is between $200 and $250. Each library that joined funded the creation of around 200 titles that would not otherwise be keyed. In return, they received immediate access to all 25,363 texts. To frame it a different way, each library paid just under $2/book—less than 10% of what it cost to produce. The success, and the value, of EEBO-TCP depends on the support of many libraries coming together. By sharing this expense, the labor-intensive work of EEBO-TCP is not too much for any single library to bear. In addition, because all of the work is done by libraries, and about 80% of it is funded by libraries,[24] we ensure that these texts remain within reach of scholars and the general public.

The History:

Early in 1997, a collaboration was proposed between UMI[25] and the University of Michigan Library to create searchable electronic text for the works represented in EEBO—which was not even on the market yet!—with the support of around 150 partner libraries. From the very beginning, this project was conceived as a library partnership with close ties to a commercial publisher, but *not* a commercial endeavor. From 1999-2000, UM library staff, including Hillary Nunn and Mark Sandler, promoted the TCP at conferences and online, encouraging libraries to commit their support. They needed to get a critical mass of support so that all the partners would agree it was worthwhile to go forward with the project. In fact, the first year of production went forward on funding from ProQuest, while partner libraries' funding was held in escrow until a sufficient number of partners had committed.

---

[22] http://www.textcreationpartnership.org/partners/

[23] The Association of Research Libraries (http://www.arl.org/)

[24] ProQuest contributed 20% to the cost of EEBO-TCP Phase I. Their contribution to Phase II is closer to 10% of the total project costs.

[25] University Microfilms, Inc., or UMI, was a branch of Bell & Howell, which later became ProQuest

Around 50 of the total EEBO-TCP Phase I partner libraries committed right away, by 2000 or 2001. The rest joined up gradually over the next nine years—including a number of partners from unexpected places. In the project's Winter 2001 newsletter, Mark Sandler wrote, "EEBO's success as a teaching tool has engendered unanticipated support from small and intermediate sized institutions. Our early thinking about ARL support was too limited."

In addition to being the backers and rights holders of the TCP texts, partner libraries also have a history of active participation in the project's governance. In the first years of the project, librarians and scholars at partner libraries ran the governing board, academic advisory board, text selection task force, encoding task force, and interface task force. All of these groups met very early in the life of the project, and disbanded after a few years (or fewer), except for the governing board, which met annually through 2007.[26] This governing board, backed by the promise of support from more than 20 partner libraries, determined that EEBO-TCP would continue on with a second phase of work. For many years, EEBO-TCP also held annual meetings at the annual meeting of the American Library Association, creating an opportunity for partner libraries to get updates on the project and mingle with project staff and with each other.

Today and in the future:

With the advent of EEBO-TCP Phase II, the governing board was re-created as a leaner executive committee, which now speaks by phone about once a year, and has occasional email contact in between. This decision was meant to lower overhead on a project that needed to put every dollar it could toward text conversion and that, by 2009, was a fairly well-oiled machine. While this change was justified, at the same time, three consecutive years of quick staff turnover and other administrative changes have left EEBO-TCP more disconnected from our partner libraries than I would like. In 2012 we revived the tradition of hosting an update at ALA, and we seek to keep partners in the loop through social media. In recent years, however, most of our contact has been with scholars working with the texts, rather than with librarians.

The work of EEBO-TCP is labor-intensive and expensive. It costs about $.68 to key and encode a kilobyte of data and, depending on who's doing it and when, how hard the book is, and how well the vendors did, between $.50-$1.00 per KB to proof and edit the text. This works out to a total cost of $200-250 per book. All told, the EEBO-TCP Phase II project is estimated to cost about $10 million. $1 million of this comes from ProQuest, and about $1.3 million (after VAT) from JISC Collections. The project also receives annual royalties from ProQuest from sales of the Phase I texts.

The rest comes from library partnerships. Since 2008, the TCP has raised more than $5.4 million, putting the total raised so far for Phase II at $6.8 million, or nearly 70% of what we anticipate needing to finish the project. We have been generously supported by many libraries and organizations over the years: in addition to the $6.8 million raised for this phase, around $9 million was raised for Phase I! But because the job is so big, we still have quite a long way to go. To finish EEBO-TCP Phase II, the project must raise more than $3 million.

[26] Meeting minutes are published at http://www.textcreationpartnership.org/

To that end, we have begun to investigate new sources of funding that would help to sustain the project without limiting existing model of distributed library ownership. In 2012, EEBO-TCP submitted a bid to the National Endowment for the Humanities proposing to create a small sub-collection of early modern travel literature. If awarded, this grant would fund the conversion of around 2,000 thousand new texts. These, together with existing related texts, would be published as a freely available, stand-alone resource, and put back into the larger EEBO-TCP archive.

In addition, it is increasingly the case that humanities research centers, consortial projects, and other campus groups or organizations have funding with which they wish to fund or subsidize EEBO-TCP partnership fees for their university. Libraries are no longer the first or only point of contact.

So our second tenet, "owned by libraries," is also subject to revision. It is shorthand for: owned, protected, and distributed by those who created it, and not owned by a commercial entity. Because EEBO-TCP partnership is built on top of a library's existing EEBO license, partnering with the library has always been the most straightforward, natural approach. But it should not be only libraries who fund, participate in, and share ownership of the EEBO-TCP corpus. After all, or so we claim, EEBO-TCP is:

III. Made for everyone

Process:

Two major factors contribute to whether we can claim that EEBO-TCP has truly been "made for everyone":

- Everyone should be allowed to use it.
- Everyone should be capable of using it.

Neither of these has yet come to fruition quite as we dream it should.

With respect to the first: EEBO-TCP's agreement with ProQuest stipulates that after work is complete, ProQuest has a five year window of opportunity to be the exclusive licensor of the EEBO-TCP content, in order to recoup their investment in this work. At the end of five years, restrictions are lifted and the EEBO-TCP texts may be used and shared by anybody, although ProQuest can and will continue to sell access to the texts through the EEBO interface. When the exclusivity period comes to an end, it does not become forbidden to sell the texts, it simply becomes possible to give them away.

With respect to the second: we must examine the project's approach to both text encoding and web interfaces. The way the text is stored, and the way it is presented to the end user, are perhaps the two most important facts in determining whether someone, or anyone, will be capable of using EEBO-TCP. Most users only access the texts through ProQuest's EEBO interface. This platform has changed and added features over the years, but (other than passing on or making our own) suggestions and requests, EEBO-TCP does not have any control over

this. In addition, the texts are hosted on a site maintained by the University of Michigan, and mirrored at Oxford. Library partners also have the right to host the texts locally, through their own interface, though only a few have taken up this opportunity. Individuals affiliated with partner institutions may also request the source files, from the text of one work to the entire corpus. These source files are encoded according to a schema based on TEI P3. In order to obtain these files, users must contact the project directly.

History:

EEBO-TCP Phase I was originally projected to finish by 2005, with the first 25,000 texts released to the public by 2010. However, this work was not completed until 2009. As a result, the exclusivity period did not begin until 2010, just when original partner libraries anticipated that restrictions would be lifted. This exclusivity period will end December 31, 2014.

The release date for the EEBO-TCP Phase II texts will depend on when the project is finished. Just like Phase I, Phase II was proposed as a five year project. Today, we have committed funding to get through about 70% of the project. Whether we go further will depend both on how much money can be raised, and how soon: whether we can ramp up production in order to get more done before the end of 2014, or whether we will be able to continue on past that date, which will delay the start of the exclusivity period. We may well see a combination of both.

Like the access model, EEBO-TCP's encoding practice was determined early in the game. The DTD working group met in Washington, D.C., March 2, 2000.[27] This group was charged, among other tasks, with determining the appropriate level of encoding and sketching out a DTD and element naming conventions. The EEBO-TCP schema they developed fits between levels three and four of the *TEI Text Encoding in Libraries: Guidelines for Best Encoding Practices* document.[28] For EEBO-TCP, the purpose of adding markup is to replicate the structure of the book, so that a user who does not have access to the page images or the original book will still be able to make sense of the text. Although of course all markup is interpretive, the aim has been to capture what is on the page, not to add new information. The guiding principle has been, "it is better to do less than to mislead."

The working group intended that TCP markup could always be enhanced by individuals with particular intentions in mind, and indeed that is what has happened. From tagging tone in *Hamlet*, as Michael Ullyot has done with English undergraduates,[29] to adding part of speech tagging to every word in the corpus, as in the Metadata Offer New Knowledge (MONK) project,[30] the encoded TCP texts are a foundation upon which more detailed encoding can be built.

---

[27] The recommendations for that meeting are available online: http://www.textcreationpartnership.org/dtd-working-group-note/

[28] The present version of this document is available on the website of the TEI Consortium: http://www.tei-c.org/SIG/Libraries/teiinlibraries/. The TCP schema is based on an earlier version of this document, but the description of "levels" of encoding is much the same.

[29] http://ullyot.ucalgaryblogs.ca/2012/01/19/encoding-exercise-description-for-english-203/

[30] http://monkproject.org/background.html

Another benefit to EEBO-TCP's comparatively light tagging is that the works are human-readable in any text editor—but of course, most users come to them through some form of web interface. The original EEBO-TCP interface task force met at Northwestern University in July, 2002, to plan for the interface that would be hosted by the University of Michigan on the DLXS platform. The attendees pooled questions, concerns and suggestions for the interface, which was launched in 2003. Some suggestions, such as the ability to constrain a search to a poetry, drama, or other types of text; and to browse the entire corpus by title and author did become a part of the interface. Other suggestions—notably, variant spelling matching, and allowing access to the underlying encoded text—have never been incorporated—largely because interface has, for better or for worse, always been seen as secondary to the completion of the text archive.

In the meantime, other libraries have incorporated the EEBO-TCP texts into their own platforms, for example, PhiloLogic@NU, hosted by the Northwestern University Library. In addition, single or small batches of EEBO-TCP texts have served as the basis for many digital editions and other kinds of projects, from the Holinshed Project at Oxford to Kirsten Uszkalo's Witches of Early Modern England.

Today and in the future:

As we collectively count down to 2015, EEBO-TCP is implementing small, gradual changes that grant as much public access to the texts as we can, and prepare for the moment when these texts are suddenly released to the public. Since spring 2010, the UM interface has allowed the public to search the EEBO-TCP texts and to see a summary of search results, although they are not able to click through to the full text. Although at first this just seems cruel, it is no worse than Google Book's or Hathi Trust's snippet view, and has the same potential to be helpful. Even without seeing the full text, searchers may be able to tell whether or not it is worth tracking down a particular book. If they can isolate what they are looking for in a single title, they may be able to request the work through inter-library loan—in fact, in August we fulfilled our first such request.

In addition, Google has now crawled these public pages, lifting EEBO-TCP out of the deep web to its surface. People who do not know EEBO-TCP are beginning to discover it when scouring the web for obscure titles, and seeking us out learn more. In the last year we have heard from a middle school teachers, a journalist, and a professor at a small liberal arts school that does not have EEBO. This can be frustrating on both sides, because we cannot always give these seekers exactly what they want. However, it is a small step toward breaking down the EEBO-TCP silo and engaging more fully with the web.

How users will engage with the texts once they are released to the public remains to be seen. Of course, the most common access point—ProQuest's EEBO platform—will not become freely accessible, nor will the EEBO page images. The platform hosted by the University of Michigan will allow public access to the texts (links through to the page images will only be available to original TCP partners). We have already experimented with this with the texts produced by the Eighteenth-Century Collections Online (ECCO-TCP) and will be ready to do the same for EEBO-TCP Phase I in 2015, and Phase II later on.

When the project began, libraries hosting their own interface were the only imaginable audience for local loading. Major cross-institutional text analysis projects had not, I think, crossed the minds of those who created the project. Yet, this is now the most exciting use of the TCP texts that I see on the horizon. Beginning in 2015, I expect to see more unrestricted distribution, manipulation, use, and re-use of the texts. This might range from distributing the books as individual e-texts in the Apple iBookstore, as James Cummings and Sebastian Rahtz have already done with the ECCO-TCP texts, to collaboration between major research institutions and small liberal arts schools that currently cannot afford to participate in EEBO-TCP, or even individual scholars for whom this is not an option.

This third leg, the right and the ability of anyone to use the texts, is the one that matters most in the long run—long after the library partnership fees are spent and the keying is done. Here, I see the greatest distance between EEBO-TCP in 2012 and where we would like to be. This is also the area where the decisions made a decade ago pinch the most, as approaches to research, to publishing, to collaboration in the humanities change faster than ever.

Conclusions and projections

As I have described it, not much about EEBO-TCP has changed since 2000—except the sheer amount of work that has been done. What is more, I do not anticipate that any of this will change very much before the end of 2014. The project has a mission to fulfill, a responsibility to partner libraries and to ProQuest, and a limited amount of resources with which to do it.

What has changed drastically, however, is the work that is being done with these texts. The original vision for the TCP was to support more, better, deeper consumption of the texts. But what I see now is not just consumption, but creation: researchers adding markup, generating new information, new platforms, new software, new tools, all based on these texts. It is very exciting.

While it is not the role or the responsibility of the TCP to pursue all of these paths—indeed, I think that would a misuse of our resources—it is our job to make sure that you can. We must develop an infrastructure for preserving, managing, hosting and distributing the texts in a way that makes sense to users and allows them to get on with their work. Judith Siefring is researching this problem with the Sustaining the EEBO-TCP Corpus in Transition (SECT) project, while at Michigan we are investigating version control and better managing releases and distribution of the texts.

EEBO-TCP has come a long way in 12 years, but there is still much to do. Transcribing by hand made EEBO-TCP worthwhile. Support from libraries made it feasible. But it is the creativity and collaboration of individuals—"everyone"—that will truly realize the potential of the EEBO-TCP corpus.