Rebecca Welzenbach
EEBO-TCP Update

[slide 1-2]
Good afternoon, and thanks for coming. This is the first Text Creation Partnership meeting we've had in several years, so I'm delighted to be here and have the chance to meet you. My name is Rebecca Welzenbach, and since last September I have been doing outreach for the TCP. Many of you may remember working with my predecessors: Hilary Nunn, Shawn Martin, and more recently, Aaron McCollough and Ari Friedlander. I have really enjoyed working with many of you remotely in the last year or so, and I'm delighted to have the chance to meet you in person.

[slide 3]
1:30 Snacks and socializing
1:40 Welcome, agenda, introductions
2:00 Update—what are we up to?
    Review of EEBO-TCP
    Progress on EEBO-TCP Phase II
    New uses and applications for EEBO-TCP
2:20 Presentations
    Erik Nebeker, University of California, Santa Barbara
    Harriett Green, University of Illinois Champaign-Urbana
2:45 Discussion
    Practical/logistical?
        Any problems with access?
        Promoting the resource?
        Questions about identification/cataloging?
        Questions about getting access to the files?

We haven't had one of these meetings for a few years. So, what are we here to do?

- Thank our supporters. Investment from libraries is by far the most significant source of funding for our work. We've now achieved more than 150% of what the initial TCP envisioned—we truly believe that this work changes the landscape of early modern scholarship, and it is thanks to you that is is possible.
- Keep you posted on our progress—partner institutions are investors in a living work, so we want to let you know how it's doing.
- News on initiatives EEBO-TCP is involved in
- Provide background and context for the TCP-curious—what is this thing, and we are we still railing on about it after all these years?
- Discussion; share ideas for what to do with the texts
    - Presentations from one librarian and one scholar; invite you all to share what's going on on your own campuses
    - Express concerns, ask questions

Let's keep this whole thing conversational—please feel free to add comments or ask questions at any point. Also feel free to get up for more food, or coffee if I'm putting you to sleep.

Help yourself to refreshments and to informational materials throughout. The “create a quire” is for the crafty or the bibliophiles among you. It's basically reverse digitization: pages from our project website

in book form, printed so as to put together a single gathering, or quire, for a book printed in quarto, like many of the Books in EEBO would have been. You can staple the gathering together, then cut your pages, or leave them uncut, as you wish.

[slide 4]
Let's start by introducing ourselves—please let us know why you're here.

[slide 5]
OK, great—a a mix of [ ] and [ ]. Really quickly, I'd just like to review what EEBO-TCP does, and why we do it—I know this will be familiar to most of you.

ProQuest's Early English Books Online contains digital page images of about 128,000 pre-1700 English books, microfilmed and scanned from libraries all over the world—with significant representation from the British Library, Huntington, and Folger Shakespeare Library.

The page images from EEBO are keyed by vendors, Apex CoVantage and SPi Global, and XML mark up is added. The result is a file that does more than provide searchability behind an image. Unlike the searchable text in most digitized collections, it can be displayed and used on its own.

[slide 6]
The text files come back to the University of Michigan where staff there, and colleagues over the ocean at Oxford, share responsibility for proofing the transcriptions and editing the mark up.

These files are delivered back to ProQuest, where they become part of the EEBO product. They are also hosted at the University of Michigan library through a slightly different interface. As well, a number of our partner institutions host the text on their own servers. Notably, Northwestern University hosts PhilLogic. I want to emphasize, since this will be important later, that for the TCP, our output is the text files—this data set. not any particular website or interface. The TCP offers one interface, ProQuest offers another, and other institutions offer yet more—these are all distinct tools with their own strengths and weaknesses.

[slide 7]
Why do we do all this? Due to the nature of these materials, it's not possible at this time to subject these texts to Optical Character Recognition. While EEBO on its own of course has searchable catalog records, it is not possible to search inside the full text of the books. Here's an image of Sebastian Brant's 1509 Shyppe of Fooles, taken from the Hathi Trust. This is a version digitzed by Google.

[slide 8] Here is the EEBO-TCP rendering of the same page.

[slide 9]

The initial scope of our project—TCP Phase I, which at the time was simply the TCP to end all TCP's--was to convert 25,000 selected texts. This was achieved in 2009, and with the support of many of you, the decision was made to continue on with a second phase. The goal this time was more ambitious: to key one edition of every unique work represented in EEBO—around 70,000 works in all, including the first 25,000.

Through the end of 2014, the Phase I texts are available exclusively from ProQuest, as an add on to EEBO (for those schools that weren't initially partners). In 2015, the final promise of EEBO-TCP will

begin to be fulfilled, when the text files for these books will be freely released to the public. (Access through EEBO will still be an add-on).

Phase II has been going strong since 2009. We've published close to 15,000 Phase texts so far, with another 4,000 or so to be published later this summer.

[slide 10-11-12] So let's pause for a minute here—in December of 2011 EEBO-TCP published its 40,000th text—that's more than 50% more books than were ever imagined possible when the project got underway. An incredible collection has been built here, and it is thanks to the support of many, many libraries that it was possible. So, congratulations, and thank you, for your support, interest, and investment in this project.

[13]
This model for sharing the cost and the reward of partnership benefits everyone. The cost for an ARL library to join EEBO-TCP Phase I was between $50,000 and $60,000 (depending on when you joined). It costs between $200-$250 to produce the encoded text for an average EEBO book. Now, there's no denying this is really expensive, labor-intensive work. But with 25,363 texts in the set, this works out to between $1.95 and $2.25 per book, per library. Our partnership fees are on a sliding scale depending on the size of your institution, so smaller schools pay less per book.

In Phase II so far, we have fewer partner libraries and we have done fewer books, so the per book cost right now it higher, but poised to drop: it will come down quite a bit as soon as we've released our next set of texts later this summer. With the funding that we currently have committed to the project, the per book cost will come down to about $1.67. The more partners join, and the more books we do, of course, the more that number goes down.

[14]
By sharing the expense of keying and encoding, each partner library enjoys rights to tens of thousands of carefully produced texts, while paying around 10% (or less) of what it costs to produce them.

[15]
So, speaking of currently committed funds, where do we stand? It works out roughly to 40% commitments from partner libraries, 20% commitments from other sources, and 30% yet to be raised. We are very grateful to have $1 million commitment from ProQuest, and a $1m pound commitment from JISC Collections in the UK which, less VAT and given current exchange rates, works out to about 1.3M. So far we have just about 67% uptake of Phase I partner libraries to Phase II, and we're seeing increased interest from libraries that were not originally Phase I partners.

In the last year, we welcomed University of Kansas, Cornell University, University of Texas, Michigan State University, and University of Nebraska-Lincoln on board with Phase II. University of Kansas, Brandeis University, University of Western Australia, Victoria University in New Zealand, University of Melbourne, all got Phase I from ProQuest, so we hope they will consider joining Phase II.

The estimated cost of the EEBO-TCP Phase II project was about $10M—we have just over $3M to go to get through the corpus using our existing methods.

Let's move on from business for now—I'm happy to return to this in the discussion or address questions individually if you like—and get to the exciting part: what can you do with these projects?

I have some ideas, but I'd rather hear now from those on the ground, actually doing this—and I suspect you would, too. So, please let me introduce our two guests, who will each share with us a bit about how they and their users are using the EEBO-TCP texts in research, teaching, and beyond.

First up is Eric Nebeker, a lecturer in the English department at the University of California, Santa Barbara. Eric drove over specifically to join us here today, so I want to thank him for that. Eric is an early modernist and works on the English Broadside Ballad Archive at UCSB.

[Eric]

Thanks!

Next up is Harriett Green, the digital humanities and English librarian at the University of Illinois, Urbana-Champaign. Harriett works closely with faculty doing really interesting work with text mining and text analysis, and also oversees the MONK—Metadata Offer New Knowledge—project within the UIUC library.

[thanks, Harriett!]

The most obvious use of TCP has always been to enhance EEBO with searchability and with the posibility of displaying modern text—a purely consumptive use.

There are other exciting things being done, taking this data apart and putting it back together again in new and exciting ways. The metaphor that I've been thinking about lately is that EEBO-TCP isn't making bread, we're making flour. What I mean by that is, bread is ready to be consumed—that's the interface through which you access and digest digital materials. That's not really what we make. What we do is pick the wheat—that is, key the text—and mill it—that is, process it, adding markup and other kinds of information. What we have to offer is flour, and we hope that it will go into not just bread, but cakes, pasta, donuts, and maybe even paste.

In building EEBO-TCP, we've created—to my knowledge—the largest corpus of encoded text data in the world. As a result, it's of interest not only to early modern scholars, but to anyone interested in text analysis, in XML, and, in fact, in improving these technologies.

The work done by the TCP has not only enabled new kinds of early modern scholarship, it has created the data needed to improve the very ways this kind of work can be done!
What do I mean by that?

- Abbot
- OCR projects at Texas A&M and MITH will use TCP texts to establish “ground truth” against new methods for developing OCR can be measured.
- Fuzzy searching
- It's worth noting that these are projects seeking and getting grant funding from Mellon, NEH, and other places. It saves lots of time and money to have a massive data set of encoded text available to you.

How can you get involved?

- Tell us about projects, classes, uses of the data at your institution. We'll post links on our site, or feature them in a blog post
- Encourage faculty and students to contact us with questions and requests
- Remind users that they can take and use these texts locally and build new kinds of things with them.
- If you're not already a TCP partner, talk about it with your library and your users. By adding Phase II, you'll bring your library's TCP access in line with the rest of the major research libraries, as well as contribute directly to the success/progress of our work. If you are at all interested in Phase II partnership, please talk to me. Our introductory rate that has been frozen for several years is expiring this year, and after July 1, our standard partnership fee will be in effect.
- If you don't have Phase I yet, contact your ProQuest representative for more information

Discussion:

Let's pause here—refresh your coffee, hit the restroom, fold a quire, and let's continue the conversation.

I'd love to hear what kind of use there is at your schools for EEBO-TCP. I'm also interested in hearing your ideas, thoughts, questions.

What do I want people to go away with?

The feeling that EEBO-TCP is a living, vibrant thing

The feeling that their investment is worthwhile

The sense that this is not just something to be consumed, but a data set; a foundation for new kinds of work.

Ideas about things to do on their campus to increase use

Sense of community around the project