Making the Most of Free, Unrestricted Texts: a first look at the promise of the Text Creation Partnership

Rebecca Welzenbach
Text Creation Partnership Project Outreach Librarian
Presentation given at Representing Knowledge in the Digital Humanities
September 24, 2011
Lawrence, Kansas
rwelzenb@umich.edu

[slide 1]
Good morning. My name is Rebecca Welzenbach, and I'd like to talk to you today about "Making the Most of Free, Unrestricted Texts: a first look at the promise of the Text Creation Partnership."

[slide 2]
In April 2011, the Text Creation Partnership announced that 2,231 XML-encoded, eighteenth century texts were now free of all restrictions, available to be used by anyone, for any purpose. Previously, these files were available only to libraries who had paid in to support their creation. This is the story of what happened next. It's a story of innovation, and experimentation made possible by not locking down these texts. It's a story of unexpected chasms between user expectations and project scope. And, I'd be lying if I didn't admit its also a story of the existential anxiety of a relatively successful, long-running project that is suddenly facing--maybe--a real shift in mission and focus.

[slide 3]
But let's step back for a moment--some context will help to explain what all of this means and why you might care. The Text Creation Partnership, or TCP, is a library-based initiative to produce XML-encoded editions of early print books, in particular, books from the Early English Books Online, Eighteenth Century Collections Online, and Evans Early American Imprints databases. If you've ever used the full text searching option in EEBO, you have used the TCP.

[slide 4]
The project has been around for around 12 years, and in that time, we've worked closely with both commercial publishers and libraries. In short, tpublishers grant us permission to use their page images. These are manually keyboarded by a vendor to create machine-readable text, and encoded using a schema derived from TEI P3. This work is funded in large part by academic libraries, who pay in to become partners in our work. These libraries are co-owners of the texts we produce, and gain access to them immediately, but all agree that for a period of time, they will not make them publicly available or share them with non-partner libraries. For a certain

part of that time, the publisher associated with the texts has the exclusive right to act as the puveyor of this resource, selling it to other libraries who haven't yet joined up. I want to make clear that this is not a copyright restriction--these texts are in the public domain and have been for many years. Rather, it's a legal agreement, a license agreement, really, that all parties will respect this rule. Central to our project, however, is the fact that after those five years, all restrictions are lifted, and the texts are freely available for anyone to use. We're still in the middle of our work, and so the release date for most of our texts is some years off. One date that many people are anticipating eagerly is January 1, 2015, when the first 25,000 EEBO-TCP texts will become available.

[slide 5]
The sudden release of the ECCO-TCP texts was not part of the plan--it just sort of came to pass, and has given us an opportunity to preview what we might expect down the road when the EEBO-TCP texts are released. What happened is this: ECCO-TCP simply never took off in the same way that EEBO-TCP did. We could not garner enough support from partner libraries to keep it going, and so in 2009, we had to call it to a halt.

One of the key factors in ECCO-TCPs demise is almost certainly the fact that, Gale already offered a full-text search option for the entire ECCO database, based on OCR. Now, certainly there are questions about the quality and sufficiency of the OCR for a project like this--there's actually an article by Patrick Spedding on this very subject in a recent issue of *18th Century Studies*. But in any case, By mid-2010, we found ourselves with 2,231 encoded texts that were, so to speak, all dressed up with nowhere to go. They'd been painstakingly created, but with only a small number of people--those who did sign on as ECCO-TCP partners--could use them.

[slide 6]
I mentioned before that for all TCP projects there's what we call an exclusivity period, where only supporting partners can acces the texts, and the commercial publishers have the exclusive right to sell them. When all goes well, this serves to benefit all those who have invested in the project: supporting libraries get the privilege of immediate access; supporting corporations get a window to make some money back on their investment, and the long-term interests of all libraries and scholars are protected by a very specific limited term before restrictions are lifted.

[slide 7]
In ECCO-TCP, we discovered, that all three of these aspects were failling: libraries were't signing up, Gale wasn't selling the product, and it turned out that in our agreement, the terms of the exclusivity period were rather vaguely defined. So, we approached Gale with our reasoning, and they gave us permission to lift the restrictions on the texts--to make them freely available, and to invite others to do the same.

[slide 8]
And there was much rejoicing, right? Yaaaaaay. Well, sort of. ^

[slide 9]
In fact, reaction were mixed. From the group that we might call power users, or who I've come to think of as textperts, there was generally jubliation and eagerness to experiment. Here, are a couple of examples, ranging from transforming the files, individually, into different formats to meet different needs, to analyzing the corpus at a very granular level--like, the level of the character--to find patterns of errors and omissions--which might then be corrected.

[slide 10]
From our average users, though, we were met mostly with frustration. This was, we learned, because there was actually a pretty significant gap between who we thought we were and what we thought we we ought to provide, and what these users wanted from us.

[slide 11]
You might be familiar with the distinction attributed to Richard Stallman between free beer--that is, it doesn't cost you anything--vs. free speech--that is, you're at liberty to do as you wish. ^ These are typically contrasted with one another when speaking about open source software. In the case of ECCO-TCP, though, we were actually offering both: the texts were available at no charge, and anyone was free to do as they liked with them. But we quickly learned that this wasn't quite what many people expected of us.

[slide 12]
What they actually wanted was free delivery. They didn't just want texts--they wanted an interface, a search engine, a URL to point to. ^^ Not data that in some abstract sense, they were allowed to use, but--not unreasonably--actually a place to retrieve it and and environment in which use it.

[slide 13]
The truth is, we were not very slickly prepared for this announcement. It all happened rather quickly, and we were eager to get the news out. This was the first time we delved into releasing texts to the public, and we weren't even sure if there'd be much interest. Perhaps not unlike the emancipation proclamation, we ceremoniously announced that the texts were free--but we didn't actually have the resources at hand to help people start using them right away--and people called us out on it, as they should. We also learned quickly that many users did not care to read our announcement (or explanatory blog post), or if they did, they actually couldn't make head or tail of it. In this example, a blogger perfectly fairly complains that it's not easy for him to get at the texts--but gets a number of other details wrong, about where and in what format the texts are available. ^ And I don't blame him, frankly because we didn't do a good job of making this information easy to find--this

is just an example of the kind of thing I'm increasingly turning up online as I dig, that make it very clear to me that we need to do this better in the future.

[slide 14]
So this is where the story sort of splits into two stories, in response two these to major categories of reaction that we got to our announcement--though the twain shall meet again a bit farther down the road. First, the innovators: what are they up to? What are all the ways the ECCO-TCP texts are already being used and represented online? Second: How do we represent our work so that in the future, all of this can be clearer and more useful, so that folks understand their rights and freedoms, and what, practically speaking, to do next? Do we need to change what we're doing? Or just how we present and disseminate information about ourselves?

[slide 15]
So, we'll start with the first and most basic. Anyone can download from the University of Michigan the original XML/SGML encoded texts and headers produced by the TCP.

[slide 16]
It is also possible to download the plain text files (stripped of XML markup) and an index containing metadata from the ckan Data Hub (Thanks to 18thConnect for distributing the plain text files, and to John Levin for making them available in a central, open location). This came out of the blog post that I just showed you--the guy who was frustrated with how difficult it was for him to get the texts.

[slide 17]
Sebastian Rahtz at Oxford has created compliant TEI P5 versions of each file, and also an epub version of each. These are publicly available to download.

[slide 18]
Anyone can search the ECCO-TCP corpus and view results via ARTFL's PhiloLogic search engine, hosted by the university of chicago. This, for what it's worth, is realy what the basic users are looking for: the ability to do fairly complex searches in a controlled environment. (Thanks to Robert Morrissey)

[slide 19]
Original ECCO-TCP partners--those who paid in to support the project--have access to the texts, along with links to the page images in ECCO, on the University of Michigan's hosted version of this resource. We haven't made this one publicly available yet because the images can only be shown to ECCO subscribers.

[slide20]
Keith Alexander took the metadata provided by John Levin at ckan, and started creating an open, linked dataset from it, that will hopefully allow information about the ECCO titles to be connected with other information--place and person names, dates, etc.

[slide 21]
And that's not all: just last week, I got a message from Perseus asking to include the texts in their digital archive. I agreed, of course, and pointed them to where they could download them (either our version or Sebastian's from Oxford). So, this brings us up to date. ^v^ What do you think? I for one, am thrilled that so many people want to use the texts, and that they are available--that this knowledge is represented--online in so many ways: different versions of XML, text, downloads, interfaces, indexed for search, metadata published, etc. The people in the TCP shop don't know how to do all of these things, and frankly don't have the time: after all, they are had at work continuing to create new texts. This, to me, feels like sort of a classic example of a DH, open access, public domain success story: by opening things up for folks to experiment with, we all wind up with so much more than you do if you just cling to it yourself.

And yet. Do you see what I see? Does your blood pressure begin to rise, as mine does, when you see all those windows piled up on each other on that slide?
OK, so I did that on purpose, to stress you out and stress myself out. But the thing is, because of this proliferation of representations, this slide is maybe not so far off of what the desktop of an ECCO-TCP user might look like. Does this stress anyone else out?

[slide 22]
It stresses out the folks at the JISC digitization programme, that's for sure. This is an excerpt from a blog post they published earlier this summer, which focused in particular on EEBO, but could easiily be extrapolated to the case of ECCO in this situtation.

[slide 23]
We have thus far been interested in data--but now there are proliferations of representations of that data--and therefore a proliferation of concerns. Is our goose cooked?

No, no, I don't think so--but there sure is a lot to think about. Our response to the JISC post was that

Our response to the JISC blog post was that the TCP texts were never intended to be a destination, and end point. Instead, they were meant to be a data set, a starting point, that would be fed into any number of applications. I still think this is right, and I'm pleased to see that it is actually happening--but I'm also dealing, with a bit of public domain anxiety: trying to balance my desire to provide a useful, high-quality product, with protecting everyone's ability to do what they want with this content. I find myself sitting on the fence. On the one hand, it's not our job to try to control everything that happens with the ECCO-TCP texts. In fact, people *should not have* to ask our permission to use the files. They're in the public domain, with no

restrictions on their use. And we feel strongly about respecting and maintaining that. And yet--would we all benefit from some way of keeping track of the different versions available? Of incorporating corrections made by one scholar in one subset of texts back into the corpus so everyone can benefit? Some way of documenting provenance so that it's clear, for example, that Ketih Alexander's Open Linked Data is based on metadata not distributed by the TCP, but grabbed from some other source? Well, yeah, I think so.

And this place of uncomfortable ambiguity is, frankly, is about where my story ends for now. This is the landscape that, after many years of pretty clearly-defined work, the TCP is trying to navigate. I don't really have the answer, but I would be grateful for your feedback--I think I might be particularly interested in the perspective of those who have published datasets--that is, information created under a very specific set of circumstances, but then sort of turned loose to the world--as I think that's sort of the most useful analogy here.

[slide 24]
Thank you!