2013-04-04

Books Without Covers: Binding the EEBO-TCP Corpus

Welzenbach, Rebecca

Rebecca Welzenbach

Renaissance Society of America 2013

Books without Covers: Binding the EEBO-TCP Dataset

[slide 1]

Good afternoon--as Michael said, I am the outreach librarian for the Early English Books Online Text Creation Partnership, or EEBO-TCP, a project based at the University of Michigan Library.

[slide 2]

As many of you will know, our mission is to create accurate, searchable electronic text for the titles represented as page images in EEBO. We're funded by a network of libraries, all of which jointly own the XML-encoded text files that we are creating.

[slide 3-7]

Our product is just that: the files themselves, almost 50,000 of them so far, and growing by a few hundred every month. Our interest is in producing the data; we leave the packaging of it into scholarly web sites—digital editions or collections—to others.

[slide 8-11]

And indeed, the EEBO-TCP texts have been indexed into many different sites, from ProQuest's EEBO, to an interface hosted by the University of Michigan, to a new platform created by JISC Collections in the UK, to special projects featuring subsets of texts, such as poetry.

In my experience, this is can be a difficult idea for new users of EEBO-TCP to grasp. People I speak to often seem troubled by these different points of access, and the fact that each looks and behaves a bit differently. I want to suggest, though, that what's going on with EEBO-TCP is actually is no different from any other digital edition or collection: the interface presents the data in a particular way and this shapes how we understand it and what we're able to do with it.

The difference is just that EEBO-TCP is one of the few instances where we actually have multiple interfaces to choose from. The opportunity for comparison makes clear what we foten take for granted about sites where content and interface are more tightly coupled: that the interface, the application, is not just a glass through which to reveal the text or data. Each one reflects the priorities, preferences, and interests of its creator (or perhaps commissioner) presenting a slightly different view on the data. Once we can see the way different interfaces put the same data into different contexts, it is seems obvious that these interfaces, must be studied and understood in order for us to make sense of these sites as bibliographic objects/artifacts; and really, for us for us to read them responsibly.

[slide 12]

There is a lot of discussion out there about how to evaluate digital editions and digital scholarship, but most of it is geared at tenure and promotion committees, to equip them with the background they need to evaluate the scholarly value of entire digital project, just as they would a book or article, when considering a candidate for promotion. Laura Mandell published an open letter on this topic through Texas A&M's Initiative for Digital Humanities, Media, and Culture. The MLA, too, provides recently updated Guidelines for Evaluating Work in Digital Humanities and Digital media, and other examples exist.

Today, I want to focus on something slightly different: evaluating these sites from the perspective of a researcher, and understanding them as cultural artifacts in the same way that a book historian might. In fact, in the realm of book history, we already are used to navigating this tension between content and its wrapper that feels a bit strange as we look between the covers of a website. We expect that a book may exist in many copies, each bound uniquely according to the preferences, budget, and values of its owner or perhaps its seller. With this in mind, I suggest that our practices for studying the bindings of early books may shed some light on how to approach the digital bindings I am describing here.

[slide 13]

I will turn now, with gratitude, to the example provided by the Folger Shakespeare Library's wonderful Bindings Image Collection. In addition to beautiful, high resolution, CC-licensed!! images, this database, launched in 2012, records a significant amount of information about each binding.

[slide 14]

Most of these fields can be adapted/interpreted to apply to the sort of digital "bindings" I'm talking about. That is not to say that every field applies, or that these are the only things we might ask about the digital bindings we encounter. But it seems sensible that the details we take for granted as being important about physical bindings, we should at least think to ask about digital ones.

In the time that remains to me, I'd like to walk through these fields, and look at how they might be applied to better understanding two of the most widely used EEBO-TCP interfaces: the one provided by ProQuest and the one provided by the University of Michigan Library. The Bindings Image Collection divides the metadata about each binding into three groups. First is summary information:

Summary Information: Where, when, and how was this binding produced?

[General Description → What is the site's purpose and major function(s)?

This might not be immediately obvious. Just as with a binding, you may need to spend some time with it first to come to an accurate conclusion.]

[Slide 15] **Binder** → Who produced this site?

In this case we can see two logos on the EEBO page, indicating a relationship between these two corporations—if there was a merger or buyout, that could be of interest. In this case, ProQuest acquired Chadwyck-Healey but felt there was value in maintaining that brand name.

In contrast, the Michigan interface does nothing to tell you who made this site! You have to infer it from the help link (University of Michigan Digital Library) and the URL, umich.edu. Further digging might reveal that this is one of many similar sites maintained by the University of Michigan Library. It is making the assumption that you know this already—but there's no reason that you should.

[slide 16] Country/Style → Is there any association with national or regional funding bodies, organizations, or networks (e.g., NEH, JISC, ADHO, RSA, Europeana)?

This may tell you something about the values, priorities, limitations, and requirements of the site. It also places it in conversation with other sites, other projects, affiliated with the same bodies. In our case, there's not much to see on the Michigan or ProQuest sites—but it is worth noting that JISC Collections is building its own platform for these texts—along with many others.

[slide 17] Period → When was the site established? When was it last updated? Is the site/interface contemporaneous with its content, or was it created later?

The ProQuest site clearly indicates that it has been in action continuously since 2000. Once again the Michigan home page falls short in this regard. But what Michigan's interface does, which ProQuest does not, is show the date that each TCP text was published online. This reveals Michigan's primary interest in the TCP (as opposed to EEBO's in the images) and in the progress of the project. As a consumer, you might ask whether there is a difference between texts completed early in the project and those that came later (the answer is yes). This interface captures that.

Bibliographical Information: Where does this artifact fit into the scholarly landscape?

[slide 18] Call Number → Or in this case, URL. Capture the URL for the main site. Are there stable URLs for sections within it the site (e.g., images, texts, or citations)? What parts of the site lack stable, citable URLs (e.g., search results pages)? This will tell you what the project thinks it is important for you to be able to cite, or not—or perhaps simply reflect the priorities,

technology, and resources at the time. EEBO allows you copy a stable link to the catalog record for an item. From there, you can get to the images, text, thumbnails, etc. In Michigan's interface, you can do something similar: our permanent URLs point to the "Table of contents" view, which the metadata for the title, and a hyperlinked table of contents derived from the structure of the book. Both sites are constructed in such a way that images or segments of text are called dynamically when a user requests them, and therefore although it would be desirable to have a permanent link to a single image or paragraph, at this point it is not feasible. You can guess at this by watching the URL as you do searches or page through books.

Provenance → Has the project changed hands, moving from one institution or publisher to another? Have the layout or functionality changed over time? How did the data wind up here? For openly accessible sites, try the Internet Archive to see how it has changed over time. Changes in publisher may mean changes in platform, which may mean changes (deliberate or accidental) in how the data looks, and how it can be used. In our cases, the interfaces have changed very little in the last ten years—to our chagrin.

[slide 19] Author/Title/Created/Published of items bound in → What data is being presented in this site? When and where was it created, when and where was it digitized? For ProQuest, it is MARC records, page images, text.

Binding Description: What kinds of use and points of view are privileged by this binding?

[slide 20] Dimensions → What/how much can you view and download (e.g., hi- or lo-res images, PDFs of complete works, full text)?

[slide 21] Covering material → How does the site present itself? Does it have any of the following: homepage, explanatory material, tutorials or demos?

[slide 22] Sewing/Headbands → How is the whole resource held together (e.g., XML, relational database, content management system, blog)? How is the data "tied" to the interface?

What are the points of entry from one to another (e.g., searching, browsing, tagging, sorting, annotating)?

[Slide 23]Edge treatment → Are there animated page-turners or other multimedia features? What purpose is served by any particularly flashy or beautiful elements?

[slide 24] Closures → What do you need to get into the site (e.g., log-in, subscription, institutional affiliation, special software or plug-ins)? Can you extract/download data from the site?

Decorative description → What impact do the look, feel and layout of the site have? What experience does it aim to create (e.g., reading room, lab, museum, social media platform?) Does the site's appearance suggest continuity with other sites or institutions?

[slide 25] Technical description → Building on the concepts you considered under "sewing," what lies beneath the surface of this site (e.g., Javascript, XML, Flash, Wordpress)? Does it use open standards, or proprietary formats? Is the anatomy of the site documented/discoverable, or not?

I think EEBO-TCP is one of the first to distribute its corpus on a large scale in the way it has. But it is by no means the last, As openly-licensed digitized materials increase and infrastructure for open linked data continue to grow, we're only going to see more of this.

[slide 26] Example: Dot Porter, Walters Museum MSS. Rather than the appearance of a coherent, unified, end-to-end product—a website that mimics a contemporary print book—I predict that we're going to see more and more of the same content bound many different ways. It will be necessary to read these interfaces on their own terms—or indeed sometimes to bind them ourselves in the way we need—in order to use these resources most effectively.