Collaboratively Curating Early Modern English Texts

Drafted by Martin Mueller

August 9, 2011

Collaboratively curating digital corpora of Early Modern English texts .. 2
The thre C's of data curation 3
Dividing the labor among people and machines and people 5
Sequential, intermittent, and cascading curation 6
Defining and controlling the quality of digital surrogates 7
Certifying individual texts 9
Curation and Annotation 11
The rewards of curation 12

Collaboratively curating digital corpora of Early Modern English texts

The following describes workflows for the collaborative curation of digital surrogates of imperfectly or incompletely transcribed early modern printed books, created in mass production environments. The texts in question are some 2,000 ECCO-TCP texts, which have recently passed into the public domain. They will be complemented by OCR transcriptions of 18th and early 19th century texts. The 18t century texts are transcriptions of Gale-Cengage images or Google books in the public domain. The 19th century texts are OCR transcriptions of texts in the British Library. The problems solved in the curation of these surrogates will solve most of the problems encountered in the curation of digital surrogates of old books, however transcribed.

Collaborative curation of digital surrogates occurs in a social and technical space where individual texts live as curatable objects continually subject to correction, refinement, or enrichment by many hands and coexist at different levels of (im)perfection. From one perspective, this is radically different from a collection of printed books, each of them finished before it is added to the collection. On a closer look, however, the digital space only speeds up (albeit by orders of magnitude) similar acts of curation in the print world, and it makes more visible the latent variance of (im)perfection in printed books.

There are two critical problems that need to be faced head-on in any discussion of such scholarly "crowdsourcing." The first has to do with acceptance by scholarly communities. The second involves reducing the time cost of curatorial labor so that scholars will find it quite literally "worthwhile" to participate at some level in such a venture.

Acceptance by scholars depends on being very explicit about matters that scholarly readers tacitly recognize and evaluate when they look at the editor and publisher of a printed text. There needs to be clarity about

1. The editorial procedures and structures of peer-review that govern the making, logging, and approval of curatorial acts
2. The current status of any text in the corpus with regard to its curation history and the quality standard that it claims to meet

There is no contradiction between saying a) that digital texts are perpetually curatable and b) that a particular text at a particular time meets certain quality standards as specified in an editorial policy and certified by one or more individuals with appropriate editorial authority. The lack of clarity about such matters is a major reason why scholarly communities are hesitant to trust texts from Project Gutenberg or Wikisource, "the free library that anyone can improve," even though in practice many of these transcriptions are quite good and serve many purposes well.

While the editing of texts in a print world is like the building and separate launching of many boats, collaborative curation of digital corpora is more like repairing a large ship at sea. Data exploration leads to moments of data curation. The easier it is to switch between exploration and curation the easier it will be to engage scholars in the work of collaborative curation. It is a challenging task for interface designer to build an environment that supports "curation *en passant.*"

The curation tasks discussed here refer to activities performed by users on data that are already in circulation. Nearly all of them can be modeled as acts of dispersed annotation. They do not include tasks of archiving or preservation, which are not easily modeled in that fashion and are likely to remain the responsibility of specialised staff. They may include the addition or correction of bibliographical metadata.

The thre C's of data curation

Basic curation tasks can be grouped under the imperatives of Complete, Correct, Connect. Higher-level forms of curation could be bundled under the heading Enrich and presuppose an environment that supports free-form annotation. Such curation is beyond the scope of this report, except to note that basic curation by user contributors is much more likely to happen if it can be made available in an environment that simultaneously supports these higher levels of curation/annotation, which

users are likely to find more interesting and more directly relevant to their projects. This point is discussed briefly at the end of this report.

Transcriptions are incomplete for any of the following reasons:

1. Letters, words, or longer passages that are illegible in the transcription source, typically a digital scan of a microfilm of the printed original. Missing single letters (marked as such) are by far the most common form of incomplete transcription in TCP texts, which contains several million instances.
2. Pages or entire sections that are missing from the transcription source. This is quite common and is typically, but by no means always, marked explicitly by the transcriber.
3. Passages that were intentionally omitted by the transcriber because they are in a foreign language or use mathematical or musical notation.

Transcriptions are incorrect because

1. The human transcriber or OCR process made errors in reading particular letters. Chief among these is the confusion of 'f' and long 's'. Incorrectly transcribed letters are the major source of error in OCR texts. Some passages are so garbled that it is easier to transcribe them from scratch.
2. In texts with structural markup the wrong elements are used. For instance in a play, a speaker label might be tagged as a stage direction and the other way round.

Connecting curation makes texts machine actionable by marking texts in ways that support apples-to-apples comparisons of related phenomena within or across texts and allow for their retrieval by query tools. It includes such things as

1. Mapping surface spellings to standardized lemmata and parts of speech
2. Identifying multi-word expressions
3. Determining sentence boundaries
4. Identifying passages written in other languages

5. Identifying names of people and places
6. Identifying citations (e.g. Biblical references)
7. Identifying longer quotations or passages that are shared across different texts
8. Identifying passages marked as spoken in written texts
9. Marking structural boundaries of chapters, scenes, and other subdivisions of texts

Acts of connecting curation are usefully seen as extending cataloguing practices from the bibliographical level of a book as a whole into its parts. Thus structural or linguistic annotation is a way of cataloguing the discursive divisions, sentences, and individual words of a book. Forty years ago, when it was a challenge to pack a million catalogue records into a mainframe of the day, that was an impossible dream. Today it is from a computational perspective not a particularly hard thing to do for quite substantial collections of primary sources.

Connecting curation is typically produced algorithmically, but should be subject to user review. User-generate corrections not only correct this or that passage but in the aggregate provide better training for iterative cycles of algorithmic curation.

The three C's of curation concern themselves with problems where the answer is nearly always obvious. They stop well short of textual criticism, although they may identify passages that require complex philological investigation, and a list of such passages is a useful thing to have. These types of curation are basic forms of textual housekeeping, but the difference between clean and dirty is as obvious in a text as in a house.

Dividing the labor among people and machines and people

Curation tasks have different levels of complexity. There are lots of little problems that can be solved by high school students who like reading old books. Transcribing a Greek passage from an early 16th century book requires not only a knowledge of Greek but some palaeographical skills. Adding or reviewing machine-generated structural markup requires some training but with proper software support is much easier than it was a decade ago and can be learned within days.

Collaborative curation allows for a division of labor that matches tasks with the skills of curators and goes beyond the scope of the proverb that many hands make light work. A more radical division of labor divides tasks between man and machine along the lines of the three R's of animal experimentation: Replace, Refine, Reduce. In a well-designed collaborative workflow you direct the crowd, or more accurately individuals in the crowd, to problems that need doing and match their skills. You reduce the time cost of getting to a particular problem, focus the attention of curators on the task that requires human intervention, but then relieve them of the burden of keeping track of what they have done and make the system do the work of reporting the curatorial act.

In the curation of old books a task that requires human intervention typically requires an inspection of the original text. Reducing the time cost of getting from the transcription to the original may be the most critical element in a good workflow architecture, because it reduces tedium, the greatest demotivator of all. Aligning a page with a page is good; even better is a horizontal or vertical alignment of lines, which puts the original and transcription instantly into a single field of vision (Typewright does this very well).

Sequential, intermittent, and cascading curation

User-based curation occurs in three work flows, which may be called sequential, intermittent, and cascading. In sequential curation users work their way through a text line by line or page by page. Users may work on a task at odd hours, and in that sense their work may be intermittent, but while they do it, they focus on it in a sequential manner.

What I earlier called curation *en passant* is intermittent curation in a stricter sense of the word. It occurs when users come across errors as they work with a text and fix it then and there. Books are full of corrections that readers made *en passant* in their own books. In a print world such curatorial acts benefit only the owner of the book (unless you corrected a library copy, which technically you are not supposed to do). In a digital world, the correction benefits all users.

Finding errors while working with data is very common and is a particularly important form of error discovery because it involves data that

are actually used. From the user's perspective, the time cost of correcting the error is a critical factor: users need to be able to switch quickly between modes of data exploration and curation. From a systems perspective, users who are too busy or just lazy and merely flag an error do almost as much good as meticulous users who record that "this 'f' should be an 's'. It will also be often the case that users can be sure that something is wrong but may not know the solution.

Cascading curation takes off from intermitten curation. Users encounter an error and wonder whether similar errors occur elsewhere in the corpus. The larger the corpus, the more productive this form of curation in which the tools for exploration double up as tools of curation. The workflow is similar to that of spellcheckers, but it requires a different interface. Users should be able to see a list of errors with keywords in context, which allows for one-by-one correction, batch mode correction, or a batch mode once a list of exceptions has been handled singly.

In the correction of orthographic errors, whether transcribed manually or through OCR, there are many algorithmic opportunities for using the three R's of animal experimentation to minimize human labor. Algorithmic corrections, in which degrees of confidence are expressed through color coding, reduce human labor by replacing the act of correction with a simple approval or choice between alternatives. The machine may also flag suspected spellings without making suggestions.

The data for such algorithmic corrections or flaggings should come primarily from a frequency-based lexicon of the corpus subject to curation but may be supplemented by other external lexica. Algorithmically suggested corrections will be more plausible if they are based on highly context-specific data. Context-sensitive algorithmically generated suggestions also work very well for supplying missing letters in incompletely transcribed words of TCP texts, especially if the missing letters are in word medial position.

Defining and controlling the quality of digital surrogates

Credible assurances about the quality of digital surrogates are essential to their full acceptance by scholarly data communities. "Good enough

for what purpose?" is a good question with variable answers. Computational linguists engaged in information retrieval from large OCR archives will put up with high levels of "noise" as long as enough "signal" comes through. Traditional scholars used to century-old traditions of meticulously transcribed diplomatic editions have a very low tolerance for error and high standards for what counts as a readable text. But computational linguists also like what they call "gold standard" texts, typically texts that involve at least some "manual" curation to reduce noise and improve the granularity of the signal. But in the trade-off between living with higher noise or putting up with the tedium of curation they typically settle for the former.

The key to quality assurance lies in some basic features of modern computers that are very familiar to IT professional but are still likely to strike ordinary scholars as quite magical. It is easy for computers to keep track of thousands of users and millions of curatorial acts they perform, logging each one carefully as who did what and when. It is equally possible for such computers to divide a large corpus into its individual words and treat each word as a distinct token with its unique ID. This can be done with corpora running into billions of words.

From the computer's perspective, the engagement of a user with a piece of text is a transaction that results in a log entry to the effect that at a particular moment in time a userID with certain properties changed, deleted, or added properties associated with one or more wordIDs . Algorithmically produced curation can also be logged in this fashion. The userID in that case is that of a machine running an algorithm. The record of such transactions is a curation log that may run into many millions of records. Think of a vast digital expansion of the multivolume *Berichtigungsliste* or "correction list" that Greek papyprologists have kept of their editorial work for almost a century.

There are many things that can be done with such a curation list, which is a permanent, cumulative, and systematic record of all user transactions. Most obviously, it can be used to change the text, but any such changes, whether they involve addition, alteration, or deletion are reversible and subject to further change. And of course such changes are themselves transaction that will be logged in the curation list and associated with the userIDs of the people or machines responsible for them.

Sorted by userID, the curation log lets you evaluate particular curators and give them performance-based privileges. A 'cadet' –to use the nomenclature in Zooniverse's Old Weather project – -might be promoted to lieutenant and acquire the privilege of reviewing the work of others while his own work could be judged to be trusted without further review.

Sorted by wordIDs, the curation log tells you about the frequency and types of error found in particular works. It may help "direct the crowd" in dealing with similar texts. For a curator working with a particular text, the curation list for that text has a variety of obvious uses.

Sorted by type of change, the curation log provides useful data about common types of error across the corpus and helps with the development of training data and algorithmic strategies to reduce the need for human curation.

The curation log is thus the fundamental management tool for maintaining quality control. It can in principle support quite different organizational choices for editorial review, whether highly centralized or widely distributed. In practice, it is likely that the best results will be found in distributed systems that give substantial control over editorial decisions to existing scholarly "data communities" in various scholarly societies and their sub-committees or interest groups. There are substantial technical problems that need to be solved in order to main a robust and flexible infrastructure that will enable such data communities to work independently while staying in touch.

Certifying individual texts

Corpus-wide collaborative data curation is likely to work best if any transcribed text in the corpus is tokenized and linguistically annotated at the point of entry. From that moment on it exists as an aggregate of tokenIDs with attributes or properties. It can then be submitted to simple statistical analyses that support reasonably accurate assertions about the quality of a text at its point of entry by comparing the distribution of tokens with their distribution in the entire corpus or some subset of it.

With TCP texts, the relative frequency of missing letters, marked as such, is an excellent proxy of the overall transcription quality. TCP texts are produced in something like a factory model of production, with little variance in the competence of individual transcribers. The biggest variable is the quality of the source. The transcriber's immediate source is a printed copy or screen image of a digital scan of the microfilm image of the printed page. Many things can and do go wrong in that chain. Confronted with a badly microfilmed image of a page printed poorly in blackletter type, transcribers will not only be unable to read some letters and words but are likely to make errors in transcribing others. The first printed edition of Marlowe's *Tamburlaine* is a good example of an error-ridden text. On the other hand, if a text has few or no missing letters, it is likely to be clean transcription in other regards as well. This is the case with a large majority of TCP texts. An initial rating, based solely on missing letters, is a reasonably good gauge of the overall quality of a text as a diplomatic transcription. Ratings can in fact be established at the page level: a digital surrogate with 150 missing letters on 120 out of 200 pages is a less trustworthy transcription as a whole than a text of similar length with 150 missing letters on three pages.

OCR based transcriptions lack a single criterion that can be used as a reliable measure of initial quality. It is, however, possible to measure the relative frequency of low-probability spellings, using as a baseline for comparisons the frequency of such spellings in TCP pages with no missing letters.

Data about the quality of a text at its point of entry would be a useful part of a "text information package" (TIP) that becomes part of its bibliographical metadata and supports a rating scheme, not unlike the eight grades from 'utility' to 'prime' that the United States Department of Agriculture uses to classify beef. By combining the initial algorithmic quality assessment with data from the curation log one can use a variety of algorithmic and human ways to upgrade the quality assessment of a text. This is not the place to discuss in detail the various types of such certification, but the equivalent of "prime beef" would be an assertion by one or more human editors with appropriate scholarly qualifications that the digital surrogate in question meets the standards of scholarly diplomatic editions familiar from the print world.

In principle, curation is like a mother's work and therefore never done. In practice one may want to cross the lyric about a mother's work with Winnicott's concept of the "good enough mother" and formulate something like a "Winnicott standard for the good enough digital surrogate." A text meeting that standard would be both machine actionable and pleasing to read. Many TCP texts already meet that standard, others can be brought up to it with relatively work, while some need a lot of attention. OCR based texts will generally require more work to produce surrogate that are pleasing to read, but may also benefit more from iterative cycles of algorithmic and human curation. Users will in the end spend their curatorial labor on texts they care about, but they may well appreciate advice on texts that would benefit most from their attention.

Curation and Annotation

Most acts of basic curation can be modeled as annotation. Changing 'afinine' to 'asinine' does not differ technically from the marginal gloss "this is the most important passage." There are obsessive people who derive satisfaction from the very tedium of prolonged stretches of data curation, but most people are not like that and find annotation more challenging and interesting than curation. If you can design an environment in which basic curatorial activities are a subset of annotation capabilities, the prospects for attracting and retaining user contributors brighten considerably.

In designing an environment that supports the simultaneous exploration and curation of textual data in a large corpus of digital surrogates it is best to start from the familiar experience of "reading with a pencil." In a single and uninterrupted workflow, the reader underlines this passage, corrects a typographical, glosses a foreign word, adds a reference, and comments approvingly or critically on a passage. All this is done with the same tool.

How close can one come to modeling the ease of such a workflow in a cloud-based environment? It may never be possible to match the informality and intimacy of reader, book, and pencil. But you can come reasonably close and offer powerful compensating advantages such as

1. Providing readers with a cloud-based private space where their annotations are available *ad locum,* but also in variously sorted forms.
2. Offering their annotations to a wider public, subject perhaps to editorial review
3. Contributing to collaborative curation as a public service

The rewards of curation

The four great motivators of human action are, in alphabetical order, duty, fame, love, money. It is possible to design workflows that support curation in exchange for micropayments in the manner of Amazon's Mechanical Turk. But in projects that aspire to being parts of some universal and public domain digital library micropayments are probably not the best way to go and more likely to undermine the willingness of volunteers to contribute some fraction of what Clay Shirky has called their "cognitive surplus" in the business of curating old books. It is likely that salaried staff and faculty will play a significant role in the technical and editorial management of collaborative curation.

That leaves duty, fame, and love: the amateur scholar, the citizen scholar, and everybody else who would like to be recognized for something useful or splendid that they have done. Any successful collaboration curation needs to find social and institutional ways of appealing to and satisfying these motives, separately and in combination. It will be important to work with scholarly and pedagogical organizations, but it is also important to think of globally scattered volunteers and bringing them together in circles of shared interest.