Back to the Future
or
Wanted: A Decade of High-tech Lower Criticism

By Martin Mueller
Northwestern University

1. About knowing the past digitally

I dislike the term 'Digital Humanities' and even more its abbreviation DH.[1] Despite its plural, it suggests that there is one thing that is properly called by that name and that the digital is its most distinctive property. If a term is bandied about a lot, it must be about something. For me the something of Digital Humanities is the trouble that the humanities have had in absorbing digital technology into their habits of work and recognition. Unlike the natural and social sciences, they have so far put the digital into a ghetto–a mutually convenient practice for those inside and outside, but probably harmful in the long run. The digital interacts in different ways with different disciplines, but in each case the interesting questions have to do with the ways in which that interaction affects the primary data, questions, methods, and working habits of that discipline.

I am very fond of the opening of Montaigne's essay about cannibals, where he writes

> We had need of Topographers to make us particular narrations of the places they have beene in. For some of them, if they have ... seene Palestine, will challenge a privilege, to tell us newes of all the world besides. I would have every man write what he knowes, and no more:... For one may have particular knowledge of the nature of one river... that in other things knowes no more than another man: who neverthelesse to publish this little scantling, will undertake to write of all the Physickes. From which vice proceed divers great inconveniences.

---

[1] The following is the lightly edited version of a talk I gave on November 18 2012 at the Chicago Digital Humanities and Computer Science Colloquium. I have liberally quoted or paraphrased stuff I have written elsewhere. I owe a great debt to the excellent papers at the conference *Online Humanities Scholarship: the Shape of Things to Come*, held at Virginia in March 2010 (http://shapeofthings.org). You can find my review of this conference at https://scalablereading.northwestern.edu/2012/08/13/the-great-digital-migration/.

So I will take you on a tour through my digital Palestine, which is the world of Early Modern English texts, hoping that the particulars of that world have some interest of their own and that some broader points will also emerge from a closer look at the particulars.

Historians are people with an interest in the past. Like Orpheus or Lot's wife they can't help looking back. For most people looking back is just a small part of their lives, and a good thing too. But imagine a world in which nobody would ever have any interest in the past. It would be like a world without either tears or laughter: a world from which some essential component of humanity had disappeared. History in the sense in which I use the term covers a much broader range of activities than are included in the work of departments of history. A deep engagement with the past is a constitutive part of the study of art, literature, music, philosophy, and religion. We engage with that past through the traces it has left. We want to "keep" them, sometimes because they are products of extraordinary skill and talent, such as Bach's preludes and fugues, but more often because they are all we have. To "keep" means to "cherish and preserve," a touching phrase that Penelope Kaiserlian used in a talk about the American Founding Era documentary editions and that, according to my colleague Barbara Newman, reaches back to the Middle Ages via a 19th century Benedictine litany (Kaiserlian, 2010). If the phrase is too sentimental for your taste, remember that it refers to the materials from which humanists construct, and the ground on which they rest, their arguments.

Many humanities scholars, myself included, are professional keepers of the past, and our interest in the digital is shaped by our sense of how it helps us with the "upkeep" of the little region of the past that is our field of expertise. The great 19th century scholar August Boeckh defined philology as "die Erkenntnis des Erkannten" or the further knowing of the already known. An iterative, cumulative, but unending engagement with a past about which we know something and will never know everything is the essence of our work. Some people think of the digital as a call to arms towards a brighter future.[2] I am all for a brighter future when it comes to health care

[2] Andrew Prescott says "It is an article of faith for me that You Tube is just as worthy of scholarly examination as an illuminated manuscript." Agreed, but it is an article of equal faith that an illuminated manuscript is just as worthy of scholarly examination as YouTube. Why then does Prescott in his blog on "Making the Digital Human" spend so much time on taking his British colleagues to task for failing to get with the program and remaining stuck in the past? It is probably not the case that attention to illuminated manuscripts is crowding out attentionYou Tube. (Prescott 2012).

or a fairer distribution of goods in the global economy. But as a humanities scholar engaged in "keeping" my little patch of the past I evaluate the digital in terms of whether it helps me towards a deeper, more granular, and more reflective understanding of the past.

As a humanities scholar you turn and return to the same things in the past. They are the "primary sources" that support your arguments or reflections and against which they are tested. Paolo D'Iorio puts this nicely in his paper from the 2010 "Shape of Things to Come" conference at Virginia:

> If a scholarly community intends to conduct research on a certain topic, it first needs to define which documents or objects to consider as its primary sources. When a research line is about to be developed and consolidated, a catalogue of primary sources is compiled, usually by archivists or librarians. ... … Catalogues of secondary sources come later … The distinction between primary and secondary has a fundamental epistemic value. According to Karl Popper, what distinguishes science from other human conversation is the capacity to indicate the conditions of its own falsification. In scholarship, the conditions of falsification normally include the verification of hypotheses on the basis of a collection of documents recognized by a scholarly community as relevant primary sources. (D'Iorio, 2010)

2. Living off a century of editorial labour

In the early 1850's Theodor Mommsen published the first volume of his Roman History and founded the *Corpus Inscriptionum Latinarum*, the systematic gathering of inscriptions from all over the Roman empire. For the next five decades he was the chief editor and a major contributor to its sixteen volumes, which transformed the documentary infrastructure for the study of Roman history. Since the early 20th century, a student of Roman history with access to a decent research library has had "at hand" a comprehensive collection of the epigraphic evidence ordered by time and place.

The *CIL* is a majestic instance of the century of curatorial labour that created the documentary infrastructure for modern text-centric scholarship in Western universities. You typically must do many things to your data before you can do anything with them. Curation or "doing to" is often tedious (not unlike the prep in painting). Doing exciting things with your data by way of exploration or analysis is much sexier. Mommsen's scholarly career combined in a quite unusual way the gifts of a brilliant narrative historian with the phenomenal energy and attention to detail that marked his editorial work.

Mommsen's work stretched across the two poles of 19th century scholarship, "Lower" and "Higher" criticism. In Classical Studies those poles are represented by the two giants of the next generation, Nietzsche and Wilamowitz, graduates of the same and very famous German grammar School Schulpforta. The tension between these two poles of scholarly work is wonderfully captured in Nietzsche's exasperated cry "Was hilft mir der echte Text wenn ich ihn nicht verstehe?" (What use is the authentic text if I don't understand it?) But their continuing war was the father of many good things. Well into the late twentieth century the documentary infrastructure of humanities research rested on what you might call a Delphic tripod of cultural memory, with its three legs of scholars who made editions, publishers who published them, and librarians who acquired, catalogued them, and made them available to the public.

In this world there was a very simple answer to the question "Who provides quality assurance (QA in modern business parlance) for the primary sources that undergird work in your discipline?" The simple answer was "my colleagues," and it might include "I do some of that work myself." Lower Criticism of one kind or another counted as significant scholarly labor and made up a significant, though gradually declining, share of the work of humanities departments. When I was a graduate student in the early sixties, a scholarly edition still made for a dissertation that would get you a good job. Comprehensive critical editions of major American authors were high prestige projects in leading universities. Edmund Wilson's attacks on these projects as wasteful exercises in pedantry were an important polemic in the early years of the *New York Review of Books*.

Harrison Hayford's Northwestern Newberry edition of Melville was one of those projects. From the late fifties through the seventies, it was the most important incubator of successful academic careers in Northwestern's English department. But when I came to Northwestern in 1976, the tide had turned. Wilson's attacks, unfair in many ways, did capture a change in outlook. A new generation of scholars felt that existing editions of most authors were good enough for most purposes. You no longer needed to take data curation seriously because a century of it had succeeded in creating a print-based data infrastructure that from now on you could take for granted. The appointments history of Northwestern's English department reflects this change of mind. The department has some members who began their career as editors or did important editorial work sometime along the way. But for forty years the Department has not looked for a scholar whose work revolved around editing. That may have been the right decision at Northwestern and elsewhere, but as a consequence two generations of literary critics in English departments all over the

country have lived off the capital of a century of editorial work while paying little attention to the progressive migration of textual data from books on shelves to files on servers or in 'clouds'.

This digital migration and the integrity of an emerging cyber infrastructure for text-centric scholarship have received remarkably little attention in the discourse of disciplines that will increasingly rely on digital surrogates of their primary sources. A decade ago Jerry McGann observed that "In the next fifty years the entirety of our inherited archive of cultural works will have to be re-edited within a network of digital storage, access, and dissemination. This system, which is already under development, is transnational and transcultural" (Quoted from McGann, 2009). A prophetic remark and especially appropriate in its emphasis on the long-term nature of this challenge. Prophetic also in its reception: getting colleagues in English or History interested in this topic is about as promising as persuading Oklahoma Republicans that climate change is real. There is not much comfort in the conviction that McGann is almost certainly right.

Leaving aside boutique projects — often of exquisite quality —the humanities disciplines have thought of the digital migration not as a challenge to a new form of Lower Criticism but as a clerical task without intellectual interest and therefore to be left to technical staff in libraries, IT departments, or publishers. As McGann put it in 2010:

> the community of scholars has played only a minor role in shaping these events... . Most scholars and virtually all scholarly organizations have stood aside to let others develop an online presence for our cultural heritage: libraries, museums, profit and non-profit commercial vendors. Funding agents like NEH, SSHRC, and Mellon have thrown support to individual scholars and small groups of scholars, ... but while these developments have increased since the public emergence of the Internet [in 1993] – the scholarly community at large remains shockingly passive (McGann, 2010)

Also in 2010 Greg Crane did some back-of-the-envelope calculations about the investment of Classics in the Lower Criticism that supports the documentary infrastructure of that discipline. Remember that it was 50-50 for Mommsen and that for much of the 20th century it was somewhere between 25% and 50%. Based on an analysis of 700 reviews in the Bryn Mawr Classical Review and 100 cv's from a job search he concluded that:

> In effect, classicists as a group have made a cost/benefit decision to allocate less than c. 5% of their labor to the production of editions and commentaries (Crane, 2010)

That sounds about right for departments of English or History that I know something about. It is possible for individuals within fields of activity to make choices that make professional and economic sense within the field but lead the field as a whole astray. The steel industry or the monoculture of corn in Iowa come to mind. Rebalancing the portfolio from 5:95 to 15:85 over the course of a decade and directing the investment towards a digitally inflected Lower Criticism would pay off handsomely even or especially for Higher critics.

3. Early Modern English texts in a digital world

Let me now turn to the particular story of Early Modern texts and their fate in a digital world. In 1883 the young A. W. Pollard joined the British Museum as an assistant in the Department of Printed Books. That was the beginning of a distinguished bibliographical career that resulted four decades later in *the Short-title catalogue of books printed in England, Scotland, & Ireland and of English books printed abroad, 1475-1640,* compiled by Pollard and his collaborator G. R. Redgrave. This fundamental inventory of Early Modern English texts was carried through to 1700 by Donald Wing in the years following World War II. An 18th century short title catalog was from the beginning conceived as a digital project. The *English Short Title Catalogue* of not quite 500,000 items published before 1801 merges these three projects in a single electronic catalogue. It has been a free resource on the Web for at least a decade, but it is useful to recall that it rests on 150 years of bibliographical labour.

The half million books inventoried in this catalogue are the central primary sources for much historical, linguistic, literary, philosophical, and theological scholarship about a 400 year span bounded on one end by the early humanists whose world Stephen Greenblatt has conjured up in *The Swerve* and on the other end by names like Rousseau, James Watt and Adam Smith. As long as universities are around in anything like their current form, this period is likely to remain a central and highly competitive field of scholarly inquiry. At my university, somewhere between 15% and 20% of faculty and graduate students work in those fields. Northwestern is probably typical in this regard.

A very small fraction of highly canonical texts from this period were edited in the 19th and 20th centuries in the Lower Criticism phase of scholarship of which Mommsen's *CIL* is a prime example. Access to most of the data was limited to schol-

ars who lived near or could travel to the few libraries that had substantial collections of such texts. Microfilms, an invention of the thirties, broadened access to Early Modern texts. By the last quarter of the twentieth century a sizable fraction of them was held in the microfilm collections of many research librarians or was available through Interlibrary Loan, an Internet of sorts, slow, but powerful, and in some ways fairer than its digital successor.

The came EEBO, Early English Books Online, digital scans of microfilm images of books published before 1700, delivered over the Web to you on a 24/7 basis, provided you were a member of an institution that could afford the hefty licensing fees. A few years ago, I conducted a little poll asking colleagues what difference digital resources made to their work. I vividly remember a colleague who answered, even before I finished my question: "EEBO has changed everything."

Humanities scholars largely think of digital technology as a convenient new tool for serving up pages for reading. Within that horizon of expectations, EEBO is a perfect end point. It gives you access to everything from anywhere and at any time. Not quite everything, but enough to make that difference meaningless. The images are not always great and sometimes quite bad. The interface is clunky, and your network connection may be spotty. But you can get to the stuff at 2 am in your pyjamas, and compared with the vagaries of hunting for books in the stacks of a real library EEBO may be the better deal much of the time.

A lot of good work has been done and will continue to be done in a mode where EEBO is a simple digital surrogate offering bibliographical data and good enough facsimiles. From a scholar's perspective, this mode of work requires very little adjustment: the computer specific routines are close cousins of ordering books from Amazon and map very readily to work routines in a print environment. And as said before, you can do it anywhere, anytime or in your pyjamas as long as your institution has paid for it.

Since 1999 the Text Creation Partnership (TCP) has been in the business of providing TEI encoded full-text transcriptions of a subset of pre-1700 books. Roughly speaking, the project aims at creating a full-text transcription of every first edition published before 1700, and transcriptions of subsequent editions if there are reasons for doing so. The project has transcribed some 40,000 texts so far and expects to reach its goal of 70,000 texts by 2015. At that point 25,000 texts completed

before 2010 will pass into the public domain, and the remaining texts will follow over a five-year period. [3]

This is a remarkably consequential project. I am not aware of another patch of the distant past that is comparable in terms of size, diversity, importance, and will have an equal density of digital surrogates. Once this resource is in the public domain it will for most scholarly purposes replace other surrogates of the printed originals. It will be free, it will often be the only, and nearly almost the most convenient source for the many look-up activities that make up much of scholarly work. Remember Hamlet:

> Yea from the table of my memory
> I wipe away all trivial fond records
> All saws of boos, all forms, all pressures past
> that youth and observation copied there...

The Google Nexus or Ipad mini will be excellent 'tables of memory' onto which you could within seconds or minutes copy every possible text that might be relevant to your Early Modern project. You can also copy excerpts from these texts and paste them into your paper, which you cannot do with the page image.

Who is in charge of quality assurance for this primary archive that will be the foundation for much future scholarship? EEBO-TCP is a magnificent but flawed enterprise, and few of its transcriptions fully meet the scholarly standards one associates with a decent diplomatic editions in in the print world.[4] Let me put my question in the slightly more aggressive form "how bad is good enough?" In the debate about data curation there is a tension between a philological and a probabilistic ethos. Hillel the Elder said that "whosoever destroys a soul, it is considered as if he destroyed an entire world. And whosoever that saves a life, it is considered as if he saved an entire world." The Hillelesque version of the philological ethos at its extreme says something like "He who fails to correct a single error destroys the entire text." The probabilistic ethos, widely followed in the world of information retrieval and Natural Language Processing, says "the noise level does not matter as long as you get enough signal."

---

[3] For more detail, visit the website at http://www.textcreationpartnership.org.

[4] For more detail see my 2009 blog "Are the TextCreation Partnership texts good enough for research purposes?" (https://scalablereading.northwestern.edu/2012/08/31/are-the-textcreation-partnership-texts-good-enough-for-research-purposes/)

Scholarly readers who have grown up in a print culture have a very low tolerance for the errors that NLP folks take in stride as "noise." Annoyance is not triggered by a failure to understand the intended meaning. Rather, the failure to clean up such messy texts are seen as forms of "dissing" the text and its readers. What kind of cherishing and preserving is it that puts up with gross errors?

Data curation and quality assurance in the digital surrogates of printed primary data take many forms, but you cannot ignore the first and quite humble task of getting the words right. The digital surrogate must do at least as well as the printed source in satisfying the human reader. It must also be machine actionable. Readers and machines have different responses to error. Readers are more on the philological side of the continuum and will feel a "yuck"' response long before the machine fails to extract enough signal.

The concern with getting the words right is an old story. Some 5,000 of the printed originals in the TCP archive have Errata pages. Some of them have simple headings like "Faultes escaped in the printing." But the author of *The Romish Fisher Cavght and held in his owne net,* printed in 1624, offers this elaborate apology:

> I Intreat thee, courteous Reader, to vnderstand, that the greater part of this book was printed in the time of the great Frost; when, by reason that the Thames was shut vp, I could not conueniently procure the proofs to be brought vnto mee, before they were wrought off: whereupon it fell out, that very many grosse escapes passed the Presse, and (which was the worst fault of all) the third part of the book is left vnpaged. This defect I finde no other means to remedy for the present, than to referre thee to the letters of the Printers Alphabet, set vnder the Page. Thus therefore, I pray thee, correct the *Errata* following.

There are two interesting properties of the TCP corpus that affect the discussion of data curation and quality assurance. Both of these have analogues in other large collections of primary materials. The TCP archive may in fact exhibit the characteristic features of the large scale surrogates of printed originals that will increasingly be the first and most widely consulted sources.

First, the TCP is published by a library. Second, in a collection of printed books, the boundaries between one book and another or one page and another impose physical barriers that constrain what you can do within and across books or pages. In a digital environment, these constraints are lifted for many practical purposes.

You can think of and act on the TCP archive as 44,000 discrete files, 2 billion discrete words, or a single file. This easy concatenability is the major reason for the enhanced query potential of a full-text archive. It also has the potential for speeding up data curation within and across individual texts.

Crowdsourcing has been a hot topic for a few years. Often it is discussed as if nothing like it had ever happened before. But I am not the first person to point out that the *Oxford Dictionary* relied heavily on crowdsourcing, with Victorian and Edwardian vicars, schoolmasters, or learned gentlemen of leisure contributing via the Internet of the British Postal service. Apologetic gratitude is a standard feature of prefaces in which authors acknowledge the countless errors that readers pointed out and whose correction makes this second edition a better book.

If you come across a simple error in a book it is usually a matter of seconds to correct it in your mind. It takes much longer to correct it for other readers of the book. You must provide the correction in a review or write to the author/publisher. The publisher must incorporate it into a second edition, and libraries must buy the second editions before the corrected passage is propagated to readers at large.

That is a typical form of data curation in a world where the tripod of cultural memory rests on the actions of scholars, publishers, and librarians. In a digital world that tripod rests on the interactions of scholars, librarians, and technologists. In a well-designed digital environment scholars (and indeed lay people of all stripes) can directly and immediately communicate with the library/publisher. If I work with a text and come across a phenomenon requiring correction or completion I can right away do the following:

1. log in (if I'm not logged in already) and identify myself as a user with specified privileges
2. select the relevant word or passage and enter the proposed correction in the appropriate form.

If I do not have editorial privileges, my proposal is held for editorial review. If I am authorized to make or approve corrections my proposal is forwarded for inclusion in the text either immediately or (the more likely scenario) the next time the system is re-indexed. The system automatically logs the details of this transaction in terms of who did what and when.

Nothing in this scenario is clearly beyond the scope of current technology. Think of it as a digital carrel in which I can work both for myself and for others. Work in this carrel should come as close as possible to "reading with a pencil," a workflow in

which the highlighting of a passage, the correction of error, or marginal comment are the work of the same hand, using the same tool, and seamlessly moving from one task to another. From a technical perspective, all these acts are "annotations," some for my own use, some for others.

The obstacles to such an environment are not primarily technical or financial. They are largely social. You need substantial adjustments in the ways scholars and librarians think about their roles and relationships. Scholars often complain about the shoddiness of digital resources, but if they want better data they must recognize that they are the ones who must provide them—although they may find it rewarding in many ways to recruit a lay public for help with those tasks. And they need to ask themselves why in the prestige economy of their disciplines they have come to undervalue the complexity and importance of "keeping" (in the widest sense of the word) the data on which their work ultimately depends. Librarians need to rethink the value chain in which the Library ends up as a repository of static data. Instead they should put the Library at the start of a value chain whose major component is a framework in support of data curation as a continuing activity by many hands in many places, whether on an occasional or sustained basis. Such a model of collaborative data curation is the norm in genomic research, a discipline that from the perspective of an English department can be seen as a form of criticism (both higher and lower) of texts written in a four-letter alphabet.

Early in this talk I used the corpus of Latin inscriptions as a model of scholarly data curation in a world of print. Let me at the end of this talk turn to Greek papyrologists as a model of a scholarly community that has thought through the problems of collaborative data curation in particularly imaginative ways. When I first read about Integrating Digital Papyrology (IDP) I was struck by the phrase "investing greater data control in the user community." Roger Bagnall, the director of the Institute for the Study of the Ancient World at NYU has talked eloquently about this model. In speaking about the technological changes brought about the Web he said that

> these changes have affected the vision and goals of IDP in two principal ways. One is toward openness; the other is toward dynamism. These are linked. We no longer see IDP as representing at any given moment a synthesis of fixed data sources directed by a central management; rather, we see it as a constantly changing set of fully open data sources governed by the scholarly community and maintained by all active scholars who care to participate (Bagnall, 2010)

What about quality assurance in such world? Bagnall faces this problem head on in his discussion of

> the *Berichtigungsliste*, a remarkable research tool in papyrology that collects periodically—there have been twelve volumes since its inception in 1915—all corrections proposed to the texts of papyrus documents ..... Before corrections are registered now, the editors of the BL do their best to check them to see if they think they are correct; if not, they are reported but with disapproval attached. How, my friend asked, will we prevent people from just putting in fanciful or idiotic proposals, thus lowering the quality of this work? (Bagnall, 2010)

Bagnall answers that collaborative systems

> are not weaker on quality control, but stronger, inasmuch as they leverage both traditional peer review and newer community-based 'crowdsourcing' models. The worries, though, are the same ones that we have heard about many other Internet resources (and, if you think about it, print resources too). There's a lot of garbage out there. There is indeed, and I am very much in favor of having quality-control measures built into web resources of the kind I am describing. (Bagnall, 2010)

A collaboratively curated *Berichtigungsliste* or curation log offers an attractive model for coping with the many imperfections of the current TCP texts. In this talk I only have time for the simplest forms of data curation, the fixing of millions of incompletely or incorrectly transcribed words. You want to fix the simple things before you tackle the hard ones. Because there are so many of them the fixing of simple things by many users poses tougher technical challenges than the fixing of hard things. And if most of the simple things were fixed, the TCP archive would be in much better shape.

The full text of TCP transcriptions is not generated from OCR but through double keyboarding. In principle this method guarantees a very high degree of accuracy: a transcription of Shakespeare that meets its quality standards would give you on average one error per play. Even the fussiest philologist can live with that. In practice things are a little different. The transcribers were asked not to transcribe stuff in foreign alphabets or in musical or mathematical notation. They did not correct the errata pointed out by printers, and they did nothing with the many errors the printers did not catch. They worked from the digital scans of microfilm images of varying and often quite bad quality. There are millions of cases where they dutifully report missing letters, words, or larger stretches as unreadable. These gaps are very une-

venly distributed. They are an excellent measure of overall quality. A page without gaps was almost certainly transcribed from a readable image and is likely to meet quality standards. A page with many gaps is likely to have other errors as well.

The correction of a simple error is in nearly all cases an atomic event that is independent of the words that precede or follow, although in many cases the preceding and following words will provide the evidence for correction even without inspection of the source text. If you "tokenize" the text and associate every word token with a unique ID, the resulting model of the text gives you a framework for a *Berichtigungsliste* or curation log for the TCP: a very long table, where each data row includes a token ID and the who, when, and what of the proposed correction. From the tokenized text you can derive word frequency lists that in the manner of a spell checker propose solutions. In some cases such algorithmic solutions do not even need human review. In the normal case, such algorithmic suggestions do require review, but for the human curator approving an algorithmic suggestion or choosing between proposed alternative requires less work than coming up with the solution in the first place.

I have sketched a scenario in which the work of many hands, supported by clever programmers, quite ordinary machines, and libraries acting consortially, would over the course of a decade substantially improve the TCP texts and move them closer to the quality standards one associates with good diplomatic editions in a print world.[5] Imagine 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. You could also imagine a system of certification for each text — not unlike the USDA hierarchy of grades of meat from prime to utility. The initial quality ranking would be based on the number of gaps, with changes based on the percentage of corrected errors. But "prime" would always be reserved for texts that have undergone high-quality human copy-editing. Such a system would build trust and would counteract the human tendency to judge apples by the quality of those at the bottom of the barrel.

I recently went to the EEBO TCP conference at Oxford ("Revolutionizing Early Modern Studies?").[6] There was much talk of projects enabled by the TCP, but little

---

[5] Compelling arguments in favour of this model have been made by Rose Holley in "Many hands make light work" (Holley 2009) and "Crowdsourcing: How and why libraries should do it?" (Holley 2010)

[6] For a fuller account see my blog entry " EEBO-TCP 2012: The future of the TCP as a public domain and collaboratively curated corpus of Early Modern English" (https://scalablereading.northwestern.edu/2012/09/26/eebo-tcp-2012-the-

talk of what the projects could give back to the TCP. A critical edition of Hakluyt's *Principal Navigations* will use the TCP transcription as a point of departure but collate them with printed copies at the Huntington Library. The project will be published by the Oxford University Press. The many textual corrections will probably not flow back to the TCP, as perhaps they should. At the University of Reading Mark Hutchings has offered seminars in which students edit short TCP texts. The projects end up as printed pages in his study. What if the editorial tasks were defined in such a way that the results of the students' labour could be directly integrated into the TCP texts, whether as textual improvements, metadata in the form of structural markup, or annotations of some form? Would students, in addition to learning something, take some satisfaction in the fact that they have added a little to a larger enterprise and that the result of their work may be helpful to others? Katherine Rowe at Bryn Mawr has asked this question and answered it with an emphatic "yes."

So have Greg Crane at Tufts and Helma Dik at Chicago. Their second-year students of Greek parse sentences from Aeschylus, Herodotus, etc, and these sentences contribute to the growing Perseus treebank of Greek and Latin texts. This is an easy case: undergraduates have always parsed Greek and Latin sentences and will continue to do so as long as those languages are taught. Feeding the output of the students' work into the Perseus treebank does not change what they are doing anyhow, but gives them the additional satisfaction of doing something useful (by the admittedly rarefied criteria of Greek scholars who love treebanks).

It is not quite so easy to reconcile the routines of undergraduate work in other disciplines with work on the TCP texts that needs doing. In the curation of TCP texts, the needed work consists first of all and for quite a while of fixing millions of incompletely or incorrectly transcribed words through some combination of algorithms and 'humint' work. There is a lot of menial work here, and the question arises whether such work teaches students anything they need to learn or whether this is a wasteful and exploitative use of their time. In "The Politics and Poetics of Transcription," one of the subtlest papers at the Oxford EEBO conference, Giles Bergel was skeptical about reducing transcription to mere "data capture" and argued that "transcription, far from being a mechanical or mundane activity, can be a demanding intellectual discipline."

Undergraduates who spend some hours or days checking a transcription against the digital page from which it is derived will probably conclude that they do not

---

future-of-the-tcp-as-a-public-domain-and-collaboratively-curated-corpus-of-early-modern-english/)

want to do this for the rest of their lives. But they will learn that it is quite difficult to copy a passage with unfamiliar orthographic or typographic habits—a point well made by Giles Bergel. They may also gain some respect and appreciation for the enormous and mostly invisible labour that makes their easy access to digital archives possible. Above all, they will learn that even for quite famous texts the ground of textual truth is rarely bedrock and quite often very thin ice. That is a very useful lesson to learn, and the reflections of bright undergraduates on this lesson are a joy to read.[7]

The grandparents of today's undergraduates include many highly educated retirees with well-defined intellectual interests, time on their hands, and the desire to do something useful. The abstract name TCP will mean little to them. But if you divide that corpus into recognizable neighbourhoods of travel narratives, military history, theological disputes, alchemy, cookbooks, childrearing, or witchcraft you may be able to recruit amateur scholars of extraordinary competence and persistence. Clay Shirky has written a fascinating book about "cognitive surplus." There is a lot of it out there, and if you do a good job of "directing the crowd" by matching things that need doing with the skills of people able and willing to do them, many scholarly tasks can be done by folks of all ages and interests, giving them the satisfaction of contributing to something worthwhile and having their work recognized.

What I have said about collaborative curation of the TCP texts applies with minor changes to other archives. Neil Fraistat and Doug Reside in conversation coined the acronym CRIP for "curated repository of important texts ." Not everything needs to be curated in same fashion, but high degrees of curation are appropriate for texts that we seek to "cherish and preserve." Large consortial enterprises like the Hathi Trust or the DPL might be the proper institutional homes for special collections of this. Somewhere in the middle distance I see the TCP collection as the foundation of a Book of English defined as

- a large, growing, collaboratively curated and public domain corpus
- of written English since its earliest modern form
- with full bibliographical detail
- and light but consistent structural and linguistic encoding

It will take a while to get there.

---

[7]" 'Fluent in Marlowe": Emily's and Sasha's successful adventures in data curation" (https://scalablereading.northwestern.edu/2012/06/05/fluent-in-marlowe-emilys-and-sashas-successful-adventures-in-data-curation-2/)

References

Bagnall, Roger. 2010. Integrating Digital Papyrology. In *Online Humanities Scbholarship: The Shape of Things to Come*. Charlottesville, Virginia. http://shapeofthings.org/RBagnall.doc

Crane, Gregory. 2010. Give us editors! Re-inventing the edition and re-thinking the humanities. In *Online Humanities Scholarship: The Shape of Things to Come*. Charlottesville, Virginia. http://shapeofthings.org/GCrane.doc

D'Iorio, Paolo. 2010. Scholarly information management: a proposal. In *Online Humanities Scnolarship: The Shape of Things to Come*. Charlottesville, Virginia. http://shapeofthings.org/papers/PDorio2.doc

Holley, Rose. 2009. Many hands make light work: public collaborative OCR text correction in Australian historic newspapers. edited by National Library of Australia. http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf

Holley, Rose. 2010. "Crowdsourcing: How and why libraries should do it?" *D-Lib Magazine* no. 16 (3/4). http://www.dlib.org/dlib/march10/holley/03holley.html

Kaiserlian, Penelope. 2010. Rotunda. In *Online Humanities Scholarship: The Shape of Things to Come*. Charlottesville, Virginia. http://shapeofthings.org/papers/PKaiserlian.doc

McGann, Jerome. 2009. "Our textual history." *TLS*. November 20, 2009

McGann, Jerome. 2010. Sustainability: the elephant in the room. In *Online Humanities Scholarship: The Shape of Things to Come*. Charlottesville, Virginia. http://shapofthings.org/papers/McGann.docx

Prescott, Andrew. 2012. Making the digital human: Anxieties, possibilities, challenges. In *Digital Riffs.* http://digitalriffs.blogspot.com/2012/07/making-digital-human-anxieties.html

Shirky, Clay. 2010. *Cognitive surplus: How technology makes consumers into collaborators*. London.