The Business of Digital Humanities: Capitalism and Enlightenment

Laura Mandell & Elizabeth Grumbach
Texas A&M University

Scholarly and Research Communication
VOLUME 6 / ISSUE 4 / 2015

Background: The Advanced Research Consortium (ARC)
The Advanced Research Consortium (ARC) began in 2005 with the launch of the Networked Infrastructure for Nineteenth-century Electronic Scholarship (NINES), the brainchild of Jerome McGann and Bethany Nowviskie. Organized around literary and historical periods, ARC is comprised of the directors of online scholarly communities that peer review digital projects and aggregate metadata for peer reviewed and other collections into an online search portal. The five ARC search portals are NINES, 18thConnect, MESA or medieval, ModNets or modernism, and ReKN or Renaissance (the latter two are forthcoming). In this article, we will discuss partnerships that ARC has established with proprietary data companies and the possible benefits for scholars and libraries from the possibility of collaborating with companies – vendors that serve data to libraries. More important, I will argue that there is a terrible threat hanging over disciplines such as literary studies and that we need to become avid archive entrepreneurs.

Keywords
Advanced Research Consortium; Proprietary data companies; Literary studies

Laura Mandell is Director of the Initiative for Digital Humanities, Media, and Culture (IDHMC) at Texas A&M University; Director of the Advanced Research Consortium (ARC) and 18th Connect; and Project Director of The Poetess Archive. Email: mandell@tamu.edu .

Elizabeth Grumbach is a staff research associate at the Initiative for Digital Humanities, Media, and Culture (IDHMC) at Texas A&M University; the Project Manager of the Advanced Research Consortium (ARC); and Project Manager for 18thConnect. Email: egrumbac@tamu.edu .

CISP Press
Scholarly and Research Communication
Volume 6, Issue 4, Article ID 0401226, 9 pages
Journal URL: www.src-online.ca
Received June 8, 2015, Accepted July 27, 2015, Published October 29, 2015

Mandell, Laura, & Grumbach, Elizabeth. (2015). The Business of Digital Humanities: Capitalism and Enlightenment. *Scholarly and Research Communication, 6*(4): 0401226, 9 pp.

Personal narrative by Laura Mandell

I went to the first or second THATCamp held at George Mason University. I was attending a panel with Daniel Chudnov, librarian at the Library of Congress and author of the blog *One Big Library*. I went into a rant, the substance of which is the following: The British Library had at some point in the mid-to-late twentieth century decided to microfilm the pages of all the books in its eighteenth-century collection. It hired a vendor (I am not sure which one) to do the microfilming, and part of the agreement was that the vendor would be able to resell the microfilm to other libraries – that is how the British Library reduced the cost of microfilming its own collection, by allowing that stipulation to be part of the contract and part of the vendor's cost recovery and profit. A company called "Research Publications" of Woodbridge, Connecticut, sold the microfilm collection to libraries. In the 1990s, that company was acquired by the Gale Group, a division of Thomson Learning which eventually became Gale Cengage Learning when Thomson divested itself of Thomson Learning.[1] Gale Cengage Learning then digitized those page images, associated the texts with metadata from the British Library's English Short Title Catalogue, and sold the package to libraries as the Eighteenth-Century Collections Online, or ECCO. Gale was now selling ECCO at such a high cost that a company closely tied to libraries in the U.K. called JISC Historic Collections bought the ECCO catalogue, repackaged it, and sold it to all U.K. universities at a much reduced rate so they could afford to buy it. My rant ended with something like, "the British people are being forced to buy back their own cultural heritage at an exorbitant cost." I will never forget Dan's response: "Welcome to my world."

Hello, world

The truth is, the British Library could not have microfilmed or digitized its collections without vendor agreements. The cost of microfilming, digitizing, cameras, scanners, servers, programmers, associating metadata with files, and OCR'ing texts is very high. Any company that performed such work would have to charge more than the British Library could afford to pay. Vendors could charge the British Library a reasonable amount for their work precisely because they were given the rights they needed in order to sell the microfilmed and digitized collections. Via those sales, they could recover the costs of microfilming or digitizing and make a profit.[2] Yes, you might say, but the total cost to universities and scholars, not just to the British Library, has increased exponentially. Exponentially? Well, there is a limit to how much a company can profit by selling collections of early modern data – there is a market, in other words, and it is neither infinite nor infinitely wealthy. But yes, money could have been saved if the British Library had not hired a vendor and had instead done the work in house – then it would have been able to sell its product to universities worldwide at a much lower cost, not needing to profit but only to recoup costs. Added to the costs, however, would be sustainability and service costs – in other words, not just the initial outlay, but all the marketing, information technology (IT) services, and backups needed to run an operation such as the one that served up ECCO, an operation just like Gale Cengage Learning. Did or does the British Library want to be in such a business, and should it? No – hence the vendor.

Right now, ARC's signature project, the Mellon Foundation-funded Early Modern OCR project or eMOP, is improving the mechanically typed text in ECCO – the textual data

Mandell, Laura, & Grumbach, Elizabeth. (2015). The Business of Digital Humanities: Capitalism and Enlightenment. *Scholarly and Research Communication,* 6(4): 0401226, 9 pp.

generated out of those microfilmed-digitized-low-quality page images by Optical Character Recognition (OCR) engines – and sending it back to Gale Cengage Learning so they can improve their product. "What?!," I hear you exclaim, "you are helping them profit!" Actually, truth be told, Gale will profit no more or less with cleaner text for searching the ECCO collection. The only people who "profit" from cleaner text and better searching capacity are the users of ECCO – *us*, scholars. And if you ask me if Texas A&M can keep clean text files of ECCO documents on its servers forever and serve them to the public forever, I would say no. And if you then ask me, "What is your sustainability plan?" for the grant deliverables you are producing for the eMOP project, I would answer, "Gale. Gale is my sustainability plan." These clean versions of the texts will be carefully preserved along with page images in the ECCO catalogue as long as Gale, or whatever company they transform into or whatever company purchases ECCO from them – as long as Company X keeps profiting from servicing and selling the ECCO catalogue – which is to say as long as we scholars, professors, and students keep wanting to use that ECCO catalogue. Would a library keep it longer than that? At the Mellon-funded UVA conference called "The Shape of Things to Come" organized by Jerome McGann to address "the elephant in the room" of long-term sustainability of digital projects, Paul N. Courant, then Dean of the University of Michigan Libraries, turned to the attending faculty and said, "Use it or lose it." I remember thinking to myself, "I'm glad no one said that about the Gutenberg Bibles." But now I know Paul Courant is right. When I worked at the old British Library, in the British Museum, I daily walked past one of those Bibles on my way to the restroom, and it was in a glass case: in digital terms, a dark archive. What we need to do is not take digital catalogues, such as Gale's ECCO, from the companies that developed them and make libraries perform all the work of their vendors, in addition to the work they are already doing. We do need to have in our contracts with such vendors that, when the company or its successors no longer wishes to sell and sustain a digital catalogue, it must pass it to libraries for dark archiving.

It's my archive and I'll cry if I want to

The eMOP project is producing better text for better searching in ECCO. Gale itself would not clean up the OCR running behind their page's images, which allows for full-text searching of the ECCO catalogue: they would not because it would not profit them to do so. Higher fees for the catalogue could not be charged, and no library would decide not to buy the catalog based upon the fact that every word cannot be searched because the typed text used for searching – hidden from the user's view – is mistyped. To generalize just slightly, the capitalist profit motive fails to produce the digital archive that scholars want and need. I am currently participating in a partnership called "Text Mining the Novel" or NovelTM, the members of which are absolutely convinced that the quality of the OCR'd, or mechanically typed text, and the machine-readable and therefore algorithmically manipulable data, does not much matter. They have a few automated cleanup routines they run, and this, in their view, is adequate for the kind of work they want to do.

It is not.

In "From Babel to Knowledge: Data Mining Large Digital Collections," Daniel Cohen (2006) expresses the standard view, the view of my cohort in NovelTM, by saying that, in

terms of the validity of data-mining results, "*Quantity may make up for a lack of quality*" (n.p., emphasis in the original). Grateful as I am for the conditional "may," I strongly disagree. In contrast to what he says here, I have learned another lesson from working with clean and messy data in conjunction with each other. That cutting-edge research requires clean data (a hygienic metaphor) was brought home to me in a recent work I have been doing determining the eighteenth-century meaning of the saying "circumstantial information." I wanted to know whether it is semantically equivalent to our locution "circumstantial evidence." In searching for the phrase in the ECCO catalogue, as currently available, via my library, I got 24 results. These results come from the 136,000 documents in ECCO Phase I, which searches through error-ridden machine-typed texts, or "dirty OCR." I then searched through a subset of those documents, 2,169 texts, which have been hand typed by the Text Creation Partnership (TCP). These hand-typed texts, not perfect themselves, are searchable by word in ARC's 18thConnect portal: one need only go to the search page and select "ECCO" and "free-culture text" as facets in order to search through the handcrafted textual data. Because only a small subset of texts has been typed by the TCP (they ran out of money to type the rest), I got a much smaller set of returns, obviously. I got four results. However, two of those returns are not on the list of 24 generated by the ECCO catalogue. What does that mean?

It means that in searching Gale's instantiation of the ECCO catalogue, at least two results that would have been returned had the text Gale is running been corrected, were missed. But think about it: we have only typed roughly one percent of the ECCO catalog – how many texts would have been returned if the whole catalogue had been as carefully typed? Presuming the other uncorrected OCR documents are as rich in returns as this one percent, the 2,700 texts typed by the TCP**,** I would get 226 more returns. That would mean 250 returns total, as opposed to 24, from just searching Phase I ECCO – that is, if and only if the data in ECCO was correct typescript. The number of search returns from clean text is more than ten times the results from dirty OCR: even presuming it is not as information rich, and it could be more so, the number of search returns is far, far greater. Most important, there is no guarantee the 24 out of the 250 results from Phase I ECCO are at all representative. In fact, I would argue they are not. The 24 returns suggest the saying in question is used in a pejorative sense in legal discourse, whereas the enhanced returns show at least one positive instance – "circumstantial information" is invaluable for literary biography, it turns out, and so when discussing literary lives is not meant as "merely" circumstantial in the legal sense but as providing accurate witness to the circumstances of a writer's life.

In terms of creating digital archives, capitalism is not working for scholars: we need to clean up the ECCO archive, need to invest labor in it beyond what the market allows. We have to partner with – yes, in some sense subsidize – these library vendors to get the digital archive that we want.

A new deal

During a Modern Language Association (MLA) convention many years back, a group of English professors got in a hotel elevator that stopped working for about 40 minutes. Inside the elevator, about 20 minutes into the ordeal, the spouse of one faculty member said to the whole group, "You all have theories about this, but nobody's doing

Mandell, Laura, & Grumbach, Elizabeth. (2015). The Business of Digital Humanities: Capitalism and Enlightenment. *Scholarly and Research Communication,* 6(4): 0401226, 9 pp.

anything."[3] To me, this emblematic tale encapsulates the point of Robert Levine's (1993, 2012) important essay in *Profession*, "The Real Trouble": we all have theories about capitalism's depredations of our intellectual work, but none of us are doing anything about it. But why should *we* make up for the damage done by profiteering? Insofar as the welfare state supports capitalism by making up for its deficiencies, is any deal with for-profit companies not complicit with neoliberalism?[4] From the perspective of those of us on the frontline of worrying about future research in the humanities disciplines, this is the wrong question: the choices are not complicity or opposition, but archive or no archive, the findable or the lost. In fact, because of the manner in which much of our cultural heritage is being digitized, a report sponsored by the European Commission on Information Society and Media worries that we could experience "a digital Dark Age" (Niggemann, De Decker, & Lévy , 2011, p. 7).

We could have all our cultural heritage dark archived and, even if it is brought to the fore by being released on the Internet, for example, if the heritage data is unreadable by machines, if it is unsearchable, we will not be able to figure out what is in it. Future scholars will think that people in the eighteenth century viewed evidence in exactly the same way as we do, and will even insert difference when there is none. I can hear the Bill Moyers of 3015, having pulled into the light and searched the dark-archived ECCO catalogue, saying to his viewers, "The eighteenth century did not even have a word for curiosity. They did have something similar, though, which they called, 'curiofity.'" When Google first launched Google books, a search of an eighteenth-century copy of *Clarissa* revealed that the word appeared in the text three times – in all 1308 pages, author Samuel Richardson only used the word three times – and, at that early time, when I visited one of the pages Google highlighted as containing the word, I found three instances of it on that page alone, the other instances not recognized as the word "curiosity," not highlighted.[5]

Keeping curiosity alive

18thConnect and eMOP are partnering with Gale and ProQuest to improve the mechanically typed texts that people search when using the Early English Books Online (EEBO) and ECCO catalogues.[6] In fact, EEBO does not contain any dirty OCR because none of those page images have as yet been run through the Optical Character Recognition process; the results have been too awful. Instead, the Text Creation Partnership has hand typed about 45,000 of those documents, and they alone are the full texts one is searching when searching EEBO. There are 123,000 items in the catalogue, but when a scholar searches it, he or she is searching some full text and predominantly metadata – only titles, authors, publication dates, and the like. The eMOP project has taught us how difficult it is, indeed, to make machine-readable texts, but we are working on it for EEBO and trying to improve the OCR for ECCO, which really is state of the art. That is, we are trying to push that state to a higher level and improve ECCO textual data.

The contracts we got with Gale and ProQuest are very good indeed. We are creating and improving their OCR; they will insert the corrected data into their catalogues as soon as we have it. Is that all? Well, all the texts are searchable in 18thConnect, so even when a scholar's library does not subscribe to EEBO or ECCO, that scholar can

perform full-text searches on both those catalogues, and metadata records from the English Short Title Catalogue are returned along with the EEBO or ECCO text results. These records also state which holding libraries have those texts. So scholars can get searching access to proprietary catalogues: is that all? No. ProQuest's contract with us is identical to Gale's.[7]

The contract has another provision in it, as does our contract with Gale, besides just accepting corrections from us. We are releasing the typed text to users via a tool called TypeWright, which can be found in 18thConnect (see Appendix). In TypeWright, users hand correct the mechanically typed texts we have generated for EEBO and ECCO.[8] Gale and ProQuest are allowing us to give the corrected document to whomever has corrected it, whether one person or a group, for their labours. Getting the digital text they have corrected allows the user to publish a digital edition of it online and make it freely available. 18thConnect is able to make it relatively easy for scholars to create digital scholarly editions,[9] thus fostering the creation of a digital environment as loaded with scholarly editions of texts as our current print environment. Those scholarly editions, if submitted back to 18thConnect for peer review, can become part of the Semantic Web – we use Resource Description Framework (W3C, 2014) metadata to ingest digital objects into 18thConnect, banking on a Semantic Web future (Berners-Lee, Hendler, & Lassila, 2001; Bizer, Heath, & Berners-Lee, 2009).

Going to the bank

In working on creating a searchable archive of cultural heritage materials, the ARC group has discovered that we need to partner with businesses in order to bank the archive we want. Bankrolling it is another matter; in my view, the Mellon Foundation is almost single-handedly shaping digital archives for the future, and we are grateful for this endowment from a businessman of the past. As much as we do not like it, we have to partner with businesses in order to perform high-quality intellectual work. Research is a luxury, one we cannot afford to lose.

Notes

1. Woodbridge Research Publications changed its name to Primary Source Media when it was ingested by Gale (http://www.cengage.com/search/showresults.do?N=197+4294904141). For Gale's relationship to Thomson, see https://en.wikipedia.org/wiki/Gale_(publisher) .

2. It is important to know that Gale did not make enough money to cover costs on some of the collections it digitized for the British Library and that profits from ECCO have in effect sponsored some other archival preservation endeavors (private communication from Scott Dawson, formerly of Gale Cengage Learning).

3. I would like to thank David McWhirter for this story.

4. To me, avoiding complicity at all costs, even if one of those costs is the livelihoods of our graduate students, is the advice given by Richard Grusin (2014).

Mandell, Laura, & Grumbach, Elizabeth. (2015). The Business of Digital Humanities: Capitalism and Enlightenment. *Scholarly and Research Communication,* 6(4): 0401226, 9 pp.

5. Google's OCR engine, one we are working with at eMOP, has since gotten much better and needs now to be run on EEBO and ECCO texts, which we are doing thanks to funding by the Mellon Foundation (http://emop.tamu.edu). Since that OCR engine's improvement, one now gets 28 returns—the long-s is more often read as it should be, as an "s." After many many critiques of dirty OCR running behind its texts, Google has removed much of the mechanically-typed, error-ridden text from view; they now substitute snippets of page images for portions of the plain text when the OCR is too "dirty."

6. http://emop.tamu.edu . We are composing the final report concerning the results of this project at the time of copyediting this article; that report will be published on the website during Fall 2015.

7. The ProQuest contract is available online, http://idhmc.tamu.edu/projects/Mellon /ProQuestContract.pdf , as is the contract with Gale Cengage Learning: http:// idhmc.tamu.edu/projects/Mellon/eMOPAppendixPublic.pdf , pp. 22-28.

8. At the writing of this article, only ECCO texts are available. The EEBO texts will be available in TypeWright by November 2015.

9. Right now, 18thConnect sends the user his or her completely corrected document in Text Encoding Initiative (TEI), the encoding format necessary for creating a digital edition. We hope to shortly have an edition builder installed in 18thConnect to make it even easier to create such an edition. Our illustrious editorial boards (18thConnect, n.d.) will peer review any edition, should students and faculty wish to have their editions findable in 18thConnect and to have line items for their curriculum vitae.

Websites

18thConnect, http://www.18thconnect.org/

Advanced Research Consortium (ARC), http://idhmc.tamu.edu/arcgrant/

Early Modern OCR Project (eMop), http://emop.tamu.edu/

Gale Cengage Learning, http://www.cengage.com/search/showresults.do?N=197

MESA, http://www.mesa-medieval.org/

Networked Infrastructure for Nineteenth-century Electronic Scholarship (NINES), http://www.nines.org/

NovelTM, http://novel-tm.ca/

One Big Library, http://onebiglibrary.net/

ProQuest, http://www.proquest.com/

Text Creation Partnership, www.textcreationpartnership.org

References

18thConnect. (n.d.). *Editorial boards.* URL: http://www.18thconnect.org/about/scholarship/editorial-boards/ [July 26, 2015].

Berners-Lee, Tim, Hendler, James, & Lassila, Ora. (2001). The Semantic Web. *Scientific American*, *284*(5), 34-44. URL: http://www.scientificamerican.com/article/the-semantic-web/ [July 26, 2015].

Bizer, Christian, Heath, Thomas, & Berners-Lee, Tim. (2009). The story so far. *International Journal on Semantic Web and Information Systems (IJSWIS), 5*(3). URL: http://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf [January 7, 2015].

Cohen, Daniel J. (2006). From Babel to knowledge: Data mining large digital collections. *D-Lib Magazine*, *12*(3) [n.p.] URL: http://www.dlib.org/dlib/marcho6/cohen/03cohen.html [July 26, 2015].

Grusin, Richard. (2014). The dark side of digital humanities: Dispatches from two recent MLA conventions. *Differences*, *25*(1), 79-92.

Levine, George. (1993). The real trouble. *Profession, 93,* 43-45.

Levine, George. (2012). The real trouble. *Profession,* 143-148.

Niggemann, Elisabeth, De Decker, Jacques, & Lévy, Maurice. (2011). The new Renaissance: Report from the Comité des Sages Reflection Group on bringing Europe's cultural heritage online. URL: http://www.eurosfaire.prd.fr/7pc/doc/1302102400_kk7911109enc_002.pdf [January 7, 2015].

Richardson, Samuel. (1784). *Clarissa: Or, the History of a Young Lady. Comprehending the Most Important Concerns of Private Life. ... By Mr. Samuel Richardson.* In Eight Volumes. London: Harrison. 1308 pp. URL: https://books.google.com/books?id=FdENAAAAQAAJ&printsec=frontcover&dq=clarissa&hl=en&sa=X&ved=0CCMQ6AEwAW0VChMI196CnsqeyAIVyXE-Ch2x9gtS#v=onepage&q=clarissa&f=false

W3C. (2014). *Resource Description Framework*. URL: http://www.w3.org/RDF/ [July 26, 2015].

Appendix
Instructions for Using TypeWright

To use TypeWright, go to 18thConnect, you can either:

- Click on the TypeWright tab, top right-hand corner, and then edit the featured text or search for a text using the search bar there. Make sure you click on "Start Editing" for the featured text or "Edit" for any other text that you find.
- Or, click on the Search tab, and select the facets on the right-hand side of the screen:
  "Other Digital Collections" and "ECCO." When those search returns come up, 182,000 of them, there will be "Edit" buttons for each item (under "Collect" and "Discuss"). Click on Edit.

In either case, you will be prompted to create an account. You need to enter a real email address, but your username and password can be anything you choose. Then you will be at the homepage for the text: again, be certain to click on "Start Editing" to see TypeWright in all its glory.

The newest feature of TypeWright is that multiple users can be editing at the same time, and they will see each other's changes as they do so.