Teaching with TypeWright:

Scaffolding Undergraduate Learning About Digitization[1]

From subscription resources like *Early English Books Online* (EEBO) and *Eighteenth Century Collections Online* (ECCO) to free resources like Wikipedia and Google Books, undergraduate students are increasingly using digital databases for their academic work. Part of our role as educators is to lead students to appropriate resources. One challenge I have faced, and one that many of us teaching literature courses do not address often enough, is getting students to consider both the practical and theoretical implications of the digital resources they use.

Susan Schreibman has argued that in a digital environment "reliability and permanence means that the digital object has a certain relation to the original and that in its transformation the trusted digital repository ensures that in its refashioning, the digital object re-presents the analogue artifact" (8). With Schreibman's comments in mind, it is questionable whether or not resources like EEBO and ECCO are, in fact, "reliable." Many respected scholarly digital archives and repositories—resources that libraries and institutions often pay significant amounts of money to access—are, as Nicholson Baker, Kevin Berland, and Patrick Spedding have noted, riddled with optical character recognition (OCR) transcription errors. As I explain later in this essay, the way search results are displayed in relation to these page images masks the underlying, error-prone layers of mediation on which many digital resources depend.

This, of course, is not to say that we should throw digital archives out the window and only ask students to use physical materials that are held at our brick and mortar libraries. Digital resources are undeniably reshaping how we define our own research. By extension, they have the

[1] [Note from the PAJ Editor: this article was accepted by PAJ in 2012, its publication delayed by the movement of PAJ from Miami University of Ohio to Texas A&M University. The article has since been minimally updated to reflect more recent manifestations of the assignment the article describes.]

potential to redefine original undergraduate research, a field that, until recently, was synonymous with the "chemical glassware" of science labs (Blackwell and Martin 2). Integrating digital pedagogy and digital resources into humanities courses "allow[s] students to engage with primary sources much earlier, and in greater depth, than was previously possible" (Cayley 211). Just as lab work introduces undergraduate students in the sciences to increasingly advanced research techniques, digital resources offer ways for undergraduates in fields such as English, History, and Classics to conduct original, experiential, and exploratory research on primary documents. However, the key is not just getting students in a computer lab or using digital sources uncritically; it is crucial to surface the problems of mediation and access at work in digital archives in a way that promotes information literacy and innovative problem solving.

My own teaching at the intersections of literary studies, book history, and digital humanities necessitates introducing students to useful digital resources and digital tools for literary analysis while also making clear their limits. This is not always an easy task, as students are liable to shut down during conversations about reliable digital resources, since many of them have had dull lectures about the "danger" of using resources like Wikipedia.[2] My experiences using TypeWright in the undergraduate classroom demonstrate that the tool facilitates student-led discussions about digitization, reliability, and access. One of the best things about TypeWright is its usability; TypeWright can be integrated successfully into undergraduate courses even when technological teaching support is minimal and students' knowledge of both eighteenth-century literature and digital research tools is limited. Moreover, teaching with

[2] These issues first became pressing in 2011 when I was teaching tutorials for a new undergraduate English course at the University of Toronto, "The Digital Text," taught by Dr. Adam Hammond. I first experimented teaching with TypeWright for the independent tutorials that I led in association with this course. Since 2011 I have used TypeWright in several undergraduates classes, from an introductory literature survey course to, most recently in 2014, a senior seminar entitled "The History and Future of the Book" at Georgia State University. For more information about Adam Hammond's course in 2011 and its final collaborative digital project, see Hammond and Brooke, *He Do the Police in Different Voices: A Website for Exploring Voices in T.S. Eliot's The Waste Land.*

TypeWright allows opportunities to design activities that meet the four stages of Kolb's cycle of experiential learning: concrete experimentation, reflective observation, abstract conceptualization, and active experimentation, , the last of which can be difficult to achieve in traditional humanities courses (Svinicki and Dixon 144).

I'll share some of my classroom experiences below, but I first want to introduce TypeWright. Publicly released in late 2011, TypeWright is a collaborative web-based tool that allows users to correct the optical character recognition (OCR) transcription errors prevalent in text versions of works in the *English Short Title Catalogue* and ECCO. TypeWright is one of the tools available through 18thConnect's Collex interface.[3] The tool follows in the steps of the Australian Newspaper Digitisation Program, which created an interface to allow the public to correct the low-quality electronic transcriptions of Australian newspapers published from 1803-1954 (see Holley). TypeWright is similarly based on the principles of open access and crowd-sourcing; any user may create a free login through 18thConnect. Once logged in, TypeWright users can compare digitized facsimile page images of works from ECCO with the text versions transcribed electronically with OCR. Users can correct, line by line, the "dirty" OCR transcriptions; these corrected texts enable more thorough searching, mining, and overall usability for scholars and students working in the eighteenth century. TypeWright offers a clear solution to the current transcription problems that plague important scholarly resources.[4]

In the classroom, TypeWright makes clear these problems of "dirty" OCR while also

[3] 18thConnect and its partner sites through ARC (The Advanced Research Consortium) employ Collex—a tool suite that "allows users to collect, annotate, and tag online objects and to repurpose them in illustrated, interlinked essays or exhibits" ("About," *Collex*).

[4] In addition to TypeWright, other projects are addressing the pressing issues of OCR errors and the publics' access to digitized texts. One notable example is the Text Creation Partnership (TCP). Phase I of EEBO-TCP was released in January 2015, giving the public access to more than 250,000 manually transcribed texts from EEBO; more than 2,000 texts are publicly available through ECCO-TCP. However, the TCP only gives access to transcriptions; the digital facsimile page images remain available only to those with subscriptions to EEBO and ECCO (see "EEBO-TCP Phase I" and "ECCO-TCP").

fostering discussions about much larger issues about digitization. I have used two different approaches to integrating TypeWright into courses. One option is to introduce problems of "dirty" OCR and then collaboratively correct a text in class with TypeWright; this approach uses the minimum amount of class time, yet still conveys poignantly the pressing problems of OCR.

Another option, and the one that I explore in more detail in this essay, is to dedicate two class sessions to issues of OCR and TypeWright: the first introducing the concept of OCR and demonstrating TypeWright and the second discussing it. TypeWright is a straightforward tool, and a session devoted to using it may not be necessary in all cases. However, my experiences teaching with TypeWright and tools such as Voyant Tools or the XML-editor oXygen suggest that students (especially humanities students) have little or no experience using interactive digital tools in an academic setting. As a result, students are excited but also anxious: they have concerns about grades, fears that a tool may crash and they will be penalized, and worries that they do not have adequate training to use or critique a tool. An initial computer lab class session alleviates many of these anxieties and, I think, ultimately results in more robust discussion in the following class. For this second option, a computer lab space works best. A computer tutorial offers a chance to troubleshoot questions that arise about the tool itself, and I make sure everyone leaves the session with a user login for 18thConnect, so that they could return to the site later to use TypeWright. This session also offers students an opportunity to ask questions about vagaries of eighteenth-century typography, which many of my students have not previously encountered. At the end of the first session I provide a list of questions that will form the basis of the following session's discussion: What type of OCR transcription errors did you notice? Were they the same errors again and again or a variety of errors? Why do you think the computer had problems reading the text correctly? What problems did the TypeWright interface

present? What advice would you give to the tool's developers? I require students to correct at least four pages of a text they have found on TypeWright and to then write a short reflection that addresses the questions I circulate.

No matter how many class sessions I'm dedicating to issues of digital mediation and reliability, I always first introduce the concept of OCR and its potential for errors. The fastest, most effective way to do this is to use Google Ngrams to demonstrate the infamous "fuck" versus "suck" problem, illustrated below.[5]

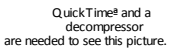

Figure 1: fuck, suck Ngram

Students are usually shocked (or at least gleefully surprised) to find out just how much eighteenth-century authors swore. Obviously, the "fuck problem" is the result of OCR mistranscriptions of the eighteenth-century long "s," something that a quick search of the actual books reveals. For example, the numerous uses of "fuck" in Josh Ash's *A New and Complete Dictionary of the English Language* (1775) actually occur in the "S" section of the dictionary, containing words related to "suck" such as "suckle," "suckling," and "suction." The sharp

[5] My use of the "fuck, suck" Ngram in my classes takes its lead from Adam Hammond's lecture "Historical Research in Electronic Archives, Pro and Con."

decline of "fuck" and its eventual intersection with "suck" in the early nineteenth century corresponds with the declining use of the long "s." While effective, such drastic examples run the risk of oversimplifying the problems of OCR and digitization. This is where TypeWright comes in.

Using TypeWright in the classroom demonstrates the imperfect transition from material book to digital resource. While in 2000 Clifford Lynch pointed to a "widespread distrust of the digital environment," such assessments are less accurate today. The distrust Lynch identified may still hold true for academics and librarians, but the students I have taught are not skeptical about the validity of digital resources. While students are often unsure about how to use digital archives and repositories effectively, the digital medium itself does not signal an alarm.[6] Moreover, latent skepticisms are often put to rest when students access digital resources through an official university gateway, be it links from a course website or a library catalogue. Even after seeing the problems with the long "s" in Google Ngrams, students who had previously heard of resources like EEBO and ECCO still classified them as reliable, accurate, and scholarly. After briefly exploring ECCO in our computer lab session, the majority of students tend to agree, indicating a willingness to accept the validity and accuracy of approved scholarly sources. In several of my classes, students have hypothesized that the transcription problems in Google Books would not be as significant for these other scholarly resources. They suggest that, if universities paid for digital resources then they must be good.

One unsettling aspect of trusted (and expensive) digital repositories like ECCO or EEBO is that the page images give the user a false sense of accuracy. That we view the scanned page image rather than read the OCR transcription inherently encourages assumptions that our

[6] Indeed, I have seen an increase in trust for digital media coinciding with the growing frequency with which online resources and readings are central to undergraduate courses.

searches are unproblematically linked to the digitized facsimiles we see. The way search results are displayed reinforces these assumptions, since, as Patrick Spedding has pointed out, computer transcriptions are "mated to the appropriate part of the photographic image of each page. This coded linkage allows users to search for a word (or a combination of words), and to call up and see a digital facsimile of the pages where each term appears, with the search words highlighted" (438). This highlighting masks the mediation at work. Of course, students were aware that the page images on ECCO were not the "real" thing and that the images on the screen mediated a material book. Still, with the page images there, it is easy to forget that "the text running behind the page images of these texts has been mechanically typed, leaving behind errors" (TypeWright). Indeed, through the ECCO interface, users are not given access to the OCR transcriptions, enforcing ideas that the texts of the original physical book, the viewed facsimile text, and the searchable transcribed text are the same.[7] A further layer of mediation not readily apparent to most of my students is the fact that the page images are digital copies of microfilm.[8] One of the best things about TypeWright is that it makes these layers of mediation more obvious.

The plethora of transcription errors that students can see and correct with TypeWright brings to the fore the mediated layers between the original material book and searchable computer transcriptions. Many errors that students find arise from ligatures and the long "s." Such errors provide more instances of the f/s problems seen readily with Google Ngrams. Indeed, after seeing the "fuck, suck" Ngram one student rightly asked, "So what's the big deal? Why not just make the computer read that s/f letter better?" This is a good question, but anyone who has looked at OCR transcriptions knows that eighteenth-century typography is not the only

[7] Tellingly, one of the questions in the FAQ section of ECCO concerns access to the underlying text files created by OCR, perhaps indicating growing awareness of "dirty" OCR. Users are informed that they cannot access the text files since they "are used solely for searching the product and the user is only able to view the digital image of the page." ("Frequently Asked Questions").

[8] For more on the problems that microfilm quality presents to digitization and computer transcription, see Spedding 440-41 and Baker 138.

problem. Hand-press period printing was a difficult, error-prone task, "so errors in [computer] transcription may be caused by broken, badly inked, or warped lines of type" (Spedding 438). Though I usually ask each student to correct only a few pages, most of them still encounter a variety of transcription errors, allowing them to see firsthand that the problems with applying OCR to older texts are numerous and complex.

For instance, students discover that the computer attempts to transcribe ornamental letters, decorative ornaments, printers' emblems, and library stamps. Unsurprisingly, the transcription of these elements results in a jumble of odd, nonsensical characters. Presented with these "transcriptions," most of my students have chosen to use TypeWright's option to delete them, but for a handful of students these errors suggest theoretical questions about what it means to transcribe a text. Was there some way to identify these marks? What if researchers were looking for illustrations in a text rather than words? Could they use a resource like ECCO to find texts they needed? Who controls what search options are available? Such questions often lead to student-centered conversations about structures of power and access to information.

TypeWright is free, but institutions must pay for access to ECCO. Thus, the content of resources like ECCO and EEBO gestures toward inclusivity, but their terms of access are severely limiting. Through their donated labor, however, TypeWright users who correct texts from ECCO may reclaim them for public use. Students typically see reclaiming these texts as a good idea. However, working with TypeWright raises contentious questions about which texts should be transcribed (and thus reclaimed) first. In many ways the discussion that my students have evolving out of their experiences with TypeWright represent a microcosm of larger debates in the humanities about access and recovery work. Usually in discussions some students will suggest that the "important works" be edited first, which prompts questions about what is

"important" and who should decide. Students hit on a sensitive problem with digital archives. Julia Flanders has argued that by digitizing collections, aggregating them, and making them discoverable, "minority literatures, non-canonical literary works, and the records of what goes on in (what appeared earlier to be) the odd corners of the universe are all given a new kind of prominence and parity with their more illustrious and familiar cousins" (Flanders 5). If we correct OCR errors in only "important" works, my students have pointed out, maybe scholars would continue to miss things that they did not know were there. Thus, on the surface ECCO texts corrected with TypeWright would seem inclusive, but the enhanced searchability of "important" or "canonical" texts would yield lopsided results. Those minority literatures, then, would have the appearance of being easily discoverable when they were actually less so than their "illustrious and familiar cousins" (Flanders 5).

The next logical step, students suggest, is to correct everything in ECCO, which is an implicit goal of the TypeWright project. Here is where the response reports that the students submitted between class sessions one and two become especially useful for guiding discussion. The majority of students that I have taught have independently proposed various developments to TypeWright that would encourage more users. Students envisioned gaming aspects with high scores and champion transcribers that, they hypothesized, would encourage large-scale public participation. Prizes and recognition might be awarded to users, and some students wanted to earn badges for successfully correcting the OCR for an entire text. They suggested adding social networking aspects with links to Facebook and Twitter, so that users could connect, converse, and collaborate with each other. While it seems unlikely that TypeWright's developers will gamify the tool, it also seems important to me that students started to think beyond the interface

in front of them.[9] In seeing the limits of a respected resource like ECCO and recognizing what a digital tool like TypeWright might become, students were also learning to challenge things as they are. In other words, to teach with TypeWright, to ask students to identify its limits and its potential, is to teach innovative thinking. While thinking of new ways that TypeWright could be extended, students were moving towards Kolb's understanding of active experimentation (Svinicki and Dixon 142). This type of playful experimentation exemplifies a hacker ethos that encourages students to consider how a tool or interface could be modified, changed, or repurposed.

Over the past several years hacking has become an increasingly important term in both digital humanities and (digital) pedagogy. One need look no further than sites such as *Prof Hacker*, *Grad Hacker*, and *Hacking the Humanities* to see hacking's contemporary importance. While historically hacking could mean breaking things (including firewalls and laws) the term has recently gained more positive associations with exploration, experimentation, and innovation. Moreover, hacking is no longer a term necessarily related to computer programming. Patrick Murray-John, for instance, has claimed that, in addition to computer coding, hacking also means "exploring what the humanities are about within digital—mostly, but not exclusively, online forms—and how that could disrupt familiar systems in interesting and useful ways" ("About," *Hacking the Humanities*). This more general understanding of hacking has also been embraced by others, including Tad Suiter and Mark Sample. Thus, a hacker ethos is related to playfulness and subversion, though not necessarily coding (Suiter). It can involve brainstorming alternative possibilities without necessarily insisting that the end goal is for those possibilities to

[9] Notably, some other crowd-sourcing transcription resources have gamified their systems. One great example is Old Weather, a resource that asks volunteers to transcribe nineteenth-century ship logs so that scientists can mine them for meteorological data. Those who transcribe texts join a particular voyage and have opportunities to be promoted from Cadet all the way up to Captain (Old Weather).

be implemented by the hackers themselves or, perhaps, even at all. Thinking about how one might "hack" TypeWright embraces this type of thinking but without computer coding playing a role. Encouraging students to think about interface design and to interrogate their own interactions with a digital tool has larger implications beyond TypeWright and beyond awareness of electronic databases. Thinking about ways of encouraging students to become hackers suggests new continuities between standard pedagogical definitions of experiential learning and contemporary discourses about undergraduate research and digital pedagogy.

While I could easily give students readings about the limitations of digital databases, TypeWright brings these issues into focus readily and interactively. Employing TypeWright in the undergraduate classroom produces learning outcomes on multiple levels. First, it gives students firsthand exposure to the problems that eighteenth-century typography can present to digitization. Second, it reveals the latent problems of OCR and databases such as Google Books and ECCO that rely on it. Third, TypeWright spurs conversations about wider themes relevant to scholarship in the digital age: What are some of the current problems with the digital tools we use? Who owns these digital tools? Who controls access to digital and digitized texts? What role does community play in scholarship in the digital age? Finally, discussions about TypeWright facilitate the development of critical thinking skills and hacking ideals and urges students to question mediation and authority in digital works. Playing with TypeWright encourages students to think outside the box or, more appropriately, to think outside the bounds of a given interface.

Works Cited

"About." *Collex*. n. pag. Web. 26 Aug. 2012. < http://www.collex.org/?page_id=2>.

Baker, Nicholson. *Double Fold: Libraries and the Assault on Paper.* New York: Random House, 2001. Print.

Berland, Kevin. "Formalized Curiosity in the Electronic Age and the Uses of On-Line Text-Bases." *The Age of Johnson: A Scholarly Annual* 17 (2006): 391-413. Print.

Blackwell, Christopher and Thomas R. Martin. "Technology, Collaboration, and Undergraduate Research." *Digital Humanities Quarterly*. 3.1 (2009). Web. 30 Oct. 2012. < http://www.digitalhumanities.org/dhq/vol/3/1/000024/000024.html>.

Cayley, Seth. "Digitization in Teaching and Learning: The Publisher's View." *Victorian Periodicals Review*. 45.2 (2012): 210-14. *Project Muse.* Web. 30 Oct. 2012.

"EEBO-TCP Phase I Public Release: What to Expect of January 1." *Text Creation Partnership*. n. pag. 24 Dec. 2014. Web. 4 April 2015. <http://www.textcreationpartnership.org/2014/12/24/eebo-tcp-phase-i-public-release-what-to-expect-on-january-1/>.

"ECCO-TCP: Eighteenth Century Collections Online." *Text Creation Partnership*. n. pag. Web. 4 April 2015. <http://www.textcreationpartnership.org/tcp-ecco/>

Flanders, Julia. "The Productive Unease of 21st-Century Digital Scholarship." *Digital Humanities Quarterly* 3.3 (2009): 1-27. Web. 7 Aug. 2012. <http://www.digitalhumanities.org/dhq/vol/3/3/000055/000055.html>.

"Frequently Asked Questions." *Eighteenth Century Collection Online*. Gale. Web. 4 April 2015.

Hammond, Adam. "Historical Research in Electronic Archives, Pro and Con." ENG287: The

Digital Text. University of Toronto, Toronto. 12 Oct. 2011.

Hammond, Adam and Julian Brooke. *He Do the Police in Different Voices: A Website for Exploring Voices in T.S. Eliot's* The Waste Land. Web. 7 Aug. 2012. <http://hedothepolice.org>.

Holley, Rose. "Many Hands Make Light Work: Collaborative OCR Text Correction in Australian Historic Newspapers." National Library of Australia, March 2009. Web. 7 Aug. 2012. < http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf>.

Lynch, Clifford. "Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust." *Authenticity in a Digital Environment*. Council on Library and Information Resources, May 2000. n. page. Web. 5 Aug. 2012. <http://www.clir.org/pubs/reports/pub92/lynch.html>.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science*. Published Online Ahead of Print: 16 December 2010. Web. 20 Aug. 2012. <http://www.sciencemag.org/content/331/6014/176>.

Murray, John. "About." *Hacking the Humanities*. n. pag. Web. 15 Sept. 2012. < http://hackingthehumanities.org/about>

Old Weather. Zooniverse: 2013. Web. 4 April 2015. < http://www.oldweather.org>.

Sample, Mark. "Hacking Campus Space." *Prof Hacker*. 3 May 2012. n. pag. Web. 10 Aug. 2012. <http://chronicle.com/blogs/profhacker/hack-your-learning-spaces/39911>.

Svinicki, Marilla D. and Nancy M. Dixon. "The Kolb Model Modified for Classroom Activities." *College Teaching*. 35.4 (1987): 141-46. Print. *JSTOR*. Web. 12 Oct. 2012.

Schreibman, Susan. "Digital Representation and the Hyper Real." *Poetess Archive Journal* 2.1 (December 2010): 1-16. Web. 5 Aug 2012. <http://paj.muohio.edu/paj/index.php/paj/article/view/7/53>.

Spedding, Patrick. "'The New Machine': Discovering the Limits of ECCO." *Eighteenth-Century Studies* 44.4 (2011): 437-53. Print.

Suiter, Tad. "Why 'Hacking'?" *Hacking the Academy, The Edited Volume.* Ed. Dan Cohen and Tom Scheinfeldt. MPublishing, 2010. n. pag. Web. 10 Aug. 2012. <http://www.digitalculture.org/hacking-the-academy/introductions/#introductions-suiter>.

TypeWright. Web. 7 Aug. 2012. < http://www.18thconnect.org/typewright/documents>.