DHQ: Digital Humanities Quarterly
2015
Volume 9 Number 4

TypeWright: An Experiment in Participatory Curation

Alan Bilansky <alanbilansky_at_gmail_dot_com>, University of Illinois, Urbana-Champaign

Abstract

TypeWright (housed on the website 18thConnect) is an experiment in participatory curation; it asks volunteers to make texts more findable, useable, and trustable. These contributions are not without rewards to the volunteer. TypeWright is part of some important trends in digitization, addressing two problems of digital texts: flawed optical character recognition (OCR) and the complicated terrain of intellectual property.

The eighteenth century was a golden age when print first came to the crowd, and TypeWright brings crowdsourcing to the texts of that century. An experiment in participatory curation, it asks volunteers to make texts more findable, useable, and trustable. These contributions are not without rewards to the volunteer. TypeWright is part of some important trends in digitization, and we might say it is part of a growing movement, addressing two problems of digital texts: flawed optical character recognition (OCR) and the complicated terrain of intellectual property. 1

TypeWright is housed on 18thConnect, one of four sister sites, all coordinated through a group called the Advanced Research Consortium (ARC). Since 18thConnect launched in 2010 its director, Laura Mandell, and editorial boards are actively reviewing projects and collecting metadata to make online objects discoverable and collectable through their site. The other three sites are NINES (Networked Infrastructure for Nineteenth-Century Electronic Scholarship, both an organization and Website), the Renaissance English Knowledgebase (REKn), and the Medieval Electronic Scholarly Alliance (MESA). The first of these, launched in 2005, is NINES. NINES allows federated searching of 124 online scholarly projects, such as the Whitman Archive and the Rossetti Archive. Its editorial boards review these projects and then the site ingests the metadata. Scholars logging into NINES can search all these resources at once, and collect results into “exhibits” that can be used in teaching and scholarship. In doing so, scholars are volunteering their labor to a good cause. Collex, an open-source collections-and-exhibits builder, was originally designed to store the data from the searches and the exhibits created and to use it to improve search results. For an excellent review of the search and exhibit functions of Collex, turn to Amy Earhart’s reflections on using NINES in a college course [Earhart 2010]. 2

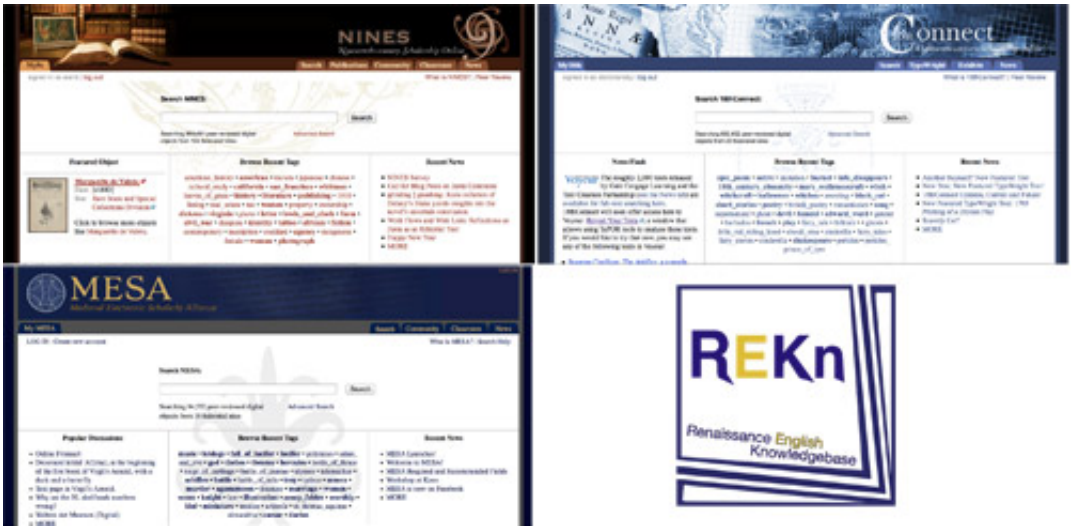

Figure 1. The four sister Websites belonging to the Advanced Research Consortium.

ARC's other three sites run on the same software and have the same editorial and publishing model. MESA (the Medieval Electronic Scholarly Alliance) came online in 2013. REKn, the Renaissance English Knowledgebase, has yet to launch. After launching in 2010, 18thConnect, the focus of this review, is actively reviewing and collecting metadata to make online objects discoverable and collectable through their site. 3

Anyone can create a login to 18thConnect, and will find tabbed links to the functions of Collex, including “Search”, “Exhibits”, and “Peer Review” (where editors can submit online resources for inclusion in searches here). There is also a link to “TypeWright”. 4

Understanding TypeWright requires a brief detour through the history of Eighteenth Century Collections Online (ECCO). It was created by Gale Cengage, scanning microfilm images of texts published in Great Britain between 1701 and 1800. Although comprehensiveness is not possible, ECCO has a substantial collection of 155,000 texts from that century printed in Great Britain and America. I know researchers who have begged, borrowed, and occasionally stolen access to it if they weren’t lucky enough to work for a school that subscribes to it. In the world of large collections of primary texts, ECCO is dwarfed by behemoths like Google Book Search, but it makes up for size with its relative accuracy in metadata such as author and date of publication. At the same time, two other publishers created similar collections: Chadwyck-Healey (later ProQuest) created Early English Books Online (EEBO), collecting images of British texts published before 1800; and Readex created Early American Imprints (Evans for short), a digital collection of American texts before 1819. All of these collections consisted of high-quality scans of microform (not always of high quality to begin with). 5

Transcriptions of these images were created by optical character recognition (OCR), mainly for the purpose of searching to locate images of the pages. OCR is prone to errors, particularly with structural information (title pages, chapter headings, etc.) and notoriously with features of eighteenth-century printing like the long S. 6

The more accurate our transcriptions are, the more we can trust our searches to locate texts to read, and the more we can trust algorithmic approaches to a body of texts. 7

The Text Creation Partnership is a consortium of more than 150 academic libraries that started at the University of Michigan and Oxford University, with the mission of producing human-keyed, accurate transcriptions of texts digitized in EEBO, Evans, and ECCO. Work continues on EEBO texts, but has ceased after transcribing 4,977 texts from Early American Imprints and 2,331 ECCO texts. The page images from microform remain the property of the publisher, but the transcriptions belong to the institutional members of the TCP, and after five years they pass into the public domain. 8

TypeWright aims to continue this effort by using scholarly volunteers to check and correct the OCR transcriptions. 9

Scholars can work through the OCR transcripts of ECCO texts that the TCP could not get to, making the texts more useable and more available. A similar project, to correct OCR transcripts of the EEBO texts, is to be launched on REKn.

An example will demonstrate how this works: Professor Y, a scholar of the eighteenth century, already has a text in mind. Perhaps she plans to teach *Edgar Huntly* or *The Female Quixote* next semester, so she needs to reread the book anyway. So she logs in to 18thConnect, goes to TypeWright, and searches for a text. 10

Figure 2. Search Results on 18thConnect.

In a list of results, if a new editor clicks on the title of a text, a confusing error message appears. That's because TypeWright has access to the texts, and users can see them only through through the TypeWright interface (otherwise, we can only access ECCO though our institutions, if they subscribe to it). So, if you click on the "Edit" button, you can begin reviewing the transcript. 11

My first impression of working in TypeWright was how different the interface is from Transcribe Bentham, a similar crowd-sourced project. This difference reflects that the two tools are conducting distinct curatorial tasks. In the case of Transcribe Bentham, there are no transcriptions yet and we humans are called upon to create them. TypeWright texts are beginning with transcriptions produced by a machine, in need of human attention to correct them. TypeWright editors generally won't be starting with a blank page. 12

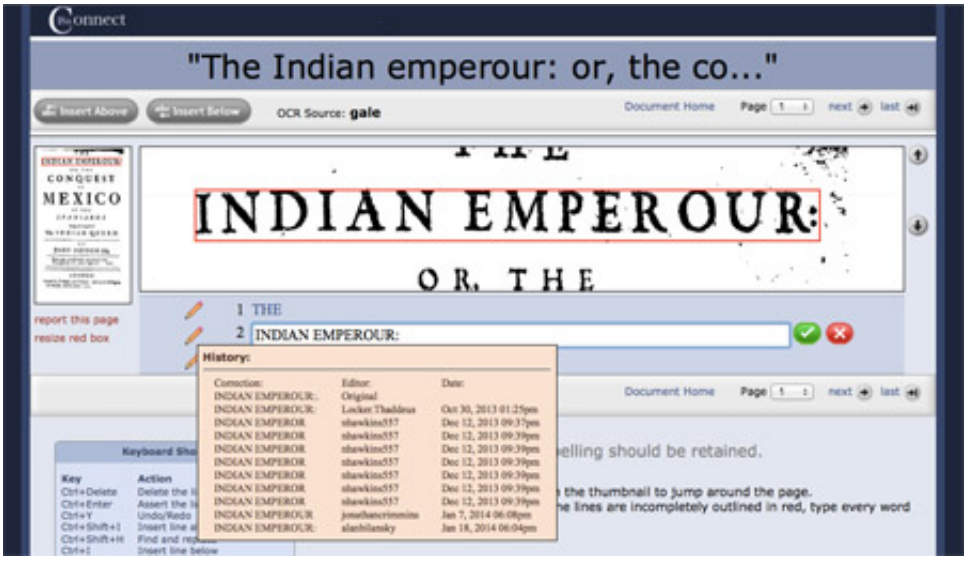

Figure 3. Editing a page in TypeWright.

After a user clicks the button to "Start Editing," their workspace appears: a blown-up image of one line of text, a space to edit the line as currently transcribed, and a thumbnail image of the page being edited, with a box locating the current 13

line. The user can edit the text, approve the current transcription or delete it entirely. One can also move around the page or around the document. The pencil icon reveals the edit history of the current line. It's worth noting here how easy it is to accidently delete a line from the transcript (an undo button silently appears, but it's easy to miss). Generally, there is an assumption that users of this site will take this work seriously and discharge their editorial duties with great care.

Once editing is complete, the user can click the "Mark Complete" button on the last page. 18thConnect has procedures in place to check the edited text and notify the user if the document has been declared complete, or what work remains to be done. The organization will also provide a letter and a certificate of completion. Documentation of a partly-edited work is also possible (e.g., "Professor X has corrected 37% of this 500pp text"). 14

This is the sort of documentation that can turn into credit in a seminar or a line on a CV. I say this expecting a *de minimis* benefit for those seeking jobs or promotions and tenure — not because digital work is undervalued, but because editorial work is generally undervalued. 15

There's more. When an editor completes a text, she owns it. 18thConnect will send the editor either of two versions of the text, one plain text, and one minimally marked up in TEI. The Text Encoding Initiative's XML standard is the basis of most contemporary textual curation efforts. The TEI version of the text is intended to serve as the foundation for a scholarly edition. Someone interested enough to correct the transcription of a text might also be moved to publish an edition. (An example of a scholarly edition that started as a Typewrite text is Jess McCarthy's edition of Daniel Defoe's "Hymn to the Pillory": http://ahymntothepillory.blogspot.com/.) The plain text version would be ideal for carrying out text mining approaches to the text, as in this case TEI tags would just get in the way. 16

As the screenshot demonstrates, the first lines of the "Featured text" (figure 4), a single line can be edited multiple times. 17

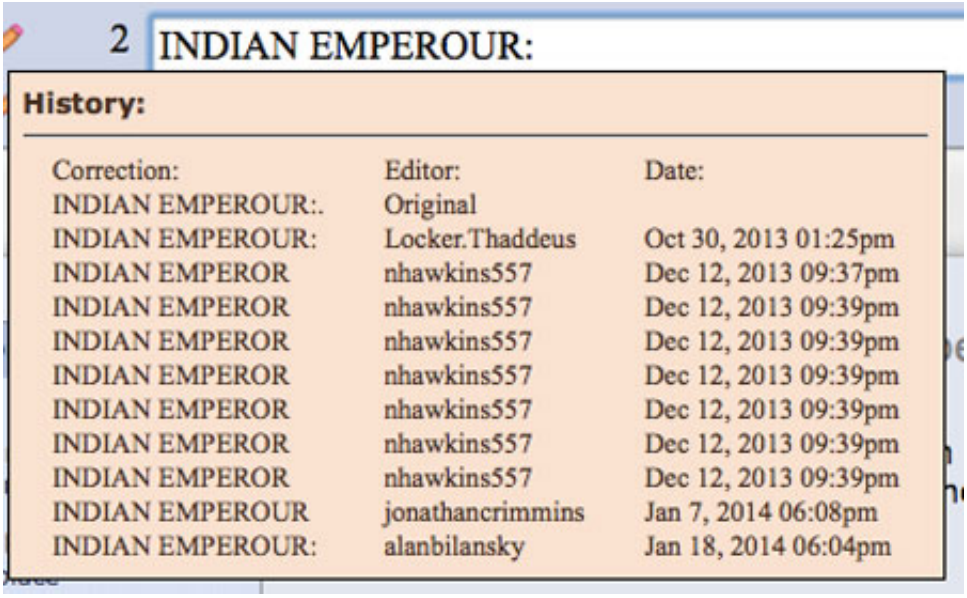

Figure 4. The edit history of a single line.

The back-and-forth we see in the version history of that line casts some doubt on the assumption that all editors will discharge their duties with the same care. 18thConnect's preference is for an editor to take ownership in a text and edit it from beginning to end, and there are good reasons good scholars would want to. 18

A few technical improvements would enhance this tool's usability. (1) It could be useful to allow scholars to formally adopt texts (e.g., a message like "User Professor X is actively working to correct this text" would display in the interface). (2) It would be very useful to bookmark a text where you left off. Or perhaps a user could automatically go to the first uncorrected page in a text, instead of always starting on the first page. (3) And as was mentioned, it seems too easy to delete a line from the transcript. (4) It would also be useful to include information on eighteenth-century printing. 19

The typical graduate student might not be certain of how to handle features like the long S or graphemes. As a benchmark, Transcribe Bentham includes basic information on paleography.

What I would most like to see is incorporation of the texts of Early American Imprints. ECCO contains some American texts but nowhere near a comprehensive collection, and 18thConnect includes Americanists of course. My other suggestions all require mere technical effort. Adding texts from Evans would require a whole new set of negotiations with a different publisher. Just the same, we are waiting for someone to come forward and take on the challenge of making the TCP-Evans texts easily available for text mining or the uncorrected OCR available for correction in the way that TypeWright does for the ECCO texts. 20

In the near future, REKn will be providing access to the TCP texts from EEBO. In the very near future, users of 18thConnect will be able to run the corrected ECCO texts through Voyant directly. Voyant is an easy-to-use text-mining tool that makes it easy to find trends in a text or body of texts. 21

In an environment where most funding comes from fixed-term grants and it is easier to fund a pilot than a sustained effort, a question we have to ask about any online resource is how long can we count on using it. The fact that 18thConnect is part of a larger organization and has seen past and planned advances serve as reasons to expect it to stick around. 22

I am a zealot. I think every student in English studies or in information science should contribute to a particular curatorial project like TypeWright. Participation would serve a valuable function in a seminar in digital humanities, text modeling, digital curation, literature, history, or scholarly editing. Students of literature and history and related fields should volunteer to learn about eighteenth-century printing, digitization, and editing. Even more important, those who study this century should volunteer for the cause of making these texts more accessible. 23

Works Cited

Earhart 2010 Earhart, Amy. “Using NINES in the Classroom.” ProfHacker (2010). http://chronicle.com/blogs/profhacker/using-nines-collex-in-the-classroom