4-28-2018

Patterns, Collaboration, Practice: Algorithms as Editing for Historic Periodicals

Elizabeth Lorang
University of Nebraska - Lincoln, llorang2@unl.edu

Elizabeth Lorang

American Antiquarian Society Spring Symposium 2018

Editorship as Collaboration: Patterns of Practice in Multi-Ethnic Periodicals

April 28, 2018

Patterns, Collaboration, Practice: Algorithms as Editing for Historic Periodicals

Good morning. Before I begin, I want to be sure to thank Jim Casey, Sarah Salter, our hosts at the AAS, especially Molly Hardy, those of you who have already shared your work and ideas thus far in the symposium, and finally, my co-panelists and our moderator. There was much I considered adding or revising, following not only on yesterday's presentations but also on conversations I had with many of you, including about extractiveness/extraction and newspapers, of collaboration, and more, but in the end I decided to leave those as opportunities for discussion.

In recent years, I have not really thought of myself as an editor, or at least a practicing editor. I have worked in a number of editorial capacities, including as the co-editor of an edition of poems published in newspapers, but that work now dates to 2013 and was published nearly 5 years ago. (Of course, until I went back and looked at the dates, in my mind it wasn't yet *that* long ago.) Based on the time that had passed and where my work had gone in the interim, I was therefore more than a little surprised by the invitation to participate in this symposium, surprise I shared with Jim Casey at the time of the original invitation. At that time, I wrote and expressed my uncertainty about my fit, explaining that my recent work focuses on "investigating methods of increased discovery and access in digitized newspaper collections of all kinds," rather than editing. But, Jim was, thankfully, encouraging, and I've appreciated the opportunity this symposium has provided to begin bringing some different aspects of my work together.

What I want to do today is position my recent work on the algorithmic "discovery" of poetic material in historic newspapers within the contexts of my various roles as an editor of periodical literature and also consider how duplicative processes and algorithms encode principles and values and function as editorial acts. Ultimately, I hope to pose a range of questions to prompt discussion around the place (or not) of machine learning in identifying and selecting texts and bodies of work; what ideas we're actually exploring/are able to explore when we enlist technology in stages of this work; and the stakes of these activities, whether human or machine, for periodicals from under-represented communities in particular. I've divided this up into three parts—which means I can't spend much time on any one part, so rather than being anything near comprehensive, I want to sketch out some broad, and at times quite rough, ideas as starting points for further conversation.

Background and Work as an Editor

First, a little about me: My experience in editing began as a graduate student working with Susan Belasco at the University of Nebraska-Lincoln. Susan invited me to join her in editing the poems that Walt Whitman published first in periodicals over the course of his career, an edition that appears on the *Walt Whitman Archive*. By definition, this edition was author-centered, though we aimed also to make it periodical-centered. Beginning with that edition, I worked in a number of editorial roles for the *Whitman Archive*, and I ended up writing a dissertation about newspaper poetry. In 2013, I undertook creating an edition, with R. J. Weir, of newspaper verse: "'Will not these days be by thy poets sung," selected poems from the *Anglo-African* and *National Anti-Slavery Standard* from 1863-1864. We sought to create an edition that was temporally-centered, periodical-centered, and community centered.

"'Will not these days be by thy poets sung"

Our work on "Will not these days" fits with many of the themes of this symposium, notably collaboration and several "patterns of practice" that we identified at the time—and which, I think, warrant further investigation: Robert Hamilton and Thomas Hamilton collaborated with one another on producing the *Weekly Anglo-African*, and we also raise the question and possibility of some degree of collaboration among the *Anglo-African* and *National Anti-Slavery Standard*. In the introduction to our edition, we wrote, "The weekly *Anglo-African* and the *National Anti-Slavery Standard* were involved with each other in telling and surprising ways. Our comparative approach at once reveals connections between titles and extends contemporary recovery projects into the neglected realm of newspaper verse."

For the period we explored, the *Anglo-African* and the *National Anti-Slavery* standard were located at No. 48 and No. 50 Beekman St. in New York. (We trace out the addresses and their history in much more detail in our introduction.) "As our work makes clear, poems regularly appeared in the *Anglo-African* one week after appearing in the *Standard*. Of course, the Hamilton brothers may have reprinted *Standard* poems from other newspapers, but in many cases the short amount of time that elapsed between printings makes it all but impossible that the *Anglo-African* reprinted from an intermediate text. Further, the *Anglo-African* expressly acknowledged the *Standard* as the source of poems at a time when popular convention dictated that reprints of reprints need not be acknowledged. (That is, if the *Anglo-African* was reprinting from intermediate texts, it was under no obligation to credit the *Standard*.) The pattern of reprinted poems and the proximity of the newspaper offices therefore suggest that the *Standard* was, at the least, one of the titles that the Hamiltons scanned for selected verse.

"On at least one occasion, a poem original to the *Standard* appeared in both papers on the same day, with the *Anglo-African* crediting the *Standard*. And perhaps even more remarkable is the fidelity with which the *Anglo-African* reprinted the text and adhered to the layout of *Standard* poems. In nineteenth-century practice, the process of reprinting commonly introduced changes to a poem, whether at the direction of an editor or at the whim or error of a compositor. Yet, *Anglo-African* reprints of *Standard* poems are almost always identical, including the perpetuation of clear errors in the *Standard* version. In one instance, the *Anglo-African* even preserved the 'For the Anti-Slavery Standard' note that preceded a poem. [Such] details raise the tantalizing possibility that the newspapers' editors or staffs may have actively collaborated [in some capacity]."

We also were interested in centering our editorial work around something other than, or at least in addition to, the author. We wrote in the introduction to our edition that "One of the aims of our edition is to expose and problematize the unit of the author as the primary organizational model and as the requisite point of access for this body of work." I think there are still some ways and opportunities to expand this idea further/further complicate the idea of organization around authorship in such editions.

There are so many threads that remain to be untangled, or at least explored, in this work, and revisiting it now, I can't help but confront the impulse to want to *do all of the things*. Over the last several years, however, the emphasis of my work, along with a host of collaborators, has been around a different set of questions related to historic newspapers and poetic content, among other genres. There is, though, a connecting question of sorts: how can we—librarians, archivists, editors, literary historians, and more (and those categories or course are not exclusive to one another!)—facilitate the work of locating content of interest in historic periodicals, to

encourage, promote, and make possible analysis of all kinds, including editorial work? Or so that others might analyze them from a variety of methods, including computational ones? Indeed, one of the reasons Weir and I set the constraints on our project that we did—two newspapers (originally the plan was three), full-length poems, weekly newspapers, roughly a 12-month period of time for the publication dates of the poems—was to deal with scale, the numbers of poems, and the time spent in finding or locating them. We had more high-minded reasons as well, but these very practical concerns are of significance.

Project Aida

Shortly after the release of the edition, I was fortunate to receive NEH start-up funding for another project, a research endeavor to explore the possibilities of computational image analysis for helping to identify poetic content in historic newspapers. Since that time, most of my research has focused on this particular challenge and idea—using digital images of historic newspapers to identify and locate visual patterns in textual materials. This research has taken me away from active editing, at least insofar as how I had been conceiving of editing, but some of the themes of this symposium, including collaboration, or what I've started thinking of more as *systems*, patterns—and patterns of practice—are fully present there as well.

In 2014, with Leen-Kiat Soh at the University of Nebraska-Lincoln, I launched the Image Analysis for Archival Discovery project, also known as Aida. Our idea was to use image processing and analysis, a subfield of Computer Science, to identify poetic content in historic newspapers based on visual features. While others were exploring topic modeling and other linguistic-based approaches, we would not consider the textual content of the poems at all, as text, but rather only as visual signals on the page, which could be represented mathematically.

The idea was that if we could model the visual features of poetic content in historic newspapers, based on their visual signals, then we could use the models to find poetic content at scale across digitized newspapers—and across languages. We published preliminary promising results in 2015.

In 2016, the scope of our project grew. We are working to develop a scalable and generalizable system. In addition, while we remain fundamentally interested in poetic content in historic newspapers, we are pursuing extending the approach to other visually distinctive textual forms as well. Our research team also grew to include collaborators at the University of Virginia—John O'Brien, in particular. With this growth, our project increased in both geographic scope (now including British newspapers) as well as temporally, to deal with newspapers from the 18th century as well as the 19th and 20th. This temporal and geographic spread also meant we would be working with a broader range of newspaper forms. In addition, John's end goal is creating an edition of poems from newspapers in the Burney and Nicholls collections at the British Library and to test whether an approach such as the one we are developing can aid in this work.

Now, I want to emphasize that the work I'm talking about deals specifically with historic newspapers that *have been digitized*. Immediately this decision leaves out many periodicals from communities that dominant white, cis-heteronormative communities have historically marginalized, silenced, and minoritized—and often continue to do. Several years ago, in a panel at a meeting of the American Literature Association—and in the follow-up sequence of essays from the panel in American Periodicals—Benjamin Fagan addressed the white supremacist model of Chronicling America, and I believe it was Jean, speaking from the audience that day,

who as well powerfully raised the issue of the un-digitized record and implications for those materials not digitized.

It's important for me to raise this limitation of my project from the outset—and to recognize the periodicals and voices it leaves out from the outset. When I say that the project is about finding poetic content in historic newspapers, what I really am saying is that it's about finding poetic content in the historic newspapers deemed significant enough for digitization, which is also layered on top of decisions made about what newspapers were significant enough for collecting and in what strategies. It also means that the models we develop and train are based only on those newspapers, which sort of becomes a reifying reality, which is something I'll say more about this in a bit.

Algorithms as Editing

Now, this work of identifying content in historic newspapers is not the same as textual editing or of creating editions, but I do want to consider algorithms and the software systems that run those algorithms as sites of editing. In doing so, I am regarding editing more broadly as a series of arguments about what matters to the editor, about what the editor values, about what the editor wants to elevate as central for conversation and understanding, and which arguments get enacted in decisions and choices throughout the system.

In a way, editing is a set of decision points about what's in and what's out, of how to deal with particular types of situations and recording those decisions in policy and methods. Positioned in this way, the type of algorithmic modeling and algorithmic selection that my colleagues and I are doing is clearly editing. For example, our work of creating rules for visual

patterns is an act of definition and selection, of saying what is in and out of scope, whether for an edition or a corpus.

As you might imagine, what we actually locate with our approach is not always poetic content according to the definitions we might bring from other contexts: some of the content we end up identifying are advertisements that present themselves visually as poems. In other cases, dialog can present as a poem, according to our rules and definitions. None of these definitions are inherently fixed, however, even if we often use them as such and can often assume a common understanding. If we begin to define—in a strict, computational science—poetic content as particular visual signals as opposed to linguistic forms or constructions according to certain rules, then we will include some content that would not heretofore have been included and will exclude other content. Also, we're not talking all visual forms of poetry, but rather those dominant in C19 and early C20 newspapers—not other time periods or periodical forms.

It might be that none of us in this room are uncomfortable with seeing algorithms in this way, but it's important both to understand that in many sectors, algorithms are understood as neutral and also to extend critiques of algorithms also to our work as editors, noting the many places where algorithmic systems and other structures have shaped possibilities for us, often before we even get to our work as editors.

The more important connection I want to highlight is to recent critical work by scholars including Safiya Umoja Noble and Cathy O'Neil, among others, who investigate and analyze the many ways in which algorithms and software encode societal systems, often reinscribing them. In *Algorithms of Oppression* and *Weapons of Math Destruction*, Noble and O'Neil (respectively) demonstrate the ways that algorithms literally and figuratively affect what we see and what we don't see, across myriad aspects of life. They also investigate how people encode into algorithms

racism and sexism, and how these structurally flawed systems and structures then create a feedback loop whereby, for example, structurally racist systems create racist outputs and these racist outputs are then used as additional inputs for the system's new "learning."

To connect these ideas, I want to share about a small-scale set of tests I've done over the last couple of weeks. At the outset, I must say that the sample size from these tests is so small that I cannot draw specific *conclusions*. Even these anecdotal results, however, prompt a series of questions—questions that should remain central, even if a larger set of images ultimately appears to challenge the preliminary results.

In order to explore and further test the generalizability of some of our current approaches, I have been experimenting with giving our system some new images that it hasn't seen before. In one of our approaches, the first step that digital page images go through is to be checked against a series of rules to see if our software can segment a full newspaper page into smaller image snippets. A good snippet looks like a clipping you might cut from a single column of a print newspaper. Currently, our feature extraction and classification system uses these snippets to determine whether poetic content is present or not—and poetic content is not only those full-length poems in the "usual places."

There are a number of known challenges to segmentation in this way—the challenges are one of the reasons we first check page images against a set of rules in the first place. Features such as the skew of a page, low contrast on the image, content spanning multiple columns—whether original to the paper itself or introduced over time through damage, etc.—all create difficulties for our software. Even with these challenges, we've had anywhere from 40% of pages pass the segmentation rules to upwards of 80% of pages pass the segmentation rules.

Having turned again to the *Anglo-African* after a few years away, I decided to test the process on images of the *Anglo-African* that I had digitized for "Will not these days." When I prepared the images to the same size and format as those we've processed to those we've used for processing images from Chronicling America and the Burney Collection, none of the images passed our segmentation rules. The images failed due to image skew and also for the darkness of the images and lack of contrast. Importantly, when I digitized issues of the *Anglo-African* from microfilm, I did very little post-processing on them. I hadn't cleaned or altered the images in a significant way, since we were not planning to OCR them, and while dark, the images are readable to the human eye, in part because we captured them at a high resolution. So, I could immediately begin to imagine some reasons why these page images might look quite different to our segmentation rules than those pages we had built the rules around.

But I also now wondered: how well had our large scale test on 20,000+ page images from Chronicling America done on newspapers from communities of color and/or newspapers in languages other than English? At the time we ran our scale test, we focused on newspapers from the 5-year period 1836-1840, then representative of the first five years of digitized newspapers in Chronicling America. We had our reasons for focusing on that five-year period, but I confess: I did not actually pay much attention to what, or whose, newspapers were represented in that corpus. Likewise, when we conducted our analysis on the test case, "what newspapers" and "whose newspapers" didn't factor into our evaluation. Instead, we focused on material qualities of the newspapers and subsequent reproductions in our analysis. Returning to that test set of newspapers now, what I see is that that corpus is entirely white (not surprising in some ways, given the composition of CA, to return to Benjamin Fagan again), and there are roughly ten newspaper titles in languages other than English, with newspapers in French, German and

Spanish. When I looked at results of our test specifically on these newspapers, only a subset of the German newspaper pages reliably passed our segmentation tests.

Having two examples in mind of periodicals from marginalized or under-represented communities that challenged our segmentation rules, I wanted to explore the function of code a bit further on some others. I went to Chronicling America and downloaded some pages from Black and Spanish-language newspapers that, at a glance, looked comparable to pages from white and English-language newspapers that had passed our segmentation rules. I ran the pages through the process, and in short, none of the pages from Black newspapers or Spanish-language newspapers passed the first steps to advance in our process. That means that none of the pages would—in our current, fully implemented system—get further processing to find their poetic content.

Now, I need to say again that the number of pages I tested is quite small overall—fewer than 30 pages of these other newspapers, in addition to the images of the *Anglo-African* I had tried. It's possible that the pattern—that is, the pattern of our software not passing Spanish-language newspapers, or those created by Black editors and publishers—would not hold up with larger deployment. Nonetheless, this experience made me confront some questions and ideas that are worth exploring, and that my team and I must be sure are at the fore of our minds as we do this work.

These newspapers did not pass the segmentation NOT because of qualities inherent to the original newspapers themselves but rather they did not pass because of choices my team made along the way and because of the post-publication histories of the newspapers' pages.

The key points I want to conclude on, then, for projects such as mine—and in relation to editing as well—are that:

1. Development set has far-reaching consequences.
    a. Despite our quote-unquote good reasons for the at-scale set used and tested, we had major blind spots, particularly given some of the commitments of our work that we brought in.
2. Machine learning learns from your biases and feeds those biases into the system.
    a. Cathy O'Neill's analysis of recidivism prediction and false equivalencies
3. Legacies of prior treatment are amplified in digitization (and reinscribed in the feedback loop of machine learning).
    a. Also considering the systems and structures that led to the way we have them now: we already know that systemic racism, misogyny, etc. affected how/what was deemed worth keeping and therefore what we have access to today. Consider as well how for those materials that did make it into libraries and archives have not been free from these systems.
    b. Papers from minoritized communities may not get the same digital treatment of Chronicling America newspapers, or the other newspapers on which we've built our software—in some cases, these periodicals are "boutique" digitized. Need to be building for this.

Like—or as editing—the creation of algorithms encodes values/beliefs into systems. This work can have us look more closely at the materials, their materiality, and how we have affected them over time and through forms, such as through microphotographic reproductions and digitization. Things we can, as humans, easily account for, in the work that happens between our eyes and our brains—the ways we can make sense of skews to the page, dark pages, or very light pages. When

needing to accommodate those challenges in algorithms, there is an opportunity to look anew at them and consider their origins.