CA Journal of Cultural Analytics

December 18, 2016

Other people's data: Humanities edition

Sarah Allison[a]

[a]*Loyola University New Orleans*

ARTICLE INFO

Article DOI: 10.22148/001c.11822

Journal ISSN: 2371-4549

ABSTRACT

Every project that uses numbers to make sense of literature seems to teach us again that in digital analysis we create more data than we can ever fully use and therefore understand. And yet, with each new project we produce more. In the Community Resource Guide to Digital Humanities Curation, Julia Flanders and Trevor Muñoz define research data as the "raw and abstracted material created as part of research processes and which may be used again as the input to further research." Computational analysis of large corpora is a time-consuming process, and a lot of analysis ends up on the cutting room floor (or on the blog, or in a footnote or an appendix). We need to make better use of that discarded data—the detritus other people shed on the way to an answer. Think of it as data recycling to combat data waste.

Every project that uses numbers to make sense of literature seems to teach us again that in digital analysis we create more data than we can ever fully use and therefore understand. And yet, with each new project we produce more. In the Community Resource Guide to Digital Humanities Curation, Julia Flanders and Trevor Muñoz define research data as the "raw and abstracted material created as part of research processes and which may be used again as the input to further research."[1] Computational analysis of large corpora is a time-consuming process, and a lot of analysis ends up on the cutting room floor (or on the blog, or in a footnote or an appendix). We need to make better use of that discarded data—the detritus other people shed on the way to an answer. Think of it as data recycling to combat data waste.

I'm not saying there aren't good reasons for starting over with every major research project—for hand-picking a corpus and developing an algorithm tailored to a specific question. Throughout academia the premium on scholarly originality is high; we in the humanities haven't had access to big literary corpora for very long, and going over the same ground can certainly suggest a failure of imagination. Moreover, it's often not fully clear what someone else's research process was. But it's worth dealing with these issues if someone else's work produces evidence relevant to a question you want answered.

Here's my case, then, for making more of other people's data: to re-appreciate the value of work that has already been published, and to look at the work of others not only as the basis for creating an analogous project, but as a set of new primary texts that merit further investigation. I lay out two examples: first, how a specific tool might serve as the potential basis of further analysis, and, second, how a set of texts that foiled a predictive tagger constitute a discovery in the context of a more specific field of inquiry. Both studies would require computational literacy and traditional reading methods—and thus suggest new ways to integrate cultural analytics with methods from literary study.

GitHub Heroes

The definition of data offered by the NEH Office of Digital Humanities Management Plans includes software code, algorithms, digital tools, and documentation. Much of the conversation around sharing work has focused on the development of code and corpora—two time-intensive elements of large-scale digital work that lend themselves to being adapted or repurposed for very different kinds of studies. I'm advocating a turn to the *byproducts* of cultural analytics—to more project-specific tools, documentation, and discoveries.

To underscore the difference between transforming key elements of someone else's work and basing a project on what they found, I turn to "The Quiet Transformations of Literary Studies," Andrew Goldstone and Ted Underwood's digital analysis of literary studies journals between 1889-2013, published in 2014 in *NLH*.[2] It used a shared corpus (they worked from JSTOR's Data for Research, a project designed to make JSTOR's metadata and texts available to researchers for largescale analysis)[3] and Goldstone subsequently made the code public, which Jonathan Goodwin then adapted to create a HathiTrust topic browser. The third element of "sharing" in this project is less familiar: the site Goldstone set up to allow the rest of us to explore their 150-topic model of JSTOR data, the Topic Model of Literary Studies Journals.

This site is Other People's Data in a form more fixed than corpus or code: it does not let you play with the parameters Goldstone and Underwood set; rather, it lets you see what they did in new ways. You will not be working with a totally different set of texts—as Goodwin did—or running a new set of models that includes all the stopwords they took out—as someone interested in the critical history of a specific novelist or theorist might want to do. You get to do just what they did—or rather, what they *didn't* do, or didn't write about. This approach requires a substantial level of methodological literacy—the site is easy to play with, but it takes some careful reading of the methodological documentation to understand just what you're looking at. But there is much to be gained from further exploration of JSTOR from this new perspective.

One possible use of the tool grows out of the argument of the article itself. When Goldstone and Underwood sketch out the emergence of criticism as a focus of literary scholarship, they draw on the distinction between critic and scholar that Gerald Graff describes in his disciplinary history, *Professing Literature*. Goldstone and Underwood's work on the critic thus also lays the groundwork for a project on the opposing figure of the scholar, naming three topics "that indicate textual scholarship" and the non-English topics that reveal a "polyglot philology of the early century."[4] Thus, their "excess" data might drive a more detailed investigation, structured by the terms of their argument, and yet distinct in its focus. The essay I've just described—an investigation of the scholar figure through the "Quiet Transformations" site—would treat the topic model more like a new manuscript: an unexplored object that deserves attention in its own right. When someone has spent months producing a tool that can make a graph that sheds light on a major conceptual question, the graph can feel like the major deliverable of the project. But we should treat a tool like this as a starting point for future research. I ran the manuscript metaphor by Goldstone, who pointed out that a new manuscript would also demand "source criticism to use it as evidence: is this an authoritative source? what process created it? what is the chain of transmission by which it reaches us?" The literacy bar, as I've suggested, is high: in the spirit of the scholar, it takes real expertise to understand the document you're examining. A topic model of selected criticism is something like an argument and something like an archive. Knitting it into a history as recognizable as Graff's transforms the "data" back into argument, which might be built on or expanded in more traditionally argumentative ways.

Re-analyzing someone else's data—following a thread suggested by their analysis—poses real challenges to the current value placed on originality in academic scholarship. The project I've described would not only develop their work but, with their aid, re-analyze their analysis along very similar lines to those they propose. This is radically new, and very interesting, in part because of the clear benefit (there is much more to learn from the tool they created) and in part because of the equally clear issues with respect to the academic privileging of originality. Its contribution would rest, not on the new worlds it opens up, but on its exploration of the world opened up by Goldstone and Underwood and what it revealed about the shape of disciplinary history. The outputs of computational models consitute new primary sites of textual analysis.

Roads Not Taken

If the Topic Model Browser is a new kind of resource for future work, I also want to consider picking up a road not taken in someone else's research. The idea of *taking* someone else's road suggests one threat of using other people's data—that it's really just stealing someone else's work. Humbly, then: there is a difference between misrepresenting someone else's work as your own and reanalyzing what you have frankly acknowledged is theirs. The metaphors of "picking up the trail" or "carrying the torch" elide what is so generative about working with someone else's abandoned graph: the discontinuity between the researchers' original goals and your own.

For example:
In their work on literary standards, Ted Underwood and Jordan Sellers used the binary decision of whether or not a work was reviewed in a set of historical journals as a proxy for literary prestige.[5] To train their tagger, they used the relative occurrence of frequently-used words. The kinds of words used most often by a text are actually a reasonably strong signal of an individual author, and even of register—speech and writing, for example, are characterized by very different elements of the set of words we use all the time. In a footnote and the methodological appendix, the authors explain their decision to exclude the radical *Tait's Magazine* from the training set on the grounds that its exclusion increased the accuracy of the model by almost five percent: *Tait's* choices looked different enough from the texts chosen by other periodicals to significantly influence the model.

As a Victorianist, I have to know: What was wrong with *Tait's*? The authors point out that, because it was a monthly, "a relatively large number of titles" would have been reviewed—"perhaps," they add, "a little indiscriminately."[6] What they had discovered was that the texts reviewed by *Tait's* were stylistically distinct from those reviewed by other journals across the nineteenth century. This is very interesting because *Tait's* reviewed poetry from an explicitly radical perspective. Odile Boucher Rivalain considers the confluence of poetic and radical motivations for reviewing poetry in *Tait's*, noting that it "made itself the champion of the Radical poets of the 1830's."[7] It's tempting to take another computational pass—to look specifically at the texts reviewed by *Tait's* in their corpus—but it's also important to recognize that another computational pass gets you a list of words that, one suspects, would echo the more general patterns the authors describe. A more illuminating way to "use" this data would be to set the parameters of a reading survey according to Rivalain's suggestive 1997 connection between politics and poetry that compares *Tait's* with the *Westminster Review*—two periodicals designed "to diffuse Radical ideas without alienating the Whigs in the struggle against the aristocracy."[8] Moreover, following Rivalain's account of the founding of *Tait's* to "counter the influence of the *Edinburgh Review* and *Blackwood's Magazine*," one could throw those two major partisan quarterlies into the mix. Despite their political differences from one another and from the Radical periodicals, the books they review seem to have more in common, stylistically speaking, with the *Westminster* than with *Tait's*.[9] Which of the works reviewed by *Tait's* were also reviewed by the other cohort? A set of samples (of poetry books reviewed by *Tait's*) that proved misleading in the larger scope of the nineteenth-century might be very instructive with respect to a specific part of it—when viewed with a human eye for patterns to supplement the stylometric perspective of the predictive tagger. The next step on "the road not taken" is thus a step back toward more established methods of literary criticism.

I've argued that it's only right to make the most of information that costs so much to produce, but let me acknowledge that what I propose is no shortcut. Drawing on other people's analysis to make a more specific argument, whether about twentieth-century critical trends or the style of nineteenth-century radical poetry about the Corn Laws, does not save work. In the first case, I'd need to pay the same careful attention to "scholar-words" that Goldstone and Underwood did to "critic-words." In the second, the study of volumes reviewed by nineteenth-century periodicals looks very much like a conventional periodical study, except that the foundational insight that frames it is drawn from a stylistic observation of great scope.

We know that many disciplines have protocols for working with studies conducted by other researchers, but, in the humanities, such work represents a new frontier. A footnote to the directions for Data Management Plans from the

NEH Office of Digital Humanities thanks "the National Science Foundation's Directorate for Social, Behavioral, and Economic Sciences for allowing us to use much of the language from its data management plan guidance."[10] Not that it's so easy in other fields, of course, and for some of the same reasons (in biology, see this and this). And as other fields have begun to make data manipulation more transparent, more instructions have emerged about leaving a clean path behind you.[11] There is something obvious about my argument that we should use other people's data more: ideally, all our work builds on the work of others. And there is something counterintuitive about it, too. The emphasis on "replication" of data in the humanities so far has largely been through projects that break new ground in order to investigate a classic non-digital argument. It's time to reconsider what it means to build on other people's work.

Notes

[1] "DH Curation Guide: A Community Resource Guide to Data Curation in the Digital Humanities," Accessed November 30, 2016.

[2] Andrew Goldstone and Ted Underwood, "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us," *New Literary History* 45, no. 3 (2014): 359-384.

[3] "Data for Research," *About JSTOR*. Accessed November 30, 2016.

[4] Andrew Goldstone and Ted Underwood, "The Quiet Transformations of Literary Studies," 22-23.

[5] "The Longue Durée of Literary Prestige," *Modern Language Quarterly* 77:3 (September 2016): 321-44. Here, I will focus on a discussion of procedure that was cut between the preprint circulated online in May 2015 and the appearance of the article in print in *MLQ* in September 2016, so please see also Ted Underwood and Jordan Sellers, "How Quickly Do Literary Standards Change?," 2015.

[6] Underwood and Sellers, "How Quickly Do Literary Standards Change?," 32.

[7] Odile Boucher Rivalain, "'Bringing out the Sympathies of Mankind': Reviewing Radical Poetry in 'Tait's Edinburgh Magazine' and 'The Westminster Review' in the 1830s and 1840s," *Victorian Periodicals Review* 30, no. 4 (1997): 350-367, 353.

[8] Rivalain, "'Bringing out the Sympathies of Mankind,'" 351.

[9] Rivalain, "'Bringing out the Sympathies of Mankind,'" 350.

[10] "Data Management Plans for NEH Office of Digital Humanities Proposals and Awards," National Endowment for the Humanities, 2015.

[11] My thanks to Inger Bergom, Senior Researcher, Institute for Democracy and Higher Education, Tufts University, for her perspective on this problem.