CA Journal of Cultural Analytics

February 27, 2017

From Reproducible to Productive

Andrew Goldstone[a]

[a]*Rutgers University, New Brunswick*

ARTICLE INFO

Article DOI: 10.22148/001c.11821

Journal ISSN: 2371-4549

ABSTRACT

The very idea of a "canonical data set" implies a whole organization of knowledge: first, the data are durably available—a quarter-century on—thanks among other things to the institutional continuity of the GSS as an important large-scale data-collection enterprise of American social science; second, the data remain meaningful, their validity underwritten by the methods of survey research; third, the disciplinary norms of sociology allow for the possibility of following on someone else's work by reusing the evidence they have already selected; fourth, that evidence can still bear on a significant research question within sociology, a testament to the fruitfulness of the research program in cultural taste and social structure which was set in motion, notably, by the Anglophone reception of Pierre Bourdieu's *Distinction*. Lizardo and Skiles's starting point, in other words, includes not simply the dataset itself but all the institutional conditions for a productive ongoing research program involving quantitative analysis of cultural data.

I recently read a paper in the sociology of culture, Lizardo and Skiles's "The End of Symbolic Exclusion?," which refers to a particular collection of 1993 General Social Survey responses on musical taste as "a 'canonical' data set providing the empirical basis for a variety of analyses (and re-analyses)" (88).[1] The very idea of a "canonical data set" implies a whole organization of knowledge: first, the data are durably available—a quarter-century on—thanks among other things to the institutional continuity of the GSS as an important large-scale data-collection enterprise of American social science; second, the data remain meaningful, their validity underwritten by the methods of survey research; third, the disciplinary norms of sociology allow for the possibility of following on someone else's work[2] by reusing the evidence they have already selected; fourth, that evidence can still bear on a significant research question within sociology, a testament to the fruitfulness of the research program in cultural taste and social structure which was set in motion, notably, by the Anglophone reception of Pierre Bourdieu's *Distinction*.[3] Lizardo and Skiles's starting point, in other words, includes not simply the dataset itself but all the institutional conditions for a productive ongoing research program involving quantitative analysis of cultural data.

I am not sure that similar conditions currently obtain for quantitative studies of culture based in the humanities disciplines, but Sarah Allison's "Other People's Data" points the way. Her essay describes a vision of cumulative research which I hope to see realized. Of course, it is particularly gratifying that the research she envisions building upon includes my own. But the real point, for Allison, is not that Ted Underwood and I said something convincing but that we—like Underwood and Jordan Sellers in their collaboration—*produced reusable evidence*.[4] She provokes us to rethink the conditions in which such reuse could be possible for individual researchers and valued in our disciplines.

There are multiple ways of understanding what the evidence *is* that Allison wants us to reconsider for other arguments. I will focus on the discussion of the history of literary studies in Underwood's and my essay "The Quiet Transformations of Literary Studies." In this case, we can conceptualize the evidence as going through a series of successive transformations:

1. the selected original documents;
2. the corpus of digital text created by OCR, with metadata;
3. the vector-space representation of that corpus: (a) as supplied by JSTOR; (b) modified by our stoplisting and orthographic normalization;
4. the probabilistic topic model of those feature vectors;
5. the interactive visualization of that model and the metadata.

As Allison points out, though there has been plenty of digital-humanist discussion about creating corpora, there has been much less attention to the latent possibilities of what she calls the "*byproducts* of cultural analytics" but which, in this case, we can describe simply as successive transformations of the data. The question raised by "Other People's Data" is: *Which* transformations facilitate further work?

As the transformations of data analysis proceed, each stage is more directed to answering the particular questions the analysts are asking. The transformation to feature vectors was already driven by Underwood's and my goal of discerning long-term thematic or conceptual trends: we developed the list of stopwords iteratively until we were satisfied that many of our topics looked, based on their most heavily-weighted words and documents, both sufficiently general and interestingly meaningful. Similarly, the choice of modeling parameters—above all the number of

topics—depended implicitly on the level of generality at which we wanted to operate. Finally, and *a fortiori*, the Quiet Transformations visualization is oriented towards our interpretive choice of talking about transformations of the quiet kind. The emphasis on visualizing topic proportions over time (by showing series of topic weights in each publication year) is an obvious consequence of this choice.

But I now realize that what seemed like quite straightforward choices were shaped by Underwood's and my interpretive aims too. The time-series plots always have the same *x* axis, ranging from 1889 to 2013; by contrast, the *y* axis limits are always scaled to the peak topic proportion, even though some topics' peaks are much higher than others'. These choices helped us to pick out individual topics whose century-long time trends are historically interesting. But there are other possibilities that are made more difficult: comparisons between topics, for example, or comparisons between particular scholarly journals, both of which could be made on the basis of the topic model and the document metadata but are not facilitated by the website.

Thus, Allison's dictum that embarking on an analysis of the Quiet Transformations website means "you get to do just what they did" cuts both ways: you inherit the limitations as well as the potential for further interpretations. Allison acknowledges what she calls the "literacy bar" for using the information we present for further work, but some responsibility rests with the authors of the transformed data—the responsibility to present the data with enough context that the next person to use that data can do so in full awareness of the choices that have been made for them. In the case of Underwood's and my work, our attempts to fulfill that responsibility are spread across several media: some of the detail about the data and our processing and modeling choices is in our essay and its appendix, and the rest remains implicit in the R and Python scripts we made available in a github repository.[5]

By using the vague term "transformation" I have been postponing a question which I now wish to take up: just what counts as data suitable for re-analysis? For Allison, "data" would seem to include, not just the initial texts under study, but the statistical model used to draw inferences in the course of Underwood's and my argument, and even the interactive graphs and tables we presented for discussion. She suggests that "a topic model of selected criticism is something like an argument and something like an archive," with the implication that it can be a "new primary site[] of textual

analysis." I certainly hoped that the Quiet Transformations website would not be mistaken for the *endpoint* of an argument about the history of literary studies (or, in Allison's humorous phrase, "the major deliverable of the project")—that it could instead generate more hypotheses of the kind she proposes in her brief discussion of the figure of the "scholar" in the early twentieth century.

Still, it is worth pausing before collapsing a statistical *model* into the category of "data." Topic models are a particularly tricky case for this distinction. Generating a single 150-topic model of 21,000-odd documents over a 100,000-term vocabulary, as we did, yields an intimidating number of parameter estimates: the model can be characterized by two matrices, the 150 x 21,000 topic-document matrix and the 150 x 100,000 topic-term matrix. In fact some of our analyses go even further and consider the final state of MALLET's Gibbs sampler, with its assignment of a topic to every *token* in the corpus. With this many numbers, understanding the model output becomes a problem in data analysis itself. Furthermore, one way of linking topic modeling approaches to classic social-science methodology is to describe it as a form of automated content analysis; in a manual content analysis, the labels applied to texts by coders, rather than the texts, would normally become the data.[6]

But a probabilistic topic model ought normally to be understood like any other statistical model, as a selective picture of data rather than primary data itself. There is nothing about Underwood's and my particular decisions that ensures their validity for all future applications; those choices went only as far as we needed for the particular arguments we wanted to make. Or, to put it another way, Underwood's and my modeling process includes all the other parameter settings we tried, not just the output we consider at length in our essay. This full process, with all of its degrees of freedom for the researcher, would need to be recapitulated: researchers who wanted to use our data should consider alternate modeling possibilities over our chosen corpus—if they are even willing to use our corpus as it stood.[7] As Allison emphasizes, other people's data need the same critical scrutiny as one's own data; that scrutiny seems more useful to me than attempting to draw more inferences from statistical models or other data summaries found in the research literature—especially in these very early days for our particular field, when all results are highly preliminary. But it is not too much to ask, *above all* in early days, that we choose data—objects of study—that are worth returning to in the way that Allison urges.

What must research be like, if the "data recycling" Allison calls for is to take place? Her argument converges with, and intervenes significantly in, discussions about so-called "reproducible research," as she hints in her final paragraph.[8] The ideal of reproducible research is that the whole process of data transformations, from the original source data to the numerical and visual outputs used in a final scholarly argument, should be accessible to others, who should be able to recapitulate it for themselves. But the apparatus of "reproduction" in this sense can also be used, Allison is suggesting, for *producing more research*. Indeed, a conversation about the conditions of what we could call *productive* (rather than "reproducible") research could advance beyond the sometimes tiresome focus, within discussions of reproducible research, on software engineering and bureaucratic "best practices."

At the same time, Allison also teaches us that the category of productive research need not be restricted to "code and corpora." The focused, purpose-designed outputs of specific research may still, she argues, be productive of further work. We can go further: they will often be *more* productive than the enterprise of building all-purpose software "tools" without any particular research end in view. We might even ask ourselves whether more specific, partial, even eccentric datasets designed with more pointed questions in mind might not prove to be more useful objects for programs of research than generalized "representative" corpora, elaborately encoded digital editions, or big digital libraries taken in bulk. As a side benefit, particularized datasets may be easier to put into circulation than totalizing text corpora; the latter tend to be too big for anyone without a server farm, even when they aren't captured by corporate data vendors. The possibilities of the former are already suggested by the contents of *CA*'s Dataverse archive of supporting data for articles, a most commendable effort at opening up research data. It is not simply, then, a question of relaxing our commitment to the appearance of "originality" in order, as Allison suggests, to "combat data waste." It might be a way for there to be *lines of research* in quantitative studies of literature and culture that extend beyond a single person or a single institution. It might be a way to distribute some of the enormous labor of making useful evidence—Allison's recognition of this labor is quite gratifying—across time and across the scholarly universe.

And it might also, finally, be a way to *prove someone wrong*. Allison is kind to the two essays she discusses: she raises issues for further investigation that would extend

their results rather than challenge them. But her proposal of "re-analyzing someone else's data—following a thread suggested by their analysis" unmistakably also implies the possibility of finding out that the original argument was not well-supported by the data or by the original analysis. This somewhat anxiety-inducing possibility nonetheless seems to me a condition for the increase of knowledge. Either quantitative studies of culture will make claims that can be defeated by evidence, or they will devolve into games with computers. If we prefer the former option, we should take heed of Sarah Allison's arguments in "Other People's Data."

Notes

1. Omar Lizardo and Sara Skiles, "The End of Symbolic Exclusion? The Rise of 'Categorical Tolerance' in the Musical Tastes of Americans: 1993-2012," *Sociological Science* 3 (2016): 85-108, doi:10.15195/v3.a5. The article goes on to join new survey data to the 1993 survey results.
2. In this case, a much-cited essay by Bethany Bryson: "'Anything But Heavy Metal': Symbolic Exclusion and Musical Dislikes," *American Sociological Review* 61, no. 5 (October 1996): 884-99, doi:10.2307/2096459.
3. Bourdieu, *Distinction: A Social Critique of the Judgement of Taste*, trans. Richard Nice (Cambridge: Harvard University Press, 1984). This line of work has already appeared in *CA*: Dan Jurafsky et al., "Linguistic Markers of Status in Food Culture: Bourdieu's *Distinction* in a Menu Corpus," *Cultural Analytics* (October 2016).
4. Goldstone and Underwood, "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us," *New Literary History* 45, no. 3 (Summer 2014): 359-84; Underwood and Sellers, "The Longue DurÃ©e of Literary Prestige," *MLQ* 77, no. 3 (September 2016): 321-44.
5. Though I share Allison's desire to shift scholarly attention away from "code" towards evidence, I do wish to mention that I subsequently developed a more adaptable version of the scripts in that repository in my dfrtopics R package, which can also generate an interactive visualization like the website Allison discusses.
6. Justin Grimmer and Brandon M. Stewart helpfully situate topic models among other forms of computer-assisted content analysis in a methodological essay: "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* 21, no. 3 (Summer 2013): 267-97, doi:10.1093/pan/mps028.
7. I should record here a warning against using just our corpus as it stands. This is possible in principle, since our feature vectors are recoverable from the model outputs we made available. But after we completed work on the article, JSTOR revised all of its exported document feature counts, as their FAQ page notes. (In fact I discovered the text-processing error mentioned there while working on another project; I discuss this in more detail in an essay in progress.) Anyone who *really* wants to revisit the questions of "The Quiet Transformations of Literary Studies" should re-request an analogous set of articles from JSTOR's Data for Research service.
8. On the general subject, see, for example, Victoria Stodden et al., "Enhancing Reproducibility for Computational Methods," *Science* 354, no. 6317 (December 9, 2016): 1240-41, doi:10.1126/science.aah6168, and Greg Wilson et al., "Best Practices for Scientific Computing," *PLoS Biology* 12, no. 1 (2014), doi:10.1371/journal.pbio.1001745.