"Tell Me Something I Don't Know!": The Place and Politics of Digital Methods in the (Islamicate) Humanities

MATTHEW THOMAS MILLER AND SARAH BOWEN SAVANT
Roshan Institute for Persian Studies, University of Maryland, College Park, Md.; e-mail: mtmiller@umd.edu; Aga Khan University, Institute for the Study of Muslim Civilizations, London; e-mail: sarah.savant@aku.edu
doi:10.1017/S0020743817001027

Debates about the value of digital methods often return to the nature of knowledge itself. Specifically, do not digital methods tell us what we intuitively already know? Or, if we do not know something yet, is it trivial or discoverable through other more traditional humanistic modes of analysis?

Similarly, there is a deep-seated suspicion that scholars working with digital methods think their methods *always* produce superior knowledge and that this knowledge is purely empirical. But digital work at its best is no call for a scientific positivism or uncritical empiricism—much less a "big data" revolution in the humanities. Rather, it is far more modest in its claims and is deeply rooted in traditional humanistic methods, including those associated with the various schools of thought that go under the banners of cultural studies and critical theory. It is built upon the ability of a computer to discern patterns, but it requires scholars to engage in interpretative acts to make sense of them.

For Middle Eastern/Islamic studies, digital methods offer a particularly promising way to study its diverse bodies of textual sources, which are among the largest, most geographically diffuse traditions in world history. These traditions are literally "too big to know" through traditional humanistic modes of enquiry based on close reading, and an exclusive dependency on these methods can result in certain macrolevel blind spots. Distant reading methods such as text reuse[2] detection, stylometry, and topic modeling can address this problem by allowing us to read the unreadable, discovering patterns at levels of scale beyond human capabilities, and thereby helping us make more well-informed choices about what we read closely. How, for example, do conventional genre classifications hold up when digital methods are added to reading practices? How widely is a text reused, and in which ways? Does it share stylistic or other similarities with other texts, and if so, which ones? Did its author rely on previous materials, and if so, from what pools? And how did he alter them?

To illustrate these points, let's consider a case involving text reuse detection methods for the study of citation practices in 10th-century Baghdad. We can detect the extent and precision of an author's reuse of previous materials and do this independently of what authors and transmitters say they are doing. Among other matters, this helps us to understand what they thought required attribution versus what they thought could be taken freely, and likewise, what they thought could be remolded versus what they thought had to be passed on verbatim. The issues involved might be familiar to us today, but our early Arabic authors almost never tell us precisely what they are doing or how they are doing it. Their modern counterparts, seeking to understand their assumptions, typically generalize from cases that are picked because they offer the clearest evidence for modes of transmission. But this is an instance of the so-called *streetlight*

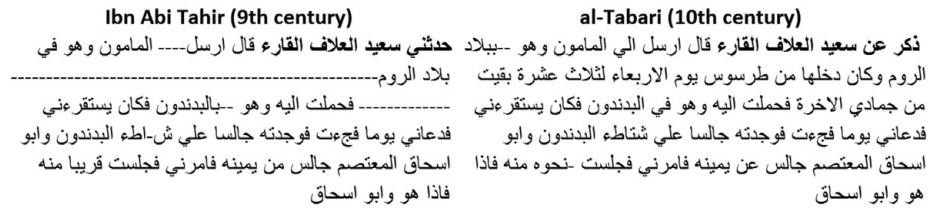

FIGURE 1. Example of reuse without naming a source. Ibn Abi Tahir says: "The Qur'an reciter Saʿid al-ʿAllaf reported to me that . . ." Al-Tabari says: "It has been related from the Qur'an reciter Saʿid al-ʿAllaf that. . . ."

effect—that is, looking for the keys you lost in a dark street only where the light is because that is where it is easiest to find them.[3] How much better to be able to see patterns across all works, before choosing instances that can be investigated as representative of something.

We are just beginning to analyze such patterns across the whole tradition by, for example, measuring the total amount of reuse in the tradition and where it occurs most. Take the case of the historian Muhammad ibn Jarir al-Tabari (d. 923). It is clear already that he represents a heavy reuser and was heavily reused. Here, the reader might be saying: "Tell me something I don't know!" But stay with us for a moment. Although al-Tabari is generally regarded as having raised the standards on citation practices, our distant reading turns up something quite interesting. Al-Tabari is very selective in what he cites and how he cites it. The key here is the unit of citation, and specifically, the individual report (which he cites) versus the sources from which he derived those reports (which are harder to discern). A most interesting example involves a book of his immediate predecessor Ibn Abi Tahir Tayfur (d. 893), which he uses without giving credit. Franz Rosenthal, in his article on Ibn Abi Tahir,[4] spoke of his mostly lost *Kitab Baghdad* (Book on Baghdad), noting that "Ibn Abī Ṭāhir's treatment agrees widely with that of the later Ṭabarī." It certainly does. Only one of the six original volumes of Ibn Abi Tahir's book survives, this volume pertaining to the reign of al-Maʾmun. It has already been digitized so we were able to use a text reuse algorithm designed by David Smith of Northeastern University, called passim, to compare this volume to al-Tabari's *Taʾrikh* (History). Passim indicated that 24 percent of Ibn Abi Tahir's book is adopted in al-Tabari's (11,700 words out of 48,222—in our one hundred–word chunking).[5] This reuse runs from the year 204 AH (819 CE), when al-Maʾmun entered Baghdad, until his death, an account of which al-Tabari narrates on the authority of an eye witness, the Qur'an reciter Saʿid al-ʿAllaf, without mentioning Ibn Abi Tahir. Smith's software identifies and aligns common passages and generates data, including relating to the precision of reuse (Figure 1).

Reading many such passages side by side, it becomes clear that al-Tabari is reusing Ibn Abi Tahir's text. Interestingly, he is citing and crediting Ibn Abi Tahir's sources, but not Ibn Abi Tahir himself. A close look at the instances of reuse shows just how precisely he uses these materials—and how he often discards Ibn Abi Tahir's own additions and adds his own (where the dashes indicate a gap in the alignment). Combined with other similar instances in other books, this suggests that up to a certain point in the

history of the written tradition, authors such as al-Tabari understood the unit for citation as portions of the book rather than the book itself. The latter was not a category with as much epistemological weight as we might expect. This is a theory worth considering, and it is one suggested by many other instances of such data (another key example involves the biography of the Prophet Muhammad by Ibn Ishaq).[6]

Regardless of one's assessment of the merits of digital humanities (DH) methods—such as the preceding text reuse example—some humanists may still prefer to keep DH at an arm's length due to its purported nefarious origins, both political and intellectual. At almost any DH panel at the annual conference of one of the major humanities organizations, you are likely to hear some audience member express some version or combination of the following two sentiments: "But isn't DH just a Trojan Horse for an assault by the sciences/tech sector on the humanities?" And/or: "I have heard that DH is the humanities vanguard in the neoliberalization of the university!"[7] Both of these accusations are scary, but also problematic.

Critics typically seek to frame DH as intellectually descendent from the "Big Data" wave in the sciences—a sort of scientific or technocratic beachhead on the already threatened shores of humanities departments. Although the threats to humanities departments are very real, the characterization of DH as some sort of belligerent foreign invader is misconceived. Most DH scholarship falls into one of two categories, both of which have long humanistic genealogies. The first group typically employs computational methods to study eminently traditional issues in the humanities, such as genre classification, authorial attribution, stylistics, and intertextuality. These scholars employ new digital methods that allow them to study their sources at a scale and level of specificity that would be exceedingly difficult if not impossible for an individual researcher to do without the aid of a computer. But these are hardly new research questions whose posing somehow aids the Big Data scientists' colonization of the humanities.

The intellectual genealogy of the second group of DH practitioners reveals—perhaps even more than the first—the shaky basis of the "DH = Big Data Invader" argument. This second group comprises a wide range of recent scholars and studies largely inspired by the more empirically inclined, so-called "distant reading" approaches to studying history and culture. The term itself was coined by Franco Moretti in 2000 and a significant amount of the pioneering early work in this field was done by him, Matthew Jockers, and their graduate student collaborators in the Stanford Literary Lab. As Ted Underwood has recently argued, distant reading as an intellectual project emerges out of the much older sociology of literature and book history research projects, seeking to address with new digital tools questions substantially similar to those that Raymond Williams and Janice Radway employed in their predigital studies.[8] Far from conservative in orientation, the approach of these scholars locates them within the Marxist and feminist camps of literary studies. Moretti, for his part, explicitly positions his project as building on Marxist literary theory, the Annales School, and World Systems Analysis. Again, as we have seen repeatedly here, DH relies upon and strengthens—even reinvigorates—existing humanities modes of analysis, and certainly not ones that we would think of as reactionary.

With its genealogy firmly rooted in a combination of traditional literary studies and feminist and neo-Marxist sociological approaches to literature and history, the assertion that DH and its practitioners are essentially some sort of "neoliberal tool" appears

much more problematic, to say the least. The features of DH work that opponents frequently point to as evidence of its purported neoliberal agenda (e.g., reliance on poorly paid contingent and graduate student labor, promotion of alt-ac, data-driven analytics) are hardly the exclusive domain of DH; they are the dominant trends in the contemporary university more broadly and have a longer and much more complex history. DH as a field does reflect these very unfortunate realities of the contemporary neoliberal university to some degree, but it is no more responsible for them than all of the humanities departments that have relied on poorly paid contingent and graduate student labor to staff their classes—a practice which has existed since well before the advent of the contemporary DH wave.

This is not to say that there are not legitimate critiques of DH. We completely agree, for example, with the criticisms of the *many DH scholars* who themselves have criticized the field for its lack of linguistic, gender, and cultural diversity and its resistance—*in some quarters*—to critical theory and cultural studies.[9] But these specific criticisms—unlike the abstract critiques of DH—evince a deep understanding of the diversity, methods, and nuance of actual DH work. To conclude, we would suggest, as Matthew Kirschenbaum has urged us, to stop talking about the "construct" of "Digital Humanities"—that favorite specter haunting all of the humanities—and evaluate "actually existing" DH work on its relative merits or demerits.[10]

NOTES

[1]In alphabetical order.

[2]"Reuse" is a technical term preferred by computer scientists. It refers to common passages between texts, whatever their origins, and is meant to be value neutral and encompass different text reuse practices, some of small scale, some of large; some involving precision, some using paraphrase; some providing acknowledgment to their sources, some not.

[3]As David H. Freedman recounts: "Late at night, a police officer finds a drunk man crawling around on his hands and knees under a streetlight. The drunk man tells the officer he's looking for his wallet. When the officer asks if he's sure this is where he dropped the wallet, the man replies that he thinks he more likely dropped it across the street. Then why are you looking over here? the befuddled officer asks. Because the light's better here, explains the drunk man." As Freedman surmises: "Many, and possibly most, scientists spend their careers looking for answers where the light is better rather than where the truth is more likely to lie." Freedman, "Why Scientific Studies Are So Often Wrong: The Streetlight Effect," *Discover Magazine*, 10 December 2010, accessed 20 September 2017, http://discovermagazine.com/2010/jul-aug/29-why-scientific-studies-often-wrong-streetlight-effect.

[4]F. Rosenthal, "Ibn Abi Tahir Tayfur," in *Encyclopaedia of Islam, Second Edition*, ed. P. Bearman, Th. Bianquis, C. E. Bosworth, E. van Donzel, and W. P. Heinrichs, accessed 11 October 2017, http://referenceworks.brillonline.com.iij.idm.oclc.org/entries/encyclopaedia-of-islam-2/ibn-abi-tahir-tayfur-SIM_3056?s.num=1&s.q=ibn+abi+tahir. For a discussion of this reuse, see also C. E. Bosworth's introduction to his translation of volume 32 of al-Tabari's *History*; *The History of al-Tabari* (Albany, N.Y.: State University of New York Press, 1987), 3–8; and (more detailed) Hans Keller's introduction to his German translation of the *Kitab Baghdad* (*Sechster Band des Kitab Bagdad* [Leipzig: O. Harrassowitz, 1908], 1:XIII–XXVI).

[5]This is based on segmentation of the texts into one hundred–word chunks. Smith's software, among others, allows different parameters to be set; different parameters or chunking will almost certainly reveal further reuse. For the texts, see https://github.com/OpenArabic, and, therein, the files 0280IbnTayfur.Baghdad.Shamela0005880-ara1 and 0310Tabari.Tarikh.JK000157-ara1.

[6]Sarah Bowen Savant explores this theory in her next monograph, *A Cultural History of the Arabic Book* (Leiden: Brill, forthcoming).

[7]A Representative summary of this view can be found in Daniel Allington, Sarah Brouillette, and David Golumbia, "Neoliberal Tools (and Archives): A Political History of Digital Humanities," *Los Angeles Review of Books*, 1 May 2016, accessed 20 September 2017, https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/#!; and Richard Grusin, "The Dark Side of Digital Humanities: Dispatches from Two Recent MLA Conventions," *Differences* 25 (2014): 79–92.

[8]Ted Underwood, "A Genealogy of Distant Reading," *DHQ* 11 (2017): http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html, accessed 15 September 2017.

[9]See, for example, the work of Tara McPherson, Amy E. Earhart, Martha Nell Smith, Alan Liu, Domenico Fiormonte, Roopika Risam, Moya Bailey, Anne Cong-Huyen, Alexis Lothian, and Amanda Phillips (the latter four representing the #TransformDH group).

[10]Matthew Kirschenbaum, "What Is 'Digital Humanities,' and Why Are they Saying Such Terrible Things about It?," *Differences* 25 (2014): 60–61.