Measuring the Uneven Digitization of Historical Literature

Evalyn, Lawrence Isaac
lawrenceevalyn@gmail.com
Northeastern University, United States of America

As literary scholarship is increasingly mediated through digital archives, it becomes important to know what subset of historical texts are in fact available digitally. Uneven digitization can impact not just distant reading research, which explicitly uses large collections of digital texts to draw its conclusions, but also the ordinary exploratory research of scholars whose first contact with historic texts is increasingly mediated through digital archives. In this paper, I present a case study of uneven digitization, measuring the availability of late eighteenth-century women's writing to shed light on broader processes of digital selection.

At the core of this paper is a comparison between four different resources: the English Short Title Catalogue (ESTC), Eighteenth Century Collections Online (ECCO), the Text Creation Partnership (TCP), and HathiTrust. For each of these databases, I collect bibliographic metadata for their holdings of works published in England between 1789 and 1799. These were years of intense and contested print publication in England, during which the eighteenth century public grappled with questions of literary value and longevity which continue to play an important role today. For this eleven-year "decade," the ESTC lists 51,090 titles in its bibliographic reference database. ECCO, a collections of full-book facsimiles of full books that used the ESTC as a reference list, has 26,848 works— barely more than half the records listed in the ESTC. HathiTrust, a similar collection of facsimiles, has 8,220 works. The smallest text repository, the ECCO-TCP collection of hand-encoded transcriptions made from ECCO facsimiles, contains only 525 records: an almost negligible 1% of the ESTC. Building on feminist scholarship which has shown the marginalization of women's writing in other forms of scholarly infrastructure, I ask: *which* half of the ESTC is in ECCO? Which 1% is in the ECCO-TCP? Specifically: has women's writing been overlooked by digitization?

To find out, I manually identified an implied authorial gender for all the works in my four bibliographic corpora described above. I expected that each smaller corpus would be more strongly influenced by opportunities for selection bias, and thus contain fewer works by women. However, this is not what I find. As shown in Figure 1, male authors grow from being responsible for 49% of the records in the ESTC to 76% of the records in ECCO-TCP. However, this rise in male authors is accompanied by a parallel rise in *female* authors, from 3% in the ESTC to 22% of ECCO-TCP. Even at 22% of the ECCO-TCP, these are far fewer women than are measured in bibliographies of novels and poetry. Nonetheless, I contend that the low numbers of women's works are not due to digitization bias at the site of gender; instead, what mass digitization shows is the preponderance of publications that are not novels or poetry. Where uneven digitization occurs, it is "authorless" writing which fails to garner institutional investment. This paper concludes with the new direction suggested by my findings: the role of unsigned literature in the eighteenth century. Ultimately, I argue, an infrastructural emphasis on a singular "author" reinforces an unhelpfully individualist interpretation of cultural production.

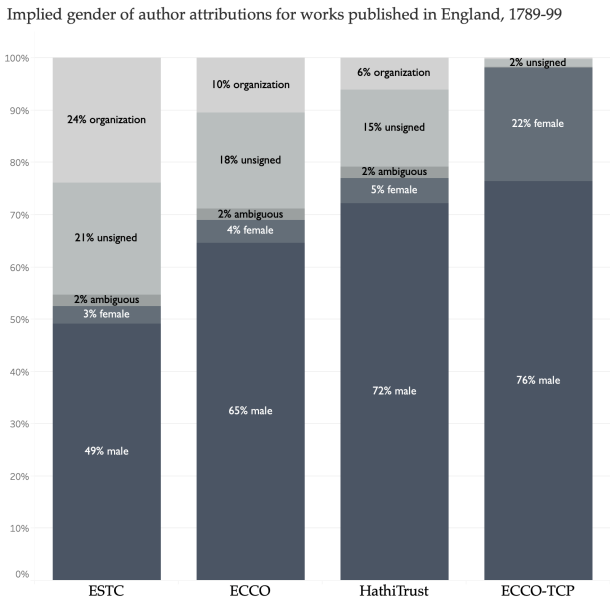

Figure 1: Author demographics in the ESTC, ECCO, HathiTrust, and ECCO-TCP, as a percentage of each resources' records of works published in England 1789-99.

Bibliography

Bode, Katherine (2018): *A World of Fiction: Digital Collections and the Future of Literary History*. Ann Arbor: University of Michigan Press.

Centivany, Alissa (2017): "The Dark History of HathiTrust", in: *Proceedings of the 50th Annual Hawaii International Conference on Systems Science*

D'Ignazio, Catherine / Klein, Lauren F. (2020): *Data Feminism*. Cambridge, MA: MIT Press.

Feldman, Paula R (2002): "Women Poets and Anonymity in the Romantic Era.", in: Clery, E. J. / Franklin, Caroline / Garside, Peter (eds.): *Authorship, Commerce and the Public: Scenes of Writing 1750-1850*. London: Palgrave Macmillan.

Gregg, Stephen H. (2020): *Old Books and Digital Publishing: Eighteenth-Century Collections Online*. Cambridge: Cambridge University Press.

Griffin, Robert J. (1999): "Anonymity and Authorship", in: *New Literary History*, 30, 4: 877–95.

Griffin, Robert J. (ed.)(2003): *The Faces of Anonymity: Anonymous and Pseudonymous Publication from the Sixteenth to the Twentieth Century*. London: Palgrave MacMillan.

Hauswedell, Tessa / Nyhan, Julianne / Beals, M. H. / Terras, Melissa / Bell, Emily (2020): "Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers", in: *Archival Science* 20: 139–165. DOI:10.1007/s10502-020-09332-1

Levy, Michelle / Perry, Mark (2015): "Distantly Reading the Romantic Canon: Quanitfying Gender in Current Anthologies", in: *Women's Writing*, 22, 2: 132-155. DOI: 10.1080/09699082.2015.1011836.

Mak, Bonnie (2014): "Archaeology of a Digitization", in: *Journal of the American Society for Information Science and Technology* 65, 8: 1515–1526, DOI: 10.1002/asi.23061.

Murphy, Kathleen S. (2011): "Translating the vernacular: Indigenous and African knowledge in the eighteenth-century British Atlantic", in: *Atlantic Studies* 8, 1: 29-48. DOI: 10.1080/14788810.2011.541188.

Riddell, Allen / **Bassett, Troy J.** (2020): "What Library Digitization Leaves Out: Predicting the Availability of Digital Surrogates of English Novels." arXiv:2009.00513v1 [cs.DL].

Singh, Amardeep (2019): "Beyond the Archive Gap: The Kiplings and the Famines of British Colonial India", in: *South Asian Review* 40, 3: 237-251. DOI: 10.1080/02759527.2019.1599562

So, Richard Jean (2017): "'All Models Are Wrong'", in: *PMLA/Publications of the Modern Language Association of America* 132, 3: 668–73. DOI: 10.1632/pmla.2017.132.3.668.

Spedding, Patrick (2011): "'The New Machine': Discovering the Limits of ECCO", in: *Eighteenth-Century Studies* 44, 4: 437-453.

Wernimont, Jacqueline / **Losh, Elizabeth Losh** (eds.)(2018): *Bodies of Information: Intersectional Feminism and Digital Humanities*. Minneapolis: U Minnesora P.