Proposed medium-length scholarly article: “What Does Author Metadata Tell Us? Women Writers and Unsigned Works in Eighteenth-Century Literary Databases”

**Abstract**

Digital archives of literary works are often used to study large scale literary phenomena. However, these archives’ collection practices are not made transparent to researchers — especially in commercial archives, which market their collections as “comprehensive.” The inaccessibility of collection data makes it difficult to identify whether a digital archive is demographically representative of the historic literature it contains. The demographics of digital archives require closer scrutiny because of the high likelihood that these repositories recreate the historic patterns of exclusion and marginalization which often shape the physical collections upon which they are based. In nineteenth century novels, for example, recent work has found that women’s works are substantially less likely to be digitized than men’s. I examined records from four digital literary resource — the English Short Title Catalogue (ESTC), Eighteenth Century Collections Online (ECCO), HathiTrust, and the Text Creation Partnership (TCP) corpus — to evaluate how eighteenth century studies might be affected by uneven practices of digitization. Because the existing metadata in these resources does not allow a direct study of gender, I took a chronological sample from each archive of all works in any genre published in England between 1789 and 1799, and manually identified the authorial gender for these 51,000 works. The four resources, each of which includes different kinds of material, all had substantially different author demographics for the same time period. Women are attributed as authors for only 3% of the works listed in the ESTC, 3% of ECCO, 5% of HathiTrust, and a disproportionate 22% of the TCP. However, works by men are not proportionately reduced; instead, the works which fail to garner institutional support are those without known individual authors. Nearly half of the works in the ESTC (46%) are by unsigned or corporate authors, declining to 29% in ECCO, 20% in HathiTrust, and only 2% in the TCP. Much of this decline is caused by the fact that, once scholars identify the author of an unsigned work, they replace its attribution in the database with the new information. In this way, “newer and better” metadata obscures our knowledge of eighteenth century print, overwriting an eighteenth century culture of pseudonymous publication with a modern emphasis on the individuality of the author. It is also the case that a work without something appealing in the “author” field of the database is less likely to attract scholarly attention. By neglecting unsigned works, we fundamentally misunderstand the world in which women published during the eighteenth century— and, ironically, we likely overlook many works by women. Encoding the metadata of attribution and authorship as separate traits— and allowing other metadata fields, such as format or publisher, to eclipse authorship in importance— enables important new ways of knowing the history of print.