

The Nature of ECCO-TCP

Stephen H. Gregg

AT ITS LAUNCH in 1999, the Text Creation Partnership was a breakthrough collaboration between libraries and commercial publishers of digitized material. It aimed to provide collections of electronic texts from the early modern period that were freely accessible to the public, transcribed to a high degree of accuracy, and encoded to enable re-use and analysis. Its initial impetus was the collaboration with ProQuest's Early English Books Online (EEBO), but other collaborations were established with Readex's Evans Early American Imprints and Gale's Eighteenth Century Collections Online (ECCO). The TCP website provides a good history of its projects, and Shawn Martin's 2009 essay "A Universal Humanities Digital Library: Pipe Dream or Prospective Future?" offers useful background as well as reflects on the possibilities and challenges of the TCP project as whole. However, the EEBO-TCP collaboration has generated most scholarly commentary (see, for example, Welzenbach, 2012; Mak, 2014; Mueller, 2018; Gavin 2019; Herman, 2020). This interest reflects a number of factors peculiar to the success and visibility of EEBO-TCP. One factor was that, on its publication in 1999, EEBO consisted only of page images; in transcribing these page images, the TCP provided the text that enabled subsequent computational analysis and electronic text editing. The other significant factor was that the large number of texts transcribed—currently now around 65,000 texts—enabled the development of several large-scale projects for exploring and analysing the literature, language, and print culture of the period, for example, [The Early Print Library](#), [PRISMS](#), [Visualizing English Print](#), the [Early Modern OCR Project](#) (eMOP), and [Linguistic DNA](#).

In contrast—and although it is also used in several of the projects just mentioned—few analyses focus on the history of the TCP collaboration with ECCO. Consequently, unanswered questions remain about the nature of ECCO-TCP which this short essay aims to answer. Why did ECCO-TCP stop after a relatively small number of texts were transcribed? What organisational pressures and individual human choices shaped the nature and biases of the ECCO-TCP collection? In

addition—and aside from academic articles like this—how do we find the answers to such questions? As Roopika Risam has argued, “the reification of canons in digital form is not only a function of *what* is there—what gets digitalised and thus represented in the digital cultural record—but also *how* it is there—how those who have created their projects are presenting their subjects” (17). In short, how are such digital collections contextualised and their histories framed?

The scale of ECCO-TCP is relatively small compared to the larger and arguably more successful EEBO-TCP. Initial expectations for ECCO-TCP were high: 10,000 texts were planned to be transcribed.¹ However, between 2004 and 2012 only 3,101 texts were eventually transcribed and encoded, comprising 2,473 fully edited texts, and 628 released without being subject to final proofing and editing.² So, why did work stop? As I have suggested elsewhere, financial factors impinged on the sustainability of ECCO-TCP (75-76). The TCP is funded according to a “quasi-commercial model” in which libraries and institutions that purchased EEBO, Evans Early American Imprints, or ECCO could become contributing partners with the TCP; these funds were then matched by the commercial publishers, ProQuest, Readex, or Gale (Martin, 4). However, in 2006 TCP’s executive board predicted budget deficits and sought to secure more funding from its partner institutions (“TCP Executive Board”). Paul Schaffner, director of the TCP, recalled that, “we never received the financial support that we hoped for” and at some time after 2009, “we ran out of money” and the ECCO-TCP project used “what was left to review and complete the books in the pipeline” (Schaffner). By 2012, these financial constraints prevented ECCO-TCP from populating its site with additional transcribed and encoded texts.

The other problem that seemed to have sapped the energy behind the ECCO-TCP project was the question of its very nature. First, what exactly were the benefits of transcribing material from ECCO? What did the project hope to achieve? As mentioned earlier, TCP’s collaboration with ProQuest’s EEBO responded to a vital need and had a rigorous rationale; namely, it provided the searchable text which EEBO lacked. However, ECCO already had searchable text, produced by OCR software. Of course, it is the accuracy of text transcriptions which underpin any digital scholarship that uses the TCP collections. One of TCP’s missions was to “Present the user with accurately keyed, modern-font texts that are faithful to the spellings and organization of the original works.” ECCO’s notoriously messy OCR-produced text, though, rendered this objective impossible (Gregg 62-66). Nevertheless, TCP’s mission was complicated by the sheer size of ECCO and which clearly presented a huge challenge: what criteria would be used to select texts that would benefit from transcription from over 180,000 titles?

ECCO-TCP, like all human artefacts of collecting, is a product of institutional and human choices. Martin Mueller describes it as “a cherry-picked collection with an emphasis on canonical high-culture texts.” But how did it become that way? The geographic and linguistic biases of ECCO itself undoubtedly shaped its bias towards canonical authors (Tolonen, et al. 22-27). To a significant extent, this legacy can be traced to the foundations of ECCO: the microfilming project which tended to favour canonical male authors and the Anglocentrism of the originary 18th Short Title

Catalogue begun in 1976 (Gregg, 12-13, 23).³ In this context, the criteria established by a TCP “selection task force” set up in August 2005 is illuminating:

1. ECCO-TCP will use the *New Cambridge Bibliography of English Literature* as a guide to begin the selection process, because this standard reference work is by no means confined in scope to ‘literature,’ but provides a good overview of writing of all kinds — philosophical, religious, travel, periodical, historical, and so on.
2. ECCO-TCP will supplement these selections with suggestions from scholars, anthologies, and other bibliographies
3. Titles in languages other than English normally will be excluded from selection in ECCO-TCP.
4. ECCO-TCP will also, as far as possible, try to include works that will benefit from the added value the project brings (titles with complex structures like encyclopedias and works with bad OCR)
5. ECCO-TCP will include authors who cross the seventeenth and eighteenth centuries, such as Defoe and Swift, and will include their political, religious, and economic texts where appropriate in order to provide complete representation of these authors in the overall TCP collection.⁴

Schaffner noted that, apart from the broad and ambitious aim of identifying “added value,” these criteria were largely workable (for example, non-fictional works by Defoe are very well represented, attribution questions aside). However, these guidelines resulted in an uneven set of texts: decisions were inevitably subject to institutional pressures and individual human choice. For example, the relatively good representation of medical texts and Irish-themed fiction reflect the demands of particular partner institutions; and Schaffner himself acknowledged that his own interest in hymn books probably resulted in the inclusion of Isaac Watts, Charles Wesley, and Philip Doddridge (Schaffner). Decisions about what to include were also influenced by the use of the *New Cambridge Bibliography of English Literature volume 2: 1660-1800*, published in 1971 (!) and its definition of “Major” authors. So, there are no works of fiction by the popular early women writers such as Penelope Aubin, Eliza Haywood, or Delarivier Manley, but—as an instance of individual choice—twenty-two works by “Minor” novelist Samuel Jackson Pratt are included. It seems the selection task force must have argued for Olaudah Equiano’s *Narrative* to be transcribed for the collection since it is not listed in the *New Cambridge* bibliography, but works by other writers of the early black Atlantic, including James Albert Ukasaw Gronniosaw, Phillis Wheatley, Ignatius Sancho, or Ottobah Cugoano, were not selected.

The challenge presented by the lack of a clear argument for the project, a wide-ranging set of criteria, and the scale of ECCO resulted in a conservative and idiosyncratic collection that seems to have reflected eighteenth-century scholarship as it stood in the late twentieth century. On top of that, the small scale of ECCO-TCP arguably magnifies ECCO’s own inherent biases. Such biases also have the potential to impact any research based on the projects mentioned earlier. Literary and historical canons change, of course, and it might seem that I have unduly fixated on the use of a 1971 bibliography to decide in 2005 what texts were valuable for a digital collection.

But while the ECCO-TCP webpage acknowledges that it is “perhaps better described as a proof of concept than as a completed project,” it avoids detailing the various factors that have shaped the nature of the collection (“Text Creation Partnership”). That is, despite TCP’s laudable claim that “Our policies were imbued with a librarian’s attitude toward content: a resolve to prepare materials without agenda or bias, and with a view toward wide use and reuse,” this oversight remains. The larger point is that we need to understand the nature of these collections and their biases, and that—without users and researchers having to carry out some additional detective work—an explicit framing of the financial, institutional, and human contexts that shape how and why they are made is essential for a more nuanced understanding and use of such digital collections.

Bath Spa University

¹ Initial estimate courtesy of Jonathan Blaney.

² Notably, Gale did not ingest the TCP transcriptions into ECCO. In contrast, the UK organisation Jisc, another partner of TCP, ingested ECCO-TCP texts in its *Historical Texts* platform in 2016 (“Developmental Roadmap”).

³ Relatedly, TCP itself is not without its racial and gendered dimensions, since transcription is outsourced to workers in the Global South. See Mattie Burkert.

⁴ I obtained this unpublished “Selection Task Force Report” (9-10 August 2005) courtesy of Paul Schaffner.

Works Cited

Blaney, Jonathan. “RE: ECCO-TCP research,” Received by Stephen H. Gregg, 2 December 2019.

Burkert, Mattie, “From Manual to Digital: Women’s Hands and the Work of Eighteenth-Century Studies.” *Studies in Eighteenth-Century Culture*, vol. 52. Forthcoming [2023].

“Development Roadmap.” *Jisc Historical Texts*, historicaltexts.jisc.ac.uk/developmentroadmap. Accessed 27 August 2020.

Gavin, Michael. “How To Think About EEBO.” *Textual Cultures*, vol. 11, no. 1–2, 2019, pp.70–105. <https://doi.org/10.14434/textual.v11i1-2.23570>.

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

<https://www.cambridge.org/core/elements/old-books-and-digital-publishing-eighteenthcentury-collections-online/058DB12DE06A4C00770B46DCFAE1D25E>

Herman, Peter C. "EEBO and Me: An Autobiographical Response to Michael Gavin, "How to Think About EEBO." *Textual Cultures*, vol. 13, no. 1, 2020, pp. 207–16. <https://doi.org/10.14434/textual.v13i1.30078>.

Mak, Bonnie. "Archaeology of a Digitization." *Journal of the Association for Information Science & Technology*, vol. 65, no. 8, 2014, pp.1515–26. <https://doi.org/10.1002/asi.23061>

Martin, Shawn. "A Universal Humanities Digital Library: Pipe Dream or Prospective Future?" *Digital Scholarship*, edited by Marta Mestrovic Deyrup, Routledge, 2009, pp.1–12

Mueller, Martin, "Collaborative Curation of TCP Texts," October 2018, <https://scalablereading.northwestern.edu/?p=565>. Accessed 13 September 2019.

Risam, Roopika, *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Northwestern University Press, 2018.

Schaffner, Paul. "Re: DCCHELP-1238 Researching a history of ECCO," Received by Stephen H. Gregg, 19 Nov. 2019.

"TCP Executive Board Meeting Minutes 2006-09-16." *Archive-It*, [wayback.archive-it.org/5871/20190806191843/http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2006-09-16/](http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2006-09-16/). Accessed 25 August 2020.

"Text Creation Partnership," University of Michigan Library, textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/. Accessed 27 September 2022.

Tolonen, Mikko S., et al. 'Corpus Linguistics and Eighteenth Century Collections Online (ECCO)', *Research in Corpus Linguistics*, 9.1 (2021), 19–34

Watson, George, ed., et al. *New Cambridge Bibliography of English Literature volume 2: 1660-1800*. Cambridge University Press, 1971.

Welzenbach, Rebecca. "Transcribed by Hand, Owned by Libraries, Made for Everyone: EEBO-TCP in 2012," University of Michigan Library, <http://hdl.handle.net/2027.42/94307>. Accessed 27 September 2022.