CHAPTER 10

Afterword: Novel Knowledge, or Cleansing Dirty Data: Toward Open-Source Histories of the Novel

Emily C. Friedman

Introduction

In this chapter I discuss the most important, most under-rewarded, and most unsexy aspect of data visualization: the production and/or usage of reliable underlying data. Indeed, visualizations are only as good as their underlying evidentiary base. As Lauren Klein noted at the 2018 meeting of the Modern Language Association, "[w]e need to assemble more corpora—more accessible corpora—that perform the work of recovery or resistance."[1] This goes for metadata as well—something that, in theory, we do not lack for in the eighteenth-century novel: massive multigenerational bibliographies of the novel, for example. And more data is coming: for example, *The Cambridge Guide to the English Novel*,

[1] Lauren Klein, "Distant Reading after Moretti," *Arcade* (blog), 2019, https://arcade.stanford.edu/blogs/distant-reading-after-moretti

E. C. Friedman (✉)
Auburn University, Auburn, AL, USA
e-mail: ecfriedman@auburn.edu

I. Baird (ed.), *Data Visualization in Enlightenment Literature and Culture*, https://doi.org/10.1007/978-3-030-54913-8_10

1660–1820 will provide synopses of every surviving English novel, produced by expert readers from within the field.

But as Laura Mandell has noted, while we have a lot of data, what it represents is still—and will always be—partial: the works of women writers are disproportionately lost compared to those by their male counterparts, to give just one glaring example. And this makes for potentially disastrous effects when creating visualizations from that partial data. Where the sources for those visualizations are clear (and they are not always), they were already obsolete at the time of their construction.

But the main thrust of this chapter is to imagine universal standards for this work, which I argue must be at the center of any future reliable visualizations about novel history. I propose guidelines for best practices in creating new data so that amendable, transformable visualizations can be produced, built on collective knowledge. I note the contributions of digital projects which have laid the foundation for such practices, including (though not limited to) massive multi-institution projects like *Orlando*[2] and small to mid-sized projects like *The Early Novels Database* (END).[3] My own small-scale project, *Manuscript Fiction in the Age of Print, 1750–1900*, creates meaningful metadata about unprinted manuscript fiction during the period, creating a parallel corpus to those of published fiction. Because I work with never-printed fiction, there are unique challenges in identification, classification, and dissemination that are beyond the scope of this chapter, but that I have written about elsewhere.[4] As an active practitioner in the field of eighteenth-century fiction studies, I am aware of the very real obstacles to full implementation of any potential standards, even without the challenges relating to unprinted texts.

[2] *Orlando: Women's Writing in the British Isles from the Beginnings to the Present*, Cambridge University Press, accessed April 15, 2019, http://orlando.cambridge.org/

[3] *The Early Novels Database*, accessed April 15, 2019, https://earlynovels.github.io

[4] See Emily C. Friedman, "Amateur Manuscript Fiction in the Archive: An Introduction," in *After Print: Eighteenth-Century Manuscript Cultures*, ed. Rachael Scarborough King (Charlottesville: University of Virginia Press, 2020), 217–36; "Must Anonymous Be a Woman? Gender and Anonymity in the Archives," *Tulsa Studies in Women's Literature*, Special Issue on "Women in Archives" (forthcoming 2021); "Ownership, Copyright, Ethics of the Unpublished," in *Access and Control in the Digital Humanities*, ed. Richard Mann and Shane Hawkins (forthcoming 2021); and "Eluding Print: Manuscript Fiction and the Survival of Scribal Practices in the Age of Print," Special Issue of *Huntington Library Quarterly* by the Women in Book History Research Group (forthcoming 2021).

I am completing this chapter in the shadow of Nan Z. Da's recent critique of computational literary studies,[5] which appeared first in *Critical Inquiry* and then in a more mainstream variation in *The Chronicle of Higher Education*.[6] The former publication almost immediately hosted an entire preplanned suite of responses within a week of the essay's publication. There is much to say about the circumstances and the circulation of Da's argument, much of which lies outside the scope of this chapter. And Da's essay is, in a sense, just one more in a very long line of critiques of the emerging discipline of digital humanities, both by those who wish to create best practices and by those who wish to discredit it utterly. Excavating all the layers of power, risk, and citational strategy present in very public fora is beyond my scope here. Nevertheless, what Da and I—and all of us—share is an investment in the verifiability of claims that are numerically based—either via counting or communicated via visualization. We will be the better for publicly stated, and preferably publicly verified, forms of accountability.

Dreaming of Ideal Data

Before I begin, I would like to briefly imagine the perfect conditions for the study of the eighteenth-century novel at scale. In that alternate universe, a copy of every work of fiction—at least!—was preserved from the flames and the privy. While we are dreaming, go one better: a copy of *every edition* of every work was saved. Or, if we would like to be profligate with our wishes, *every copy*. In this dream world, the Stationer's Company requested and required the author's name, even if the work was pseudonymously or anonymously published. Expand the vision further: their address was also required, and maybe some demographic data (age, gender identity, occupation).

Moreover (to ride a personal hobbyhorse), every manuscript of fiction, published or unpublished, was somehow preserved according to the same principles. Perhaps that goes too far—we do not want to intrude on the actual material conditions of production too far, and manuscripts were

[5] Nan Z. Da, "The Computational Case against Computational Literary Studies," *Critical Inquiry* 45, no. 3 (2019): 601–39, https://doi.org/10.1086/702594

[6] Nan Z. Da, "The Digital Humanities Debacle: Computational Methods Repeatedly Come Up Short," *The Chronicle of Higher Education*, March 27, 2019, https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986

often simply a casualty of the ways books became printed books. But, at least, we can dream that no manuscript was ever disconnected from its author or authors, nor crossed borders without documentation.

We can dream that all of this information has been consolidated in one place: a single complete bibliography and checklist of fiction, produced through unprecedented collaboration, available not only in the durability of print, but in an open access relational database, allowing users to search by all of these different known fields of information, creating book lists researchers could use to dive into a given author, publisher, neighborhood, profession, or more.

In this world of unlimited—enthusiasm? money? time? investment?—, optical character recognition of both eighteenth-century type and handwriting is perfect, and at least minimally encoded digital surrogates of each work are available. Not only the bibliography but also the digital corpus itself would then be truly comprehensive and reflective of the fiction produced.

Dream still bigger and imagine every circulating library, private library, bookseller's records or catalogs also survived, preferably with the kind of detailed demographic information of, for example, the New York Society Library's records, now digitized in the *City Readers* database.[7] Imagine every mention of any literary work from correspondence, reviews, and the like somehow gathered together, a macro version of Cardiff University's *British Fiction, 1800–1829: A Database of Production, Circulation, and Reception*.[8]

Luxuriate in what we might be able to know, to glean. What we might be able to ask such a dataset, how we might represent different aspects of it.

Waking Up

And then, reluctantly, wake up.

Face what we actually have: a fraction of what we know to have been produced. We are haunted by a Great Forgetting (to put a new spin on Clifford Siskin's phrase) that we cannot possibly recover from. We know

[7] *City Readers. Digital Historic Collections of the New York Society Library*, The New York Society Library, accessed April 15, 2019, http://cityreaders.nysoclib.org/

[8] *British Fiction1800–1829. A Database of Production, Circulation & Reception*, Cardiff University Centre for Editorial and Intertextual Research, accessed April 15, 2019, http://www.british-fiction.cf.ac.uk/

this, for example, from periodical advertisements for books that we have yet to recover copies of. For what we do have, we face incomplete records, innumerable challenges of attribution, and varying levels of reliable transcriptions and scans. Enormous amounts of digitized microform surrogates are still only partially representative of what survives, let alone what once existed. And that's what is available to those with institutional access to proprietary databases.

It is worth noting, for the record, that these are all known losses from *print* fiction; the losses from manuscript culture during the age of print are still more profound because they are largely unknowable. Unlike the print marketplace, which was largely controlled by the Stationer's Company, works of manuscript fiction in the eighteenth-century and thereafter were not necessarily known, named, and documented for commercial exchange. The nature of their survival varies wildly from that of the print work. For that reason, we do not yet, nor may we ever know how much manuscript fiction survives in inaccessible or nearly inaccessible private collections. We certainly will never have a full sense of how much literary work was produced in manuscript and subsequently intentionally destroyed or accidentally lost. Critical consensus for a substantial amount of time was that these were not significant losses: they were unimportant precisely because manuscripts with no connection to print culture have no value. While attempting to create a database and digital corpus of manuscript fiction, I have seen the difficulties associated with such critical dismissal. Such dismissal, and by extension poor reckoning, also occurred to fiction that appeared in so-called "ephemeral" forms: in periodicals, pamphlets, cheap editions, and so on.

From this perspective, one's work is tinged by the regret that no matter how long or how hard we labor, we will still be faced with what has been obliterated. To dream of the perfect dataset is to dream away the very real working conditions we hope to study: the fact that a manuscript of a printed work is nearly always destroyed in the process of production, the fact that selection and disposal themselves are part of history. I am reminded of Aden Evan's description of the digital as "calculably imprecise," by which he means the digital is capable of weighing and measuring "to a given level of accuracy and no more"—that "no more" being the very fuzziness of objects and individuals whose borders are never entirely

demarcated.[9] We can enumerate and visualize and analyze more texts than ever before, but unless the end results show us something in soft-focus, slightly blurred, we have missed a key aspect of our knowledge, that is, what we cannot know.

Even with all such caveats, we have access to spectacularly more information, that is more rapidly retrievable, than any prior scholarly generation ever had to hand. It is important to recall that we are luckier than some other fields: we do not face the challenges attendant with, for example, reassembling page by page (often scattered across the globe) the surviving Coptic corpora described by Schroder and Zeldes.[10] Moreover, in addition to what has been digitized, we have enormous datasets of surviving eighteenth-century prose fiction produced in the twentieth century. We tend to call these datasets "bibliographies"—the monumental works of Beasley, Forster, Garside, Letellier, McBurney, Raven, Schöwerling—understandably, perhaps, given that they officially exist exclusively in print editions.[11] These works of enormous intellectual labor are preserved and stabilized by their instantiation in print, but to use them as something other than a reference requires laborious transformation, which due to copyright regulations cannot be then shared publicly. I often wonder how many times researchers (or their student assistants) have privately done this kind of transformative work that allows for dynamic analysis and visualization of these bibliographies, and how many *sub rosa* exchanges of that information occur. In an ideal world, these bibliographies would be undergoing the same kind of digitization project as that currently led by Mattie Burkert to transform *The London Stage* from print to a truly usable database.[12] But, as Burkert has noted of *The London Stage* and Katherine Bode

[9] Aden Evans, *Sound Ideas: Music, Machines, and Experience* (Minneapolis: University of Minnesota Press, 2005).

[10] Caroline T. Schroeder and Amir Zeldes, "Raiders of the Lost Corpus," *Digital Humanities Quarterly* 1, no. 2 (2016), http://www.digitalhumanities.org/dhq/vol/10/2/000247/000247.html

[11] The University of Michigan has digitized *The English Novel 1770–1829: A Bibliographical Survey of Prose Fiction Published in the British Isles* (Oxford: Oxford University Press, 2000) and placed it on HathiTrust, allowing for search-only access by non-UM users.

[12] The project is currently in beta at https://londonstagedatabase.usu.edu/. Burkert wrote about the project's origins in the 1960s London Stage Information Bank in "Recovering the *London Stage Information Bank*: Lessons from an Early Humanities Computing Project," *Digital Humanities Quarterly* 11, no. 3 (2017), http://www.digitalhumanities.org/dhq/vol/11/3/000321/000321.html

has reminded us about the fiction bibliographies of the last half-century, these datasets are themselves still partial and decontextualized.[13]

From one perspective, eighteenth-century scholars of the novel are swimming in data: massive periodical runs survived in some form or another full of reviews and advertisements about new fiction, more novels survived than can possibly be read by one human, and the Stationer's Company trade records provided enormous amounts of information about the marketplace for commercial fiction in the period. Another essay could easily be written about the challenges attendant in wrangling the "big data" aspects of eighteenth-century studies. Certainly, the challenges attendant with our search-driven research milieu have been discussed before. As Ted Underwood has noted, most literary scholars have transitioned into practices of searching, often using tools with proprietary algorithms that are opaque even to those with the technical expertise to understand them. The implicit "understanding" of what search is and how it operates is now so established that there are tenured faculty whose careers have taken place entirely within its mindset.[14] Our scholarly habits have been transformed by a practice that seems so simple that the vast majority of practitioners, even very sophisticated ones, have not built up a theoretical lens through which to understand what is "found" and what is overlooked in such practices—shockingly considering this is a now decades-old research practice. But as Underwood's work with classification algorithms shows, there are many places where algorithmic search and the substantial capacities of the analysis break down: while they are good at making broad category distinctions between "fiction, drama, poetry, and nonfiction prose," they are poor at detecting parody, fine distinctions between subgenres, or genres that exist across a long period of time (and thus across large shifts in generic expectations).[15] Thus, researchers of eighteenth-century fiction face challenges of simultaneously *not*

[13] Katherine Bode, *A World of Fiction: Digital Collections and the Future of Literary History* (Ann Arbor: University of Michigan Press, 2018), 20–21.

[14] Ted Underwood, "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago," *Representations* 127, no. 1 (2014): 64–72, https://doi.org/10.1525/rep.2014.127.1.64

[15] Ted Underwood, Michael L. Black, Loretta Auvil, and Boris Capitanu, "Mapping Mutable Genres in Structurally Complex Volumes," *Proceedings of the 2013 IEEE International Conference on Big Data*, ed. Xiaohua Hu et al., 95–103. Silicon Valley, CA: IEEE, 2013. https://doi.org/10.1109/BigData.2013.6691676

enough and *too much*, which are also the challenges attendant with fields from particle physics to genomics and beyond.

That said, to throw up our hands, turn our backs on large-scale collaborative projects, and cultivate our small, impressionistic gardens seems to be falling off the other side of the horse. If computational work and the visualizations used to communicate that information to larger audiences do not save us, they probably will not damn us either. My call here is twofold: to afflict the comfortable and comfort the afflicted, as it were. We should absolutely challenge and critique work that makes too sweeping claims based on macro-analysis from already available data and text sets and call, instead, for more work in making such sets more robust, and representing their outputs more transparently and with more attention to the unknowable. Simultaneously, the perfect cannot be the enemy of the good: we should recognize and reward more robustly digital editorial work that refines our existing knowledge and allows for future iteration and improvement.

Case Studies: Where We Are Now

In this section, I want to survey the landscape of available data and related public-facing projects that tackle enumerating, describing, analyzing, and/or visualizing works of eighteenth-century fiction. It is worth noting at this point that the vast majority of such projects are in article or monograph form, and either do not make their data or work materials publicly available, or work from curated samples of one of two major data and text sets, which I will describe below. Leah Orr's groundbreaking *Novel Ventures: Fiction and Print Culture in England, 1690–1730* is an excellent example of the former.[16] Her work considers 475 works of fiction printed between 1690 and 1730 in order to assess more precisely the fiction market of the period. Many of the texts Orr considers had not been previously included in any of the standard bibliographies; she lists those titles within an appendix in her monograph. Her monograph also includes tables and charts describing trends in genre over the period. But it would be difficult to test her arguments because there is no space to reproduce her entire list of works, alongside how she coded them by genre. For those (like me) who would love to see other scholars adopt Orr's devotion to precision

[16] Leah Orr, *Novel Ventures: Fiction and Print Culture in England, 1690–1730* (Charlottesville: University of Virginia Press, 2017).

and careful assessment decade by decade throughout the eighteenth century, this is a disappointment.

But this is not just Orr's challenge: it is a challenge that all of us face when we work (as Orr did) with bibliographies and other very large datasets. Almost all such projects were produced at least in part within for-profit publishing structures. I have already touched on the nature of bibliographies in this respect, which are all printed books of recent production, published by major presses, unlikely to take on an open access transformation of the data within those volumes, and even less likely to do so together. Could Orr have released her spreadsheets? It seems unlikely.

Nor is this simply the challenge attendant when one's research is extending existing print resources. Proprietary databases have significant affordances and limitations of their own. Many research projects, including the ones that would not identify themselves as "digital," are built on data that emerges from Gale Cengage's *Eighteenth-Century Collections Online* (ECCO), a subscription database that claims to include "every significant English-language and foreign-language title printed in the United Kingdom during the 18th century, along with thousands of important works from the Americas."[17] Even for scholars who will never attempt to visualize large amounts of data, ECCO has become an essential component of academic work in eighteenth-century studies, at least for those who work or study at institutions that can afford its cost. ECCO's own platform allows for full-text searching of the corpus, but because the full text is derived from optical character recognition (OCR) of digitized black and white copies of microform surrogates, there are limitations to the capabilities of search due to errors.[18]

In order to improve the full-text searchability of their database, Gale Cengage partnered with the Text Creation Partnership (TCP),[19] which also partners with ProQuest's *Early English Books Online* (EEBO) and Readex's *Evans Early American Imprint*s. In exchange for human labor providing human-transcribed texts, these publishers allow the TCP to

[17] This claim appears on the Text Creation Partnership's page describing "ECCO-TCP: Eighteenth-Century Collections Online," 2019, https://www.textcreationpartnership.org/tcp-ecco/

[18] This has been ably discussed by Patrick Spedding in "'The New Machine': Discovering the Limits of ECCO," *Eighteenth-Century Studies* 44, no. 4 (2011): 437–453, https://www.jstor.org/stable/41301590

[19] "EEBO-TCP Phase I Public Release: What to Expect on January 1," *TCP*, December 24, 2014, https://www.textcreationpartnership.org/

make the data and metadata of these texts publicly available in XML-encoded electronic editions and plain text. In a related endeavor, Typewright, a tool hosted by the aggregation site 18thConnect.org, allows users to correct the OCR-generated transcriptions in ECCO. If any users contribute to the correction of a given text, they will be given access to the completed transcription and XML file for their own use, even if they do not have institutional access to ECCO. Gale Cengage can afford this generosity because ECCO-TCP is unlikely to put ECCO out of business any time soon: at last publicly reported count, the ECCO-TCP corpus included 2231 texts—roughly 1.5% of the over 136,000 titles in ECCO.

Thus, while there is a pathway to facilitate access to a human-scaled corpus of texts (say, all editions in ECCO of the work of Samuel Richardson), ECCO-TCP is still comparatively tiny. Depending on the kind of questions you want to ask—for example, about the frequency of a word in print usage across the century—you would find ECCO-TCP both too small and too unrepresentative of the whole of ECCO, to say nothing of print culture, to be very certain of your findings. More texts, transcribed and ideally encoded, are needed for such work.

This is the goal of the HathiTrust (HT), a collective project to create a "shared repository of cultural heritage materials" by combining efforts of various library and collection digitization projects. HT facilitates what it calls "non-consumptive research," including data analysis and visualizations, from its corpus of millions of titles, of both in-copyright and out of copyright works. "Non-consumptive" means that the researcher is conducting computational analysis without reading or displaying a "substantial portion" of a volume if it is in copyright.[20] In other words, HathiTrust Research Center (HTRC) walks a very careful line that avoids any suggestion of copyright infringement by restricting the kinds of data exports that can be performed by a user. Anything that could recreate an in-copyright text would be *verboten*, understandably, while summaries, token counts, topic models, and the like can be exported. Datasets and text sets have been created for users, but more importantly users can create work sets that can be shared and cited.

Compared to ECCO-TCP, HTRC's corpus is both smaller (only 505 volumes from ECCO directly) and far larger: over 34,000 books in English produced between 1700 and 1800, roughly 26,182 that describe

[20] "Non-Consumptive Use Research Policy," HathiTrust Research Center, accessed February 20, 2019, https://www.hathitrust.org/htrc_ncup

themselves as either prose, fiction, novel, tale, or story. This is an imperfect reckoning of what is available in a few ways: first, not all items are available full-text for all users, and second, the only metadata connected to genre for the user interface is based on subject headings.[21]

HTRC is a model of transparency, keenly aware of the gaps in its underlying scope. As the documentation for one project, *Word Frequencies in English-Language Literature, 1700–1922*, notes, "[c]ontributing institutions are mainly located in the United States. So, while the collection contains volumes from around the globe, coverage of works published in the U.S. is more complete. Also, because books before 1800 may be held closely in Special Collections, digitization of that period is less predictable. We don't necessarily recommend this dataset as a source for literary research before 1750."[22] Thus, projects using HTRC's metadata or text sets must be exceptionally careful in their construction of work sets, or else their conclusions (here, visualizations) will reflect more the holdings of contributing institutions than anything else.

Both the HTRC and ECCO-TCP emerge from massive ongoing digitization and transcription efforts. They allow for their material to be cited and used at enormous scale. But while both have text and data sets that can be used, they are not extensible by a given user. Researchers can carefully select which items go into their corpus as they prepare to do visualization work within the HTRC ecosystem, but should they transcribe and encode their own additional items, ingestion (as far as I understand it) would not be possible. And, as I will argue in the final section of this chapter, extensibility is absolutely essential to the robustness of our research moving forward.

Mark Algee-Hewitt et al. remind us to distinguish between "the published, the archive, and the corpus," that is, "the totality of the books that have been published," "that portion of published literature that has been preserved," and "that portion … that is selected, for one reason or another,

[21] *Word Frequencies in English-Language Literature, 1700–1922* is specifically not a text, but a dataset concerned with providing word frequencies at the volume and page level for fiction, drama, and poetry. For their documentation, see Ted Underwood, Boris Capitanu, Peter Organisciak, Sayan Bhattacharyya, Loretta Auvil, Colleen Fallaw, and J. Stephen Downie, *Word Frequencies in English-Language Literature, 1700–1922*, [Dataset], HathiTrust, 2015, https://doi.org/10.13012/J8JW8BSJ

[22] Ibid.

in order to pursue a specific research project."[23] While the authors imagine a convergence of all three that "may soon be a reality"—a theoretical imaginary that they themselves walk back as they move further into their argument, they do not indicate explicitly that their sampling procedure is "an ideal model of research" but note that "dirty hands are better than empty."[24] Dirty data is better than no data, but it is potentially misleading in the same way that a dirty window not only obscures what is present, but might suggest shadows of things not present at all.

Given the right partnership, and a different intellectual framework, it is possible to create a meaningful corpus without impossible comprehensiveness. In her analysis of the Stanford Lab model, Katherine Bode notes that without a robust knowledge of the "bibliographic and editorial practices" of the historical moment in which texts were produced, there is an "inadequate foundation" upon which to build one's argument.[25] In response, her latest book, *A World of Fiction: Digital Collections and the Future of Literary History*, walks through her work focusing on a subset of the 16,500 works of fiction identified in The National Library of Australia's TROVE database. This work is meant to create what she calls a "scholarly edition of a literary system"[26] that emphasizes the constructedness of any corpus, paralleling the traditional scholarly edition, which is an "argument" present through a "curated text" that is "designed to enable and advance rather than to decide or conclude—investigation."[27] To that end, her dataset, *Reading by Numbers*, is openly available on several locations on the web, including TROVE's site, so that her choices can be fully understood and engaged with.[28] As a sampling method, it has the virtue of being transparent and thus correctable/iterative over time.

Creating cleaner, open-source, extensible data is also more achievable when one turns to projects that do not digitize, transcribe, or encode, but focus on the production of metadata. *The Early Novels Database* (END),[29]

[23] Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser, "Canon/Archive. Large-Scale Dynamics in the Literary Field," *Literary Lab Pamphlet 11* (January 2016): 1–13, https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf

[24] Ibid.

[25] Bode, *A World of Fiction*, 5.

[26] Ibid., 4, and further described on 6.

[27] Ibid., 6–8.

[28] *Reading by Numbers*, AusLit, accessed April 15, 2019, https://www.austlit.edu.au/specialistDatasets/ResourcefulReading/ReadingByNumbers

[29] *The Early Novels Database*, accessed April 15, 2019, https://earlynovels.github.io/

led by Rachel Sager Buurma and Jon Shaw, has produced over 1800 records in a MARC-based schema,[30] representing novels published between 1660 and 1850. Its object of focus is individual copies of a given novel, describing a variety of features that require close human attention: prefaces, introductions, tables of contents, and so on. END generates new metadata for existing collections, which creates a beneficial and mutual relationship between librarians and archivists, faculty specialists in literature, and their student collaborators. That literary scholars and their students have learned, in essence, one of the fundamental structuring data schemas of librarianship (i.e., MARC), means that the project is far more likely to grow beyond its original scope. And it has: while its own dataset is largely bounded by the institutional and geographic limitations of its Philadelphia-based team, the data is openly available and the schema extensively documented on GitHub, providing a model for other research teams to add on their own, site-specific data. I have experimented with having my own students in eighteenth-century novel courses prepare descriptions of Auburn's small collection of relevant eighteenth-century printed texts for incorporation into the END, and I have found it takes only a minimal amount of ramp-up time to do so.

A project need not be massive or even large to be open, extendable, and useful. A good example is *Reading With Austen* (RWA),[31] a project led by Peter Sabor and managed by Catherine Nygren and Megan Taylor, which creates a visual representation of the shelves in the library of Godmersham Park, the estate of Jane Austen's brother, Edward Austen Knight. They were particularly fortunate to have an 1818 handwritten catalog[32] that included the shelf locations of each book. The majority of the books, collectively named The Knight Collection, are still at the estate (now Chawton House Library). The longer-term goal of the project is to find the "lost sheep" via the distinctive Knight family bookplates, and either purchase them for reincorporation into the Knight Collection or virtually "reunite" them through photographs of the book's spine, bookplate page, title page, and any pages with marginalia. This project requires support from the

[30] MARC, or MAchine-Readable Cataloging Record, was developed by Henriette Avram, Sally McCallum, and Lucia Rather at the Library of Congress in the 1960s to create shared standards for libraries. While it has been revised over the years, MARC continues to be the *lingua franca* of structured data for libraries.

[31] *Reading With Austen*, accessed April 15, 2019, http://www.readingwithausten.com/

[32] This catalog was turned into a spreadsheet by Deborah Bygrave and Hugh MacKay under the direction of Stephen Bending and Stephen Bygrave.

public, and thus it is extremely transparent in its process. While its core dataset has not yet been published, the project website notes that they will release that clean dataset under a Creative Commons License, understanding that many queries can be run with such a dataset that would not be possible with their library-shelf visualization or catalog search interface.

RWA is able to be open with its data because of two reasons. First, it received funding from Social Sciences and Humanities Research Council (SSHRC) Canada Research Chairs Program, which did not put any requirements of exclusivity or monetization on the output. Second, the Knight family already had an entire apparatus of agreements that made their family home and the Knight Collection accessible to scholars through the creation of Chawton House Library and a long-term loan of the Knight Collection. Without the cooperation of Richard Knight, it would have been far more difficult, if not impossible, to have created and disseminated the underlying spreadsheet, to say nothing of the visualizations and photographs that are part of the front end. Other projects on historical libraries face such restrictions as they come to fruition.

Orlando: Women's Writing in the British Isles from the Beginnings to the Present (*Orlando*), led by Susan Brown, Patricia Clements, and Isobel Grundy since 2008, is a rich database on the history of women's writing in Britain. Its encyclopedia entries and other data allows users to search for women across an enormous sweep of interpretive tags, to move between and among them through densely hyperlinked entries on people, primary and secondary material, and contextual information, and to create timelines for their own use. It is also an important example of a massive scholarly project that is largely controlled by a publisher (i.e., Cambridge University Press) that requires that it be held behind a paywall, and while its tag structure can be revealed through the user interface, the user interface cannot be superseded: the site's data cannot be analyzed outside of the framework of the interface. While the latter obstacle is unavoidable, *Orlando* provides two workarounds for the challenge of the paywall: free access during March (Women's History Month), and a scheme in development that allows contributors access to the project in exchange for their labor. This is an important initiative, and one that is very much in line with the ideals of the editorial team.

It is important to note that *Orlando* here stands in for innumerable other projects, large and small, that now try to balance the needs of the editorial and technical team with those of the end users. It is easy to call for open access, but someone must pay for the labor that goes into the

work of transcribing, scanning, encoding, coding, building, and then sustaining and continually updating a large-scale text set or database. As scholarly publishing must increasingly justify its expenditures using the value systems of the market, it is little surprise that publishers would look to support subscription-based digital projects or create public interfaces but retain valuable back-end data. As I will note in the next section, we should not necessarily assume that this is needed when it comes to small- and medium-scale datasets.

Conclusion: Toward Best Practices

In short, our work would be much easier if we were more able to work together. Here, I propose what I believe are reasonable best practices for those of us that are working with or creating large datasets connected to Anglophone fiction in the long eighteenth century. I suspect I am not saying anything new to the majority of practitioners, but unlike (for example), the Text Encoding Initiative (TEI), which makes collective standards for the sustainable and interoperable encoding of individual texts, there is no collective or regulatory body for the interoperability of datasets in a similar fashion. Thus, it is worth reiterating and making explicit practices that make our work iterative—not just for today, but for generations to come. It is my hope that this is only the starting point of a larger conversation about creating durable and interoperable datasets, and that refinements and additions will come to address omissions or new challenges or affordances that come with new technology.

Clarify the Definition of "Novel" and "Fiction" in Our Work. As a collective action, I realize this may well be as much of a pipe dream as any articulated in this chapter. I can barely imagine a singular agreed-upon definition that all scholars of the novel or even prose fiction would agree upon. But ensuring that we articulate our individual project's boundaries in this respect is critical; also, ensuring that the datasets, bibliographies, or other sources of lists employ the same definition is absolutely vital.

Acknowledge Sample Size and Sources. As literary scholars, we need to get into the habit of thinking about how our evidence arrives to us. We need to be explicit early and often about where we get our data, especially when we are extrapolating them into numbers. Most (if not all) of us work from either small text sets we have created ourselves, or from preexisting

large-scale corpora, and neither is without its attendant blind spots and challenges. As I have discussed already, large-scale corpora give the illusion of comprehensiveness, which can be misleading. And small text sets potentially suffer from selection bias. It is almost surprising that Da does not take on the question in her tests of statistical methodology from extant projects, but it is possible that she sees that as well-trod territory.

I want to affirm that there is not one standard threshold to determine the appropriateness of a conclusion. If we deal with eighteenth-century fiction, we always deal with a small slice of the printed material produced in the period—a reality we do not always acknowledge, but we should. As Leah Orr notes in her study of the early eighteenth-century print fiction market, when we talk about new titles in a given year, we talk about very small numbers: even including reprints, translations, and the like, just 475 different works of fiction were printed in England in the period 1690–1730.[33] If you chose to focus on new works each year, you would be referring to a much smaller set still. There's lots we can say about fiction and its readers, but it is crucial to keep scale in mind.

Contextualize. As noted above, gaps and silences in data are nearly inevitable when dealing with eighteenth-century fiction. Those who read our work or engage with our data should have a clear understanding of the curatorial choices that went into selection and those choices should be defensible through evidence drawn from knowledge about the production of texts in the period. To the extent possible, a project should note the gaps in data. As Nathan Yau has pointed out, missing data is itself a useful piece of information; therefore, visualizations should use white space, variable scale, and treat absence itself as a category in order to represent this absence effectively.[34]

Open Dataset. Ideally, every peer-reviewed publication that relied on enumerative or visualized datasets would require either a publicly available attachment of the corpora and tools used or, at minimum, to have that information stored in institutional repositories (even if not publicly available). As I have discussed in this chapter, some do—many don't. One of the most searing footnotes in Da's essay for me was the following: "the process of requesting complete, runnable codes and quantitative results

[33] Orr, *Novel Ventures*, 103.

[34] Nathan Yau, "Visualizing Incomplete and Missing Data," FlowingData, accessed April 15, 2019, https://flowingdata.com/2018/01/30/visualizing-incomplete-and-missing-data/

(tables, output data, matrices, measurements, and others) took me nearly two years. Authors and editors either never replied to my emails, weren't able or willing to provide complete or runnable scripts and data, or gave them piecemeal only with repeated requests."[35] Da's experience is common across disciplines. Digital humanities are not out of step with all other data-heavy fields in this respect: while collaboration and data sharing is common in a handful of fields, a 2002 study published in *The Journal of American Medical Association* showed that 45% of geneticists withheld data because of the attendant expenses, and 80% noted that the effort required to share data made them unlikely to do so.[36] A 2012 study that interviewed researchers in various fields revealed that lack of disciplinary standards and repositories shaped whether an individual researcher was likely to share his or her data with others.[37]

For a variety of reasons that I have discussed in this chapter, publicly accessible datasets or text sets are not always achievable. The major bibliographies are all decades away from entering the public domain. The costs of permissions of various sorts become unwieldy at anything approaching large scale. While the published fiction we work with is officially out of copyright by every conceivable measure, the photographs and scans of those texts are not. Moreover, not all fiction produced in the period was "published": as of the time of this writing, never-published literary manuscripts produced before the mid-twentieth century will remain in UK copyright for another twenty-one years and declaring such a work "orphaned" requires up-front fees with no assurance of success. Even when institutions waive such fees and grant permissions, making these images available to a larger public still incurs the recurring costs of server space and maintenance. Even among scholars, data is power—but power of a strange kind. Because we, as a scholarly community, do not value dataset production in the ways we would monographs: publication of data comes with no reward or incentive. Thus, data is too often only valued as

[35] See note 2 in Da, "The Computational Case against Computational Literary Studies," 602.

[36] E. G. Campbell, B. R. Clarridge, M. Gokhale, L. Birenbaum, S. Hilgarten, N. A. Holtzman, and D. Blumenthal, "Data Withholding in Academic Genetics: Evidence from a National Survey," *Journal of American Medical Association* 287, no. 4 (2001): 473–80, https://doi.org/10.1001/jama.287.4.473

[37] Youngseek Kim and Jeffrey M. Stanton, "Institutional and Individual Influences on Scientists' Data Sharing Practices," *Journal of Computational Science Education* 3, no. 1 (2012): 47–56, https://doi.org/10.22369/issn.2153-4136/3/1/6

the raw material from which publications come—publications which cannot be thoroughly vetted, peer-reviewed, understood, much less built upon, without that data being available. Moreover, because there may be another publication that might emerge from it, there is every reason to guard one's data until a sufficient number of publications have been wrung out of it—if indeed the data is ever made public. This is a value system that we must collectively transform if we are ever to succeed.

Extendable. Some gaps are the product of data that can never be retrieved. But gaps like those of ECCO-TCP (and, indeed, ECCO itself) can ultimately be filled in, though the scale of the task is daunting. No one research team, no matter how well-funded, is likely to transcribe all known fiction from the period. Instead, we should work and create documentation for our data that begins with the assumption that our work will be picked up and continued by others. And we must make space—physical, digital, conceptual—for that work to occur in.

Indeed, data visualization is not a static process: today's network visualization or graph can and should be expected to be superseded by new information gained by the team, by new ways of asking questions of existing data or text sets, or ideally by both. Many of our current ways of building data collections and conducting data analysis still fail to wholly tackle these challenges.

Bibliography

Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2016. Canon/Archive. Large-Scale Dynamics in the Literary Field. *Literary Lab Pamphlet 11* (January): 1–13. https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf

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

Burkert, Mattie. 2017. Recovering the *London Stage Information Bank*: Lessons from an Early Humanities Computing Project. *Digital Humanities Quarterly* 11 (3): n.p. http://www.digitalhumanities.org/dhq/vol/11/3/000321/000321.html

Campbell, E.G., B.R. Clarridge, M. Gokhale, L. Birenbaum, S. Hilgarten, N.A. Holtzman, and D. Blumenthal. 2001. Data Withholding in Academic Genetics: Evidence from a National Survey. *Journal of American Medical Association* 287 (4): 473–480. https://doi.org/10.1001/jama.287.4.473.

Da, Nan Z. 2019a. The Computational Case against Computational Literary Studies. *Critical Inquiry* 45 (3): 601–639. https://doi.org/10.1086/702594.

———. 2019b. The Digital Humanities Debacle: Computational Methods Repeatedly Come Up Short. *The Chronicle of Higher Education*, March 27. https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986

Evans, Aden. 2005. *Sound Ideas: Music, Machines, and Experience*. Minneapolis: University of Minnesota Press.

Friedman, Emily C. 2020. Amateur Manuscript Fiction in the Archive: An Introduction. In *After Print: Eighteenth-Century Manuscript Cultures*, ed. Rachael Scarborough King, 217–236. Charlottesville: University of Virginia Press.

Garside, Peter, James Raven, and Rainer Schöwerling, eds. 2000. *The English Novel 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles. Volume I: 1770–1799*. Oxford: Oxford University Press.

Kim, Youngseek, and Jeffrey M. Stanton. 2012. Institutional and Individual Influences on Scientists' Data Sharing Practices. *Journal of Computational Science Education* 3 (1): 47–56. https://doi.org/10.22369/issn.2153-4136/3/1/6.

Orr, Leah. 2017. *Novel Ventures: Fiction and Print Culture in England, 1690-1730*. Charlottesville: University of Virginia Press.

Schroeder, Caroline T., and Amir Zeldes. 2016. Raiders of the Lost Corpus. *Digital Humanities Quarterly* 1 (2): n.p. http://www.digitalhumanities.org/dhq/vol/10/2/000247/000247.html

Spedding, Patrick. 2011. 'The New Machine': Discovering the Limits of ECCO. *Eighteenth-Century Studies* 44 (4): 437–453. https://www.jstor.org/stable/41301590.

Underwood, Ted. 2014. Theorizing Research Practices We Forgot to Theorize Twenty Years Ago. *Representations* 127 (1): 64–72. https://doi.org/10.1525/rep.2014.127.1.64.

Underwood, Ted, Michael L. Black, Loretta Auvil, and Boris Capitanu. 2013. Mapping Mutable Genres in Structurally Complex Volumes. In *Proceedings of the 2013 IEEE International Conference on Big Data*, ed. Xiaohua Hu et al., 95–103. Silicon Valley, IEEE. https://doi.org/10.1109/BigData.2013.6691676.

Underwood, Ted, Boris Capitanu, Peter Organisciak, Sayan Bhattacharyya, Loretta Auvil, Colleen Fallaw, and J. Stephen Downie. 2015. *Word Frequencies in English-Language Literature, 1700–1922*. [Dataset], HathiTrust. https://doi.org/10.13012/J8JW8BSJ

Yau, Nathan. 2019. Visualizing Incomplete and Missing Data. FlowingData. https://flowingdata.com/2018/01/30/visualizing-incomplete-and-missing-data/. Accessed 15 Apr 2019.