Working around Walled Gardens: The Princeton Prosody Archive as Workflow

Meredith Martin, Rebecca Koeser, Mary Naydan

The Princeton Prosody Archive (PPA) is an open-source, full-text searchable database of 6,000+ English-language digitized works about the study of poetry, versification and pronunciation. But the PPA is also — technically and conceptually — a workflow that brings together materials from both HathiTrust Digital Library and Gale/Cengage's Eighteenth Century Collections Online into one searchable interface (figure 1). This workflow is necessary because today's digital research landscape consists of silos or "walled gardens" of scholarly materials controlled by proprietary vendors and digital libraries with layers of copyright restrictions. Designed to keep scholars inside them, these "walled gardens" limit scholars' ability to work across collections. What if scholars could analyze materials from multiple collections at once? What discoveries could scholars make if they weren't limited by a particular vendor's metadata, OCR quality, indexing practices, or built-in Natural Language Processing tools? [1]

The PPA has asked and answered these questions over a sixteen-year process of negotiating with vendors, collecting and curating materials, building the public-facing web application, and making the data usable for computational analysis. Our short paper will discuss how we gained access to additional levels of data by securing Memoranda of Understandings from HathiTrust and Gale/Cengage and worked within their data infrastructures to pull in bibliographic metadata, full-text OCR, and page images from their APIs and servers. Although our work improved their infrastructures and data, our workflows did not always align: for instance, we waited over a year in Gale's development queue for minor yet necessary enhancements to their API to access basic metadata. Similarly, we spent years correcting HathiTrust's inaccurate metadata for hundreds of works but could not feed those corrections back into HathiTrust because of HathiTrust's understandably limited ability to develop individual workflows with each partner research library.

The roadblocks we encountered while trying to integrate our workflows starkly illustrate how Big Tech and for-profit companies circumscribe how scholars conduct academic research, even on public-domain materials. For instance, despite HathiTrust being a not-for-profit collaborative of academic and research libraries that own the physical copies of the digitized works, we needed special permission from *Google* to display page image thumbnails because Google digitized 95% of HathiTrust works in the mid-2000s [2]. We had to prove that our archive of highly specialized texts about prosody — miniscule compared to HathiTrust's 18+ million volume collection (see figure 2) — wouldn't compete with Google Books! This example is one of many that exposes the fragility of the current research ecosystem, in which scholars encounter often invisible limitations around who can access these materials, how they can be used, and even which materials get included in the first place [3].

The PPA imagines an ideal scholarly information landscape in which the curation of data is driven by research questions, rather than by profit or by who happens to have digitized what. Rather than developing new resources from scratch, we imagine sharing workflows, ideas, and code with other scholars to remove barriers and work across — rather than just within — our research library's subscriptions.

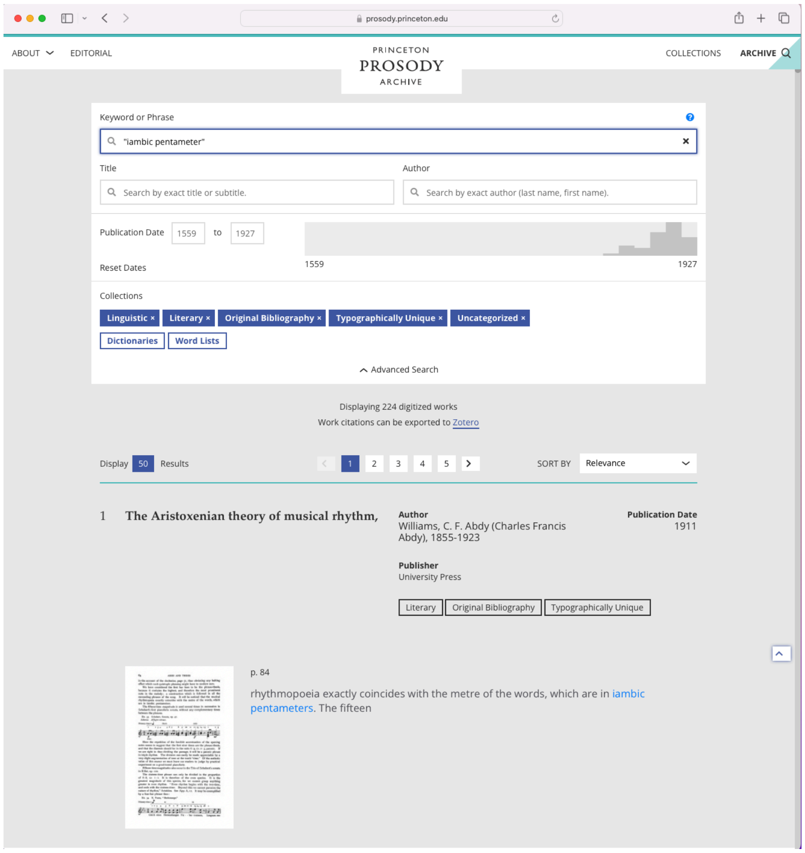
Bibliography:

1. While we are talking about openness *across* resources, there is also a desire for more openness *within* them. As an example, a very recent paper from *CHR2023: Computational Humanities Research Conference* (December 6–8, 2023, Paris, France) discusses the challenges of building out HathiTrust Research Center's Extracted Features dataset to make the non-consumptive analysis it offers more flexible and customizable (John A. Walsh et al., "'The Library is Open!': Open Data and an Open API for the HathiTrust Digital Library," https://ceur-ws.org/Vol-3558/paper7875.pdf).

2. "About the Collection," *HathiTrust*, https://www.hathitrust.org/the-collection/.

3. In *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy* (Northwestern UP, 2019), Roopika Risam discusses the "uneven digitization of literary texts" and calls scholars to recognize "the ways that existing digital archives bear traces of colonialism" (39-40).

Figure 1: Screenshot of the PPA search interface



Keyword or Phrase

"iambic pentameter"

Title

Search by exact title or subtitle.

Author

Search by exact author (last name, first name).

Publication Date 1559 to 1927

1559 1927

Reset Dates

Collections

Linguistic × Literary × Original Bibliography × Typographically Unique × Uncategorized ×

Dictionaries Word Lists

Advanced Search

Displaying 224 digitized works

Work citations can be exported to Zotero

Display 50 Results 1 2 3 4 5 SORT BY Relevance

1 **The Aristoxenian theory of musical rhythm,**

Author
Williams, C. F. Abdy (Charles Francis Abdy), 1855-1923

Publication Date
1911

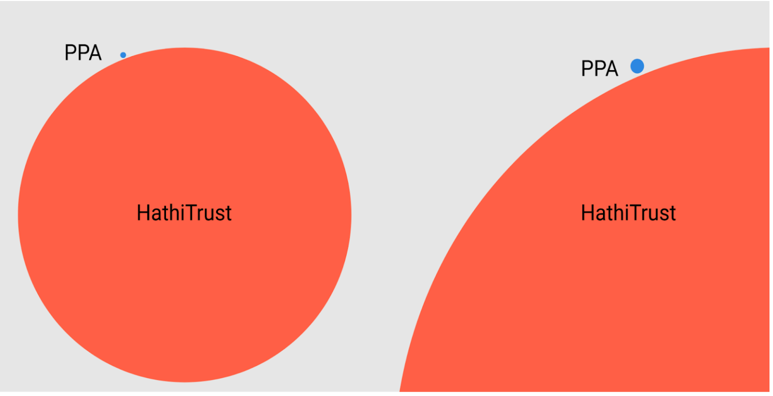
Publisher
University Press

Literary Original Bibliography Typographically Unique

p. 84

rhythmopoeia exactly coincides with the metre of the words, which are in iambic pentameters. The fifteen

Figure 2: Visualization showing the relative sizes of HathiTrust (orange) and the HathiTrust items within the PPA (blue)


PPA
HathiTrust
PPA
HathiTrust