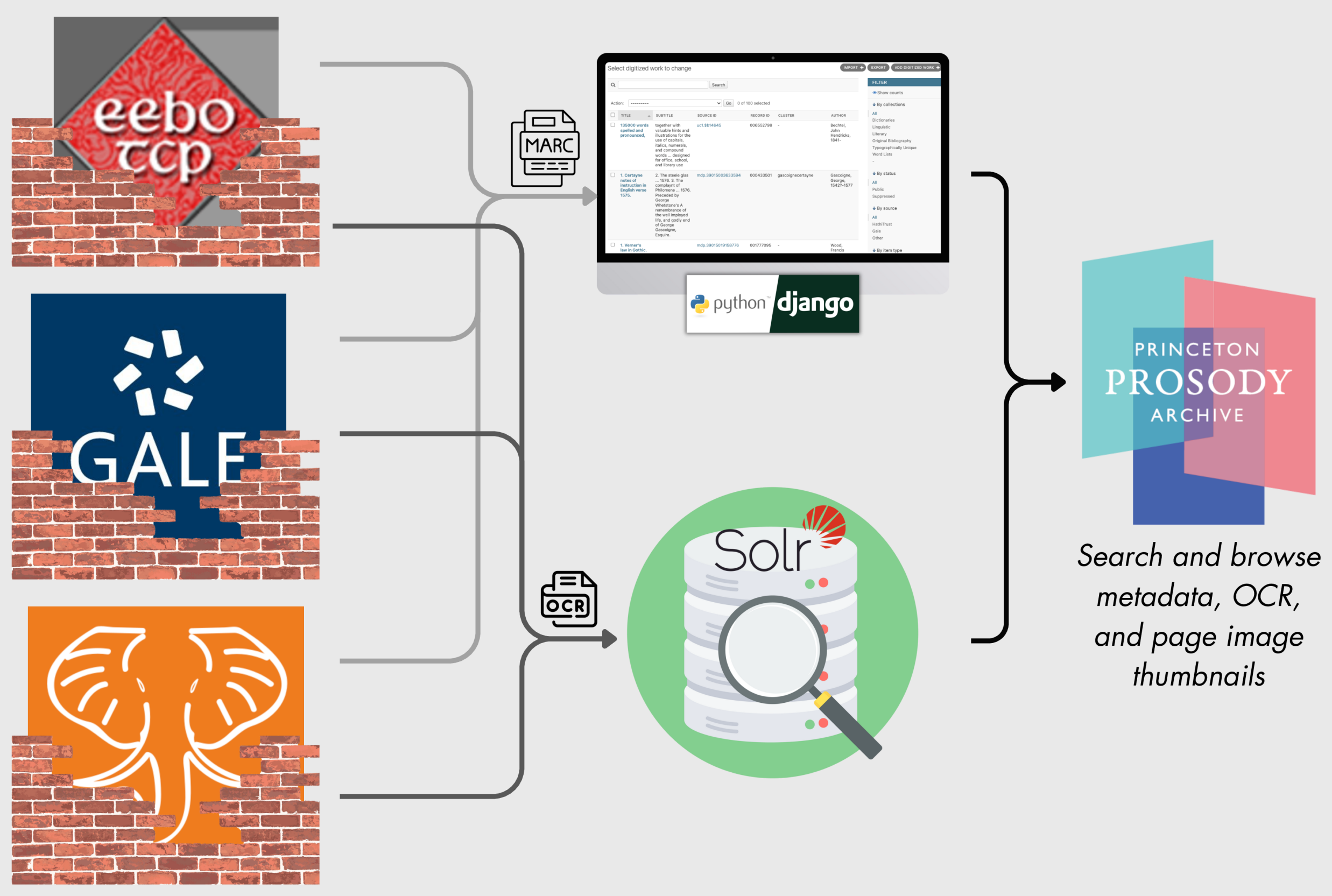
Beyond the Walled Gardens
Reinventing the Digital Research Landscape with the Princeton Prosody Archive

Introduction

Walled gardens are **closed platforms** that control and restrict access to cultural heritage material via subscriptions. Since the 2008 Google Books lawsuit, they have come to dominate our digital research landscape and shape the kinds of projects and analysis that digital humanists conduct [1]. While vendors have increasingly offered **TDM services** *within* their own walled gardens, this **"uncooperative, siloed approach"** poses a number of problems for researchers [2]. For instance, researchers often cannot export or reproduce their results, and they must either **limit their TDM to a single vendor** or perform it multiple times across vendor platforms, which can be inconsistent, cost prohibitive, and block research advances. Additional downstream effects include decreased international collaboration and **avoidance of innovative projects** for fear of litigation [3]. As a result, literature-focused digital humanities projects have typically focused on **digital editions and text corpora from one data source**, such as The Walt Whitman Archive and ECCO-TCP. Other projects, like the Modernist Archives Publishing Project, work with libraries to bring together archival material from a range of special collections into a single interface.

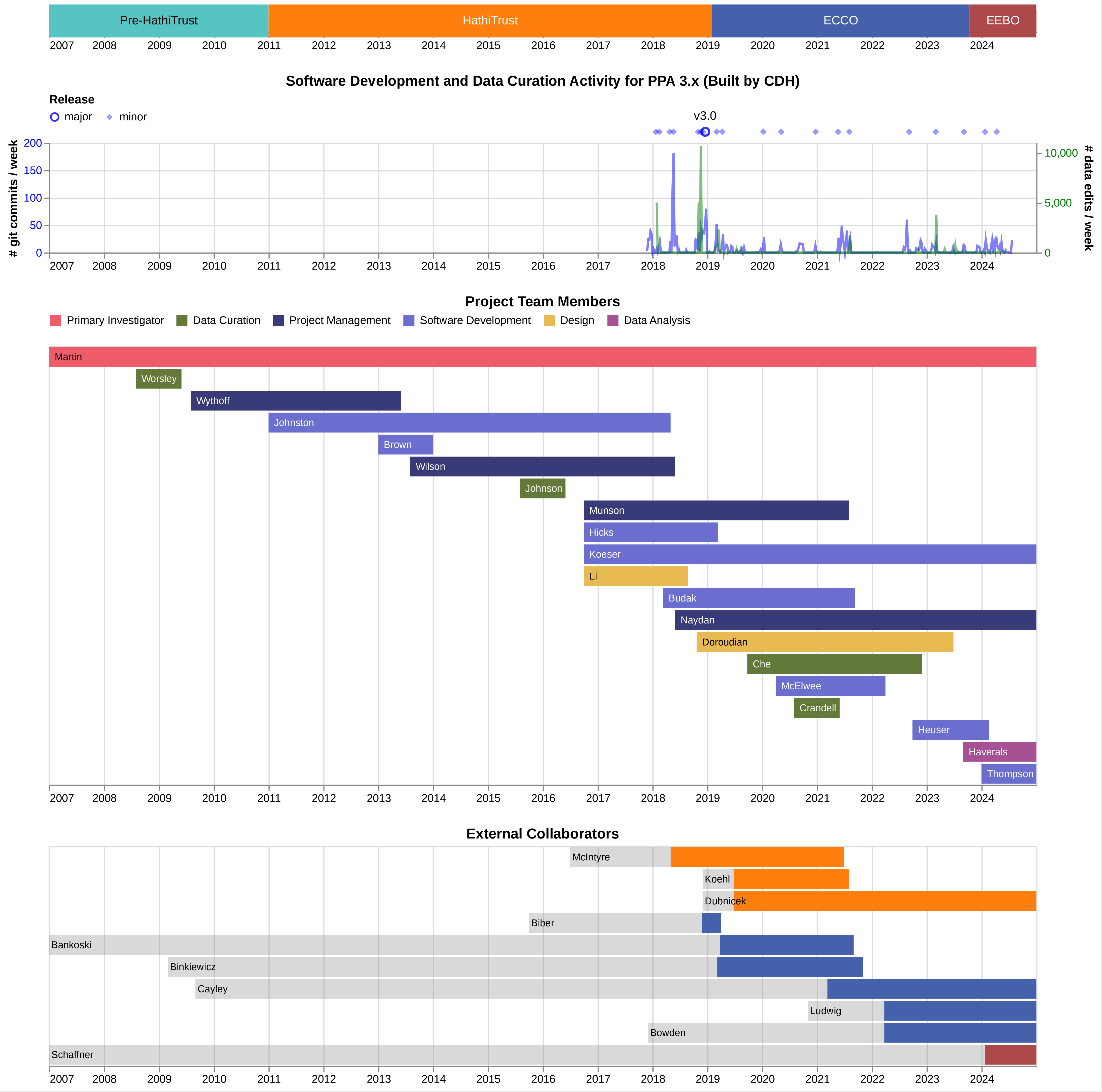
PPA Architecture

The **Princeton Prosody Archive** (PPA) is one of the first humanities resources of its kind to collect, curate, and analyze materials across 3 separate proprietary vendors: **HathiTrust Digital Library** (Google), **Eighteenth Century Collections Online** (Gale/Cengage), and **EEBO-TCP** (ProQuest). It is an open-source, full-text searchable database of nearly 7,000 English-language works published between 1532 and 1928 about the study of versification and pronunciation.


eebo tcp
GALE
MARC
python django
Select digitized work to change
OCR
Solr
PRINCETON PROSODY ARCHIVE
Search and browse metadata, OCR, and page image thumbnails

Our **code** is available on GitHub and could be adapted to combine materials from these sources on *any* topic; there is nothing prosody-specific in the code [4]. We **imagine a future** with many different "PPAs" on many different topics, all cutting across walled gardens. If many project teams seek out their own MOUs with HathiTrust, Gale/Cengage, and ProQuest, or if university libraries withhold their subscriptions until these companies allow cross-platform TDM, we might start to use our **cross-institutional leverage** to incentivize them to reinvent the digital research landscape that our scholarship relies upon, so that we don't have to do it all ourselves.

Scales of Labor by "Era"



The above visualization attempts to surface and credit the amount of **invisible labor** that went into creating the PPA. It is organized by dominant "era" corresponding to the **main source** the project team was focused on integrating between 2007 and 2024. The top double line graph overlays GitHub commits with data curation edits in the django administrative interface; note the major spikes before and around the 3.0 release in March 2019. The middle bar chart displays PPA project team members over time, color coded by primary role and showing the continuity of core members Martin, Koeser, and Naydan. The bottom bar chart shows external collaborators from HathiTrust, Gale, and EEBO; the gray portion displays the full dates of their employment (gathered from LinkedIn), while the colored portions indicate active involvement with the PPA. This visualization is limited and incomplete; there are many more collaborators, including undergraduates and librarians, that we did not have space to include, and project team members often engage in many different kinds of work not captured by their primary role.

Access Comparison

		GALE	eebo tcp
MARC records	open access (bibliographic API)	with library purchase	with MOU
OCR	with MOU	with MOU	open access (TCP)
Page image thumbnails	with MOU	with MOU	☒
Elevated API access	☒	with MOU	☒
Custom API adjustments	☒	with MOU	☒
TDM rights	with MOU, on campus only	with hard drive purchase	open access (TCP)
Restrictions	non-compete with Google	–	–

The Path through the Hedge Maze

Strategies for working with commercially owned/restricted data:

- Leverage the **power of your institution**, its name, and its subscriptions
- Devise **MOUs**
- Cultivate **relationships** with librarians and vendors who can help you negotiate and advocate for your research needs
- Prepare for **change**, such as contacts leaving unexpectedly and changes to the data, interfaces, and technical architecture
- **Document** everything: agreements, meeting notes, email chains
- Cultivate team **continuity**

Poster Authors

All authors are from the Center for Digital Humanities at Princeton University.
Mary Naydan, Project Manager, mnaydan@princeton.edu
Rebecca Sutton Koeser, Lead Research Software Engineer, rebecca.s.koeser@princeton.edu
Meredith Martin, Faculty Director, mm4@princeton.edu

References

[1] Giancarlo Frosio, "Google Books Rejected," *Santa Clara Computer and High Technology Law Journal* 28, no. 1 (Nov. 2011): 81-141.
[2] Peter McCracken and Emma Raub, "Licensing Challenges Associated with Text and Data Mining," *JLSC* 11, no. 1 (2023): 1-14.
[3] Patricia Aufderheide, Brandon Butler, and Kimberly Anastacio, "The Chilling Effects of Obstacles to Accessing, Using, and Sharing In-Copyright Data for Quantitative Research," *Information & Culture* 59, no. 1 (2024): 44-65.
[4] Code: https://github.com/Princeton-CDH/ppa-django

bitly

PRINCETON PROSODY ARCHIVE

THE CENTER FOR DIGITAL HUMANITIES @PRINCETON