

Data management plan.

Classifying a large collection by genre is partly a service for other scholars, and partly a way to start interesting arguments. We don't expect other critics to passively adopt the boundaries we draw for subtle subgeneric categories like "gothic romance" or "the sensation novel." On the contrary, our goal is to highlight interesting ambiguities, and disseminate code that other scholars can use to do their own classification. But we do think it will be possible to establish broadly reliable generic metadata for categories like "verse," "drama," "prose fiction," and "prose nonfiction," and those results should be easy for other researchers to borrow.

Our goal is to make the products of this research as public as possible given United States copyright law. We are using public-domain data wherever possible. Even the portion of the work after 1923 will be done through HathiTrust Research Center, an institution that seeks to facilitate public non-expressive access to works otherwise in copyright.

The collection produced by this project will reside at HTRC. All the metadata we produce will be public; it will be saved both as individual .json files associated with each text, and as a .csv file for the whole collection. Where we identify article divisions or internal boundaries between prose and poetry, these will be marked on the text using TEI-lite. TEI files before 1924 will be public-domain, although special arrangement with HathiTrust may be necessary to access a subset of those files that were originally digitized by Google. After 1924, the metadata will still be public, but the texts themselves will have to reside at HTRC.

The software tools we propose to develop will be open-source, and available on github. Some portions may be built in Python, but the core will be written in Java for reasons of performance and scale; it will be designed to accept HathiTrust data structures. More importantly, all the tools and resources we develop will be embedded in HathiTrust Research Center as parts of a research infrastructure immediately available to other scholars. This will include:

- data-cleaning modules that remove running headers and segment volumes,
- lists of rules for OCR correction in particular periods, and algorithms for contextual correction of phrases ("mortal fin forgiven" => "mortal sin forgiven").
- general-purpose software for automatic genre classification,
- a model specifically trained to classify works in the period 1800-1949,
- manually-classified genre metadata (and author gender) for a collection of roughly 10,000 volumes 1800-1949,
- automatically-classified genre metadata for a collection of more than a million volumes 1800-1949,
- a list of volumes in the collection that proved impossible to classify, which might be one of the most interesting products of this research — providing useful leads even for literary scholars who are not interested in distant reading.