Program Era Project: ACS Final Report

Overview of Project

The academic discipline of creative writing was born at the University of Iowa in 1936 with the founding of the Iowa Writers' Workshop, a program that flourished in Iowa City, paving the way for the Nonfiction Writing Program, the International Writing Program, the Translation Workshop and the Playwriting MFA, and laying the groundwork for the expansion of the discipline across the United States and around the globe. And yet, until recently, literary historians had paid little attention to the history of this creative writing program's growth, or the rich and complex global network the resulted among the alumni and faculty over the course of the century. The objective of the Program Era Project is to ensure that the documentation and curation of this history begins at the University of Iowa, by digitally archiving and cataloging our records and research such that they can be easily accessible and understandable not only to scholars but also to the general public.

We have built and are expanding a database of graduates and instructors and their professional itineraries. Our ultimate objective is to develop a web application for students and scholars to generate interactive data visualizations. Current plans include charts that document changes in the size and demographic composition of creative writing cohorts, maps that can show where writers came from and went to, and network graphs that show the connections amongst advisors and advisees.

Along with data exploration based on archival information, we wish to track, visualize, and compare various features of written work. As all of the written work we are interested in collecting and analyzing is presently under copyright, the HathiTrust collection and HTRC were ideal partners. In applying for the ACS grant we aimed to build a corpus of all available published materials written by Iowa-affiliated writers.

Alternatives Identified

When considering possible methods for assembling a corpus of Workshop writing, the Program Era Project team's options were extremely limited. The chief reason for this is that the overwhelming majority (if not all) of the works produced by IWW authors remain under copyright. Therefore, access to full text digital volumes is limited by law.

The PEP team explored the possibility of building its own corpus of IWW works in collaboration with the University of Iowa Libraries. However, a number of difficulties soon became apparent in collecting such texts at scale. First, the cost of building a large corpus of e-texts was a subject of concern. Secondly, when purchasing e-texts, the formatting and degree of user access to texts held by the libraries was not uniform. Lastly, and most significantly, such texts are not typically available in formats that are readily compatible with computational text analysis methods. Furthermore, converting such texts would present not only additional technical hurdles and project team labor but, more importantly, could present issues with legal restrictions on digital texts.

Therefore, after consideration, the idea of assembling an exclusive corpus was abandoned. A collaboration with HTRC via the ACS program was identified as the only viable option to gain non-consumptive access to the digital texts needed for the Program Era Project.

Data query and preprocessing

Our ideal objective was to capture everything written by everyone in our database, but it quickly became apparent that this would not be feasible. The database contains about 3700 individuals, most of whom graduated and never gained great prominence as writers, and some of whom never published at all. Additionally, several names in the database are fairly common. Fact-checking whether books were false positives under those conditions seemed impractical, so we decided to start off with a shorter list of prominent authors.

We assembled a list of 380 writers, who were either listed as thesis advisors in our database or included in a list of prominent Workshop writers on Wikipedia. After an initial round of querying, we provided additional information such as lifespans to cut down on false positives. Ultimately, we received a capsule containing about 2500 volumes.

Next, we went through a preprocessing stage. In a spreadsheet of all the volume records, we marked out duplicates and false positives (eg, non-Iowa authors with the same name as a Workshop writer). For now, we have marked out non-English works and anthologies or multi-authored works (more on this shortly). We manually added genre information, so that we could focus on items of interest such as poetry, fiction, and essays, and filter out items such as books about writing (unsurprisingly, many creative writing professors have written books about writing).

After preprocessing and filtering, we have a corpus of 1500 unique volumes representing 234 authors. The representation of authors varied widely – Robert Bly accounted for nearly 50 of the volumes, while other prolific authors had one or no volumes included.

Since we were interested in analyzing the writing styles of individual authors, anthologies presented a problem on two counts. Approximately 8% of the volumes were multi-authored volumes, often anthologies of poetry or short stories, and several of the single-authored volumes were collections of short works by that author. Both present interesting challenges. Single-author collections often include works that are repeated across multiple collections, disproportionately amplifying the traits of a single work. On the other hand, a multi-author anthology needs to be divided up before being useful for individual author signal identification.

In our early stages of analysis, we've sidestepped these challenges by starting with a subset of novels. However, despite difficulty of use for single-author style analysis, anthologies present interesting opportunities in other directions of research, such as tracing networks of who included whom in their edited anthologies.

Corpus Processing and Data Collection

The ACS corpus provided to the PEP consisted in a collection of compressed archives, each containing individual page files for each volume found and assembled by the HTRC team. Additionally, for each volume a JSON was included, offering selected metadata on the volume. As the PEP team's aim was to collect a data set of text metrics on each volume in the ACS corpus, it was necessary to develop a tool that would automate the extraction of each volume archive, the ordering and concatenation of its pages into a single volume, and the collection of metrics from each volume. Because the HTRC metadata included, in many cases, information on each volume such as publisher, publication date, and ISBN, the tool was also designed to draw such data from each volume's associated JSON file and include it alongside the metrics gathered by the PEP text analysis tools.

During the process of extracting and concatenating the individual volumes in the ACS corpus, our PEP tools collect metrics on content and stylistic features in each volume. Three specific sets of metrics are now collected. First, the PEP tools measure various stylistic features such as sentence length, punctuation choices, relative use of nouns and verbs, and comparisons of pronoun usage such as male and female or first-person and third-person. Second, the PEP tools track and disambiguate location references, measuring which states and census regions in the United States are most frequently mentioned. Finally, in our latest update, our text analysis software is now able to measure sentiment. By comparing the frequency of words found in a text that are defined as either positive or negative by Stanford's Natural Language Toolkit, the PEP software can assign a score to a text, ranking it in terms of positive or negative sentiment. Coupled with the metadata extracted from the JSON files accompanying each text in the ACS corpus, these collected measurements are then analyzed to identify trends in IWW-affiliated writing.

After the ACS corpus volumes are extracted and concatenated and PEP metrics are collected, three archives are produced. The first archive contains the individual page files for each ACS corpus volume as well as the full text assembled from these page files, which are placed in order and then concatenated. Second, a single archive containing each concatenated volume is produced. Using our list of duplicate or false positives as a guide, volumes which should not be in the archive are then culled from the collection. Finally, metrics for each volume are assembled in a third archive. Using the same list of works utilized to cull the full text collection, false positives and duplicates are removed from this collection before it is concatenated into a single collection of data to be exported out of the HTRC capsule.

Results

We now have a database of metrics on stylistic features and location references for each volume in our corpus. This has allowed us to rank and compare volumes in our database and find intriguing trends which hold the potential to drive future scholarly interventions. We can, for instance, discern clusters of authors whose works feature male pronouns over female, or identify writers and volumes whose works feature the longest sentences or most adjective-rich prose. These stylistic standouts pose interesting questions and suggest the significant potential the database has for future exploration, particularly were specific standout writers to be compared to their advisors or advisees.

More significantly, we have used our data to develop a full-length scholarly article which has been submitted for publication. Thanks to our location-tracking metrics, we were able to examine whether an authors' time at the Iowa Writers' Workshop influenced what locations they mentioned in their work. We discovered that, in fiction produced by authors affiliated with the IWW, Iowa was mentioned significantly more than in a comparison corpus assembled by the TxtGeo project. Moreover, when we assembled a list of the authors who mention Iowa most frequently, we found that only a single author on that list was a native Iowan. We have given the title of "Squatter Regionalism" to this literary phenomenon where writers write about the place in which they taught or studied Creative Writing. Thus far we have received positive impressions from the editorial staff at *Cultural Analytics*. We await further reports from readers.

Future Work

We anticipate further evolution of our location-tracking methodologies based on discussions with *Cultural Analytics* and its readers. Furthermore, we are currently exploring potential methods for how we might develop scholarly work based on the metrics drawn from the poetry in our corpus. We also

anticipate collecting results from our new sentiment analysis feature and how we might correlate that data with our previously collected metrics. Lastly, we are interested in exploring authorship metadata on volumes in our corpus and comparing the inclusion of IWW authors in collections assembled by IWW faculty members to networks of advisor/advisee relationships at the University of Iowa.

Related Publications

https://dsps.lib.uiowa.edu/programera/2017/11/20/collaborating-with-hathitrust/

https://dsps.lib.uiowa.edu/programera/2018/02/09/breaking-down-the-htrc-data-capsule/