PROJECT SUMMARY

Jasmine Wong, MSLIS ‘20

IS590 Open Data Mashups, Fall 2019

OBJECTIVES

Aggregate author citation metadata from literary databases to explore prevalence

Build a corpus of author citations for use in literary history research

DATASETS

English Short Title Catalog (HathiTrust)
Early English Books Online (ProQuest)
Open Syllabus Project
Project Gutenberg
Oxford Text Archive

RESULTS

CORPUS SIZE

1,427 authors

TOP AUTHORS

Daniel Defoe
William Shakespeare
Jonathan Swift

METHODS

COLLECTION

Identify viable datasets

Extract data in machine-readable form

CLEANING

Hand-edit character encoding errors

Standardize dataset structure

TRANSFORMATION & PROCESSING

Combine datasets using Python Pandas

Consolidate matching author records across datasets

Calculate relative frequency of each author in full corpus

TIMELINE

August

Project development and dataset exploration

September

Dataset selection and data extraction

October

Dataset extraction and cleaning

November

Data processing and analysis

December

Final analysis and submission