Book Sleuth:

A Search Engine for Concept Discovery

HARVARD UNIVERSITY

Project Overview

- Search large corpora of digitized texts and find specific passages directly relevant to your research
- Rank-ordered results are easy to review, sort, and filter, so that researchers can find material that matters most
- Small, 1000-word passages make it easy to identify which parts of books to examine first
- Perfect for researchers wanting to discover new, non-canonical material

Current Corpora

- Gale's Eighteenth Century Collections Online (207,000 volumes)
- Proquest's Early English Books Online (30,000 volumes)

Corpora to Acquire

- Google Books
- HathiTrust non-Google set
- JSTOR
- Corpora directly relevant to research needs

Upcoming Project Phases

- Acquire more corpora from vendors and prepare for search
- Structure data to allow for increased volume of data and scaling search demands
- Mount search engine on a Harvard-facing prototype website, for the whole research community to use
- Acquire funding to support further development and expanding compute demands
- Communicate with other departments, divisions, and faculties, both to encourage use and to plan for future stages of exapansion
- Brand the tool and seek media coverage