Implementing Ngram Search across Yale Digital Collections

Peter Leonard  
Librarian for Digital Humanities Research  
Sterling Memorial Library  
203-432-5077  
peter.leonard@yale.edu

Lindsay King  
Public Services Librarian  
Haas Arts Library  
203-436-8052  
lindsay.king@yale.edu

PURPOSE AND EXPECTED OUTCOME

Yale University Library acquires, creates, stewards and preserves many large electronic text collections as part of its ongoing digital initiatives. We propose to cast new light on this vast amount of digitized material by implementing Bookworm, an open-source Ngram visualization tool. Bookworm visualizes trends in word frequency within collections as diverse as the Library of Congress’ Chronicling America Historic Newspapers and ArXiv. With the support of SCOPA, researchers in New Haven — as well as across the globe — will be able to derive these insights from Yale University Library collections.

From Google Books to Yale University Library Collections

The concept of “Ngram search” first attracted widespread attention in late 2010 as part of the Google Books Ngram Viewer, which allows comparison of the relative frequency of multi-word phrases. Shortly after the launch of the Google tool, however, the Harvard Cultural Observatory announced a parallel implementation in an open-source software packaged called Bookworm. Implementing and extending Bookworm to work with YUL collections will help establish models of use and support for text mining projects by Yale researchers. An example of what can be uncovered in an Ngram tool is this visualization of the phrases “need to” and “ought to”:

![Ngram Visualization](image-url)
As this example from Google Books shows, one of the powerful aspects of an Ngram tool is its ability to work on in-copyright material by showing the relative frequency of specific terms without necessarily exposing the underlying texts. Although the majority of materials we imagine using Bookworm on at Yale are either already online (in either open or subscription-only format), this Ngram tool could potentially show broad patterns from restricted collections, such as the Kissinger and Fortunoff archives, without requiring that the entire archive be made public. The ultimate decision to use such a tool would rest with those projects’ administrators; however, we wish to point out the potential utility of an Ngram tool even on closed archives.

**METHODOLOGY AND TIMELINE**

**Stage 1: Internal Test Implementations** Fall 2013 (already underway)
The code for Bookworm is openly available at [https://github.com/bmschmidt/Presidio](https://github.com/bmschmidt/Presidio). We have already begun to investigate the data transformations required to ingest some sample corpora (Vogue and Yale Daily News) into the tool. Peter Leonard will do as much of this integration work as possible before inviting the tool’s author to campus (see Stage 2 below), in order to have test implementations to work with and therefore more specific questions to ask. Because the Vogue archive has been identified as a large (120-year) corpus, Lindsay King, who has a track record of publication and presentation on fashion history, will serve as an evaluator of the Ngram tool’s utility for fashion research.

**Stage 2: Customization and Extension with Tool Author** Spring 2014
We propose to bring one of the main authors of the tool, Benjamin Schmidt, to Yale for a two-day engagement to:

- work with Yale scholars and librarians on implementing Bookworm
- brainstorm future directions and extensions to Bookworm
- implement a Yale-hosted “build-your-own-Bookworm” service

Schmidt is a newly-appointed Assistant Professor of History at Northeastern University, where he serves as a core faculty member of the NULab for Texts, Maps, and Networks. As Northeastern University is one of the hubs in the New England Digital Humanities Network, bringing one of the core digital humanities faculty to Yale to work on a concrete project will help cement ties in this emerging regional network. We will engage with the Whitney Humanities Center-sponsored Digital Humanities Working Group, in order to attract the widest possible audience for Schmidt’s time on campus.

**Stage 3: Implementation across selected Yale digital text collections** Summer 2014
We have identified a number of “shovel-ready” digital collections at the library on top of which Bookworm can be implemented. Our goal is to be able to deploy Bookworm across two different kinds of collections: Yale-digitized and vendor-digitized.

- **Potential Yale-Digitized Collections**: Yale Daily News, Yale Indian Papers Project, various Arcadia grants, Yale Dissertations, Jonathan Edwards Papers
- **Potential Vendor-Digitized Collections**: Vogue (ProQuest), Liberty (Gale Cengage), others.
Stage 5: SCOPA talk about using Bookworm on Yale projects  Spring 2015

EXPENSES

<table>
<thead>
<tr>
<th>Item</th>
<th>Amount</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travel</td>
<td>$200</td>
<td>Train</td>
</tr>
<tr>
<td>Lodging</td>
<td>$500</td>
<td>Two nights at The Study</td>
</tr>
<tr>
<td>Honorarium</td>
<td>$500</td>
<td>For Ben Schmidt, co-author of Bookworm code</td>
</tr>
<tr>
<td>Testbed Mac Mini Server</td>
<td>$1100</td>
<td>For intensive text processing</td>
</tr>
<tr>
<td>Sprout Server</td>
<td>$480</td>
<td>For web serving</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$40/month x 12 months</td>
</tr>
<tr>
<td>TOTAL</td>
<td>$2780</td>
<td></td>
</tr>
</tbody>
</table>

BENEFIT

This project will serve as a model for text mining large digitized collections at Yale University Library. Bookworm will add substantial scholarly value to both open and restricted digital collections at Yale, including Yale-digitized content and vendor-digitized content. Finally, deploying the Bookworm code on YUL collections and bringing Ben Schmidt to campus will connect Yale scholars with an emerging digital humanities network in the Northeast.

Impact on Library IT and Sustainability

It is important to note that the addition of a Bookworm interface to existing online collections does not impose any changes or disruptions to the existing digital infrastructure – the tool merely links to existing stable URLs (often called permalinks) that already exist in most digital repositories. This includes a number of current and legacy systems in use at YUL, such as Fedora/Hydra/Blacklight and ContentDM. Likewise, Bookworm links out to subscription services licensed through vendors such as ProQuest and GaleCengage, utilizing those databases’ existing security and authorization features (Yale IP address or VPN requirement, etc.)

In our budget we have identified, and budgeted for, two different types of digital infrastructure necessary for this project.

First, we have specified a higher-performance (non-virtual) Mac Mini, to be used for the intensive data processing and indexing required during the initial data import phase. We
anticipate connecting this machine to the 8-terabyte RAID array that Library IT is providing to Peter Leonard, which is intended to be used for short-term high-performance and high-capacity scratch storage space of this kind.

Second, we have specified an ITS Sprout Server (a virtual machine) to serve as the production web and database server for Bookworm. Peter Leonard already maintains one of these servers, provided to him by Michael Friscia in Library IT, and has experience with running various kinds of web applications in this environment. The Sprout Server is a pay-by-the-month service, and following the expiration of the 12-month period specified in this grant, we will evaluate the utility of the Bookworm project and decide upon next steps to ensure its future service. We anticipate that in a year’s time we will have adequate information about the technical needs of this tool to be able to advise Library IT about the costs and overhead of maintaining the tool on an ongoing basis as part of the production infrastructure of the Library.